A STATISTICAL MODEL OF BACKGROUND AIR POLLUTION FREQUENCY DISTRIBUTIONS

M.Ya. Antonovsky
International Institute for Applied Systems Analysis, Laxenburg, Austria

V.M. Buchstaber
All-Union Research Institute of Physiotechnical and Radiotechnical Measurements, Mendeleevo, USSR

E.A. Zelenuk
Natural Environment and Climate Monitoring Laboratory, Moscow, USSR

RR–91–9
June 1991

Reprinted from *Environmental Monitoring and Assessment*, **16**:203–252, 1991.

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
Laxenburg, Austria

Research Reports, which record research conducted at IIASA, are independently reviewed before publication. However, the views and opinions they express are not necessarily those of the Institute or the National Member Organizations that support it.

Printed by Novographic, Vienna, Austria

Preface

This paper describes an approach for identifying *statistically stable* central tendencies in the frequency distributions of time series of observations of background atmospheric pollutants. The data were collected as daily mean values of concentrations of sulfur dioxide and suspended particulate matter at monitoring stations in the USSR (3), in Norway (1), and in Sweden (1).

In their approach, the authors use well-developed statistical techniques and the usual method of constructing multimodal distributions. The problem is subdivided into two parts: first, a decomposition of the observations in order to obtain a description of each season separately and second, an investigation of this description in order to derive statistically stable characteristics of the entire data set. The main hypothesis of the investigation is that dispersion processes interact in such a way that in the zone of influence of one process (near its mode) the "tails" of the other process should not be taken into account. This permits illumination of interrelations between the physics and the chemistry of the atmosphere.

During the last 15–20 years, a wide range of monitoring programs has been initiated at national and international levels including, for example, the European Monitoring and Evaluation Program (EMEP) under the auspices of the ECE, and the Background Air Pollution Monitoring Network (BAPMoN) under the auspices of the WMO.

The flow of data from the system of monitoring stations has led to national and international projects for the development of extensive environmental data bases such as NOAANET (NDAA)/ GRID/GEMS/UNEP/ NASA, etc. The degree of information obtained should be sufficient for the goals of the analysis but often there is an overabundance of such data. The methods discussed here therefore help in air pollution assessments, particularly for distinguishing the baseline components and their trends over decades.

BO R. DÖÖS
Leader
Environment Program

A STATISTICAL MODEL OF BACKGROUND AIR POLLUTION FREQUENCY DISTRIBUTIONS

M. Ya. ANTONOVSKY

International Institute for Applied Systems Analysis, Laxenburg, Australia

V. M. BUCHSTABER

All-Union Research Institute of Physiotechnical and Radiotechnical Measurements, Mendeleevo, U.S.S.R.

and

E. A. ZELENUK

Natural Environment and Climate Monitoring Laboratory, Moscow, U.S.S.R.

(Received December 1988)

Abstract. This paper describes an approach for identifying *statistically stable* central tendencies in the frequency distributions of time series of observations of background atmospheric pollutants. The data were collected as daily mean values of concentrations of sulfur dioxide and suspended particulate matter at five monitoring stations – three in the USSR, one in Norway, and one in Sweden.

The approach uses statistical techniques and methods for constructing multimodal distributions. The problem is subdivided into two parts: first, a decomposition of the observations in order to obtain a description of each season separately and second, an investigation of this description in order to derive statistically stable characteristics of the entire data set. The main hypothesis of the investigation is that dispersion processes interact in such a way that in the zone of influence of one process (near its mode) the 'tails' of the other process are not observed. This permits illumination of interrelations between the physics and the chemistry of the atmosphere.

1. Introduction

The present study is devoted to the statistical analysis of background air pollution monitoring data, having as its objective the design of a statistical model of background air pollution and its application for the determination of statistical characteristics describing the probability laws governing the behavior of impurities in the atmosphere. This study is an extension of notes by Izrael *et al.* (1985), Izrael *et al.* (1987), and Antonovsky *et al.* (1985).

Statistical models of air pollution distribution have been widely discussed in the literature (see, for example, Augustinyak and Sventz (1982); Berlyand (1975); Berlyand (1984); Benarie (1982); Mage (1981)). However, background monitoring data possess certain specific features, creating difficulties in the use of traditional models (such as, for example, the two-parameter lognormal distribution LN2 (Harris and Tabor, 1956; Larsen, 1961). Measurements of background air pollution levels are conducted in areas where the direct effects of strong pollution sources are practically excluded. This implies that the observed data variability is to a considerable degree due to the effects of large-scale atmospheric processes, that determine the mode of occurrence of different concentration levels in the area, rather than to the effects resulting from point sources of pollution. Most of the air pollution models employed are designed for use under the

Environmental Monitoring and Assessment **16**: 203–252, 1991.

assumption of the existence of point sources. Studies of the probability concentration distribution laws for the atmosphere of normal regions allow one to get an idea of the qualitative mechanisms governing the formation of different concentration levels. Statistics, describing these laws, reflect certain regularities in the formation mechanisms and can be used for assessment of background air pollution. Such an approach enables one to validate statistically the intuitively derived concepts of the normal (background) level as the mean of the minimal measurements for a given time-interval (Rovinskii and Buyanova, 1982), or as the minimal but most distinctly expressed concentration level, typical of the region (Izrael, 1984; Rovinskii and Wiersma, 1987). The derived statistics represent an informative description of the time series of background air pollution monitoring data and, in turn, can be used to obtain explicit inferences bearing on the nature of the measurements and their behavior.

The major stages in designing, analysis and application of the statistical model of background air pollution are as follows:

(1) Statistical analysis of background air pollution monitoring data. Studies of the logarithmic concentration distribution functions for data series of different time-intervals.

(2) Investigation of the possiblities of describing the logarithmic concentration series by multimodal distributions, and the physical prerequisites for the origin of multimodality.

(3) Simulation of data series in terms of composite distributions of a specific type, and development of graphical methods for estimation of performance parameters.

(4) Description of seasonal observational data series by central tendencies of multimodal frequency distributions. Development of techniques for identification of statistically stable grouping intervals.

(5) Analysis of statistically stable grouping intervals and their manifestations in seasonal and multiyear data series.

(6) Analysis of the air pollution components described by statistically stable grouping intervals; comparative analysis of the components and their manifestations at different background monitoring stations; development of recommendations for the assessment of background concentration levels.

2. Present Status of Statistical Analysis of Air Pollution

Descriptive air-quality models have been employed in routine investigations since the 1950s. A review of existing models can be found, for example, in Mage (1981).

One of the earliest models to be used is the two-parameter lognormal distribution LN2, with a density function:

$$f(x) = \frac{1}{2\pi\sigma} x \exp\left(\frac{-(\ln x - \ln a)^2}{\sigma^2}\right).$$

Lynn (1976) was among the first to study the applicability of several probabilistic models to air pollution data. The analysis involved the normal law, LN2, the three-parameter lognormal distribution LN3, the I and IV types of the Pearson distribution and

the Gamma-distribution. The conclusion was drawn that the LN2 was the best of all the above-cited distributions. Here a situation occurring frequently in statistical analysis was observed. Namely, in many cases a distribution can be selected (even among those cited above) that most closely approximates the distribution of the sampled data. However, not one of these distributions can be applied to the description of all types of samples of aerometric data. For their description, several distributions should be employed. However, the LN2 distribution is of greatest value.

For instance, in Mage and Ott (1975) the authors conclude that all air pollution data studied by them reveal a common behavior in their deviations from the LN2 – their distribution functions plotted on lognormal probability paper demonstrate characteristic 'curving'. In order to take account of this effect, they suggest using the LN3 model – a three-parameter lognormal distribution.

In de Nevers *et al.* (1979), after analytical treatment of a large number of event-data on atmospheric particulate matter, the authors distinguished not one (as in the former example) but four types of deviations from the straight line, typical of distribution functions plotted on lognormal probability paper. The authors analyzed in detail the reasons for such deviations and proposed to describe them by a combination of two LN2 distributions. In the same work, an example is given illustrating how in reality such a meteorological situation leading to a 'composite' distribution can arise, and an analytical treatment is presented of real data corresponding to such a situation. It is obvious that, from the point of view of increasing model applicability, the last line of attack on the problem is best. By retaining the well-studied and convenient LN2 distribution as the base-distribution, one may perform a uniform description of practically all observed deviations from LN2 by postulating that several different types of meteorological processes affect the concentrations.

We have chosen the two-parameter lognormal law of pollutant concentration distributions – the LN2. Of all the laws studied, this is most widely used, owing to the fact that it performs well for all pollutants within any observational area, and for various time averages and, most likely, reflects certain general conditions in the formation of different air pollutant concentration levels. Taking into account the fact that we are often confronted with the necessity of studying distributions that deviate from LN2, we adopt here the hypothesis postulating an increase of model applicability by the use of combined LN2 distributions.

3. Construction of a Statistical Model Simulating Background Air Pollution Frequency Distributions

3.1. Estimates of Background Concentration Levels

Air-pollution background monitoring stations have been established in the USSR and in many other countries within biosphere reserves, also in localities not subjected to the influence of any apparent source of pollution. These programs involve measurements of air-pollutant concentrations. Since 1976 such aerometric data have been accumulated in

the U.S.S.R. which makes it possible to estimate background concentration levels for particular regions, to analyze the data for different regions and for the world as a whole, to study the principles governing the formation of different concentration levels, and to obtain estimates of normal air pollution concentrations over continents (Burtseva, Lapenski *et al.*, 1982); Burtseva, Volonseva *et al.*, 1982; Pastukhov *et al.*, 1982). Annual data publications have begun (see, for example, Bulletin of background pollution of the natural environment in the region of East-European Members-Countries of CMEA, 1982, 1983).

The data on heavy metal concentrations in the area of the 'Borovoe' station are discussed in Burtseva, Lapenko *et al.* (1982). In the case of lead, the lower limit of measurement error was found to be 0.5 ng m^{-3}, the coefficient of variation not exceeding 20%. According to the data presented in Burtseva, Volosnea *et al.* (1982), lead concentration measurements at background monitoring stations are performed within an accuracy of about 10%. The data represent daily mean concentrations in the lower atmosphere. Analyusis of the histograms of daily mean values for lead concentrations measured over a four-year period, 1977-1980, shows a strong asymmetry in the frequency distribution, with a pronounced concentration maximum in the left lower quartile and a long 'tail' in the right upper quartile. Burtseva, Lapenko *et al.* (1982) used the histograms for simple statistical inferences on the possibility of obtaining relatively stable estimates of lead concentration levels, the major maxima in the frequency distribution being chosen. For the samples in Burtseva, Lapenko *et al.* (1982), such an interval included 65–85% of the observations. The upper limit of the interval was taken as the upper estimate of the background concentration level; thus, according to the authors' estimates, the background concentration level in the atmosphere for lead in the area of the 'Borovoe' station is between 0.5 to 30 ng m^{-3}. For the four years studied, no clearly evident time changes in the concentration distributions occurred; during 230-310 days per year, the concentrations varied within the limits typical of normally pure continental areas.

The proposed method for estimation of the background concentration level has a number of shortcomings. One of these is that the method does not explain the behavior of the concentrations in the frequency distribution. For instance, in Burtseva, Lapenko *et al.* (1982), the authors could not offer a plausible explanation for the increase in the frequency of lead concentrations in the interval of 30–60 ng m^{-3} in 1979, or the presence of arsenic concentrations in the interval of 3–6 ng m^{-3}) for 30% of the observations in 1980 (the arsenic background level being defined at 1–3 ng m^{-3}). Analysis of the possible various types of effects of meteorological and other conditions on concentration variations fails to explain the observed events (Burtseva, Lapenko et al. 1982). Analysis of background monitoring data for sulfur dioxide was performed in Pastukhov *et al.* (1982), the average monthly concentrations varying between 0.3 to 18.9 μg m^{-3} during the period of investigations – from 1977 tot 1981. The highest values were recorded during the winter, the lowest – during the summer, which is a general result found also in data from the Repetek and Berezina biosphere reserve, background monitoring stations. The annual cycle is associated with two factors – the considerable increase in anthropogenic emissions from fuel-burning during the cold periods of the year, on the one hand, and the drop in the

rate of oxidation of sulfur dioxide, on the other hand. Analysis of the monthly concentrations of sulfur dioxine, separately performed for the warm and cold seasons, made it possible for the authors (Pastukhov, 1982) to estimate the sulfur dioxide concentration level in the area of the 'Borovoe' station at 0.5–1.0 μ m^{-3} – for the warm period and at 3.2– 13.7 μg m^{-3} for the cold period. Similar analysis of the verage monthly values at the 'Berezina B.R.' and 'Repetek B.R.' background monitoring stations gives the values 1.0–2.4, 10 μg m^{-3} – for the first and 0.3, 1.0 – for the second. Analysis of meteorological conditions and trajectories indicated that the extreme concentration values cannot be unambiguously correlated with the vector wind directions in the 'Borovoe' station area. The derived estimates for different observational areas are incommensurate and doubt arises concerning their possible use in estimating characteristics of continental and global background concentration levels.

The data to be used are from three background monitoring stations – Borovoe, Berezin biosphere reserve, and Repetek biosphere reserve in the U.S.S.R. Descriptions of the data are given in bulletins (Bulletin of background pollution of the natural environment in the region of East-European Members-Countries of CMEA, 1982 and 1983). The techniques used to derive the data and a discussion of their reliability can be found in Burseva, Lapenko *et al.* (1982), Burseva, Volosneva *et al.* (1982) and Pastukhov *et al.* (1982).

In the present study, three pollutants have been selected – sulfur dioxide, lead, and total suspended particulates, for which daily observations were available during 1976-83 at the Borovoe station and 1980-83 at the Berezina and Repetek stations. The three pollutants differ according to their physical-chemical behavior, and the stations are located in different physical-geographical areas. A joint analysis of the sampled data with a view to finding common statistical characteristics can enable one to define some common principles governing the behavior of air pollutants, and can provide a basis for designing techniques for evaluation of background pollutant concentration levels on a wide scale – both in space and time.

The first stage of statistical analysis should be the construction of the statistical data model. Then, the statistical characteristics describing the data series can be investigated, and their applicability for obtaining non-statistical conclusions can be explored. Techniques for designing statistical models and the use of the statistical information in hydrometeorological and geophysical applications are described in Aivazyan *et al.* (1983), Gruza and Reitenbach (1982) and Kleiner and Gradel (1980). In Aivazyan *et al.* (1983), some general techniques used: in designing statistical models are presented. In practice two different methods of analysis are used: mathematical, relying on theoretical-probabilistic considerations, and computational – by way of direct reproduction of the model function on a PC. The first method calls for hypotheses and *a priori* assumptions concerning the data that should serve to validate the choice of model; the second requires some preliminary formalized knowledge of the data, that could be reflected in algorithmic form, and could be used to develop or refine the theoretical-probabilistic method. In the present study, both of these mutually complementary methods are employed: the first stage, presumably, should involve the development of certain general theoretical-probabilistic concepts of the model.

In Figures 1, 2, 3, plots are shown that characterize the lead concentration distributions at the 'Borovoe' station during the four-year period of observations. Because much of the subsequent analysis is based on studies of these plots, we shall dwell upon them. These plots portray graphically the empirical density and comulative distribution functions (1 and 2), and depict the deviation of the empirical density function from the theoretical one (3).

Histograms are often constructed when the number of observations becomes large. The length of the interval is taken equal to

$$h = \frac{x_{\max} - x_{\min}}{10 \lg (N)+5}, \tag{1}$$

where $x_{\max}$ and $x_{\min}$ are the maximal and minimal points on the logarithmic concentration scale for the given sample, N-the number of observations in the sample.

The distribution function is plotted on normal probability paper as distribution quantiles against the observed variable,

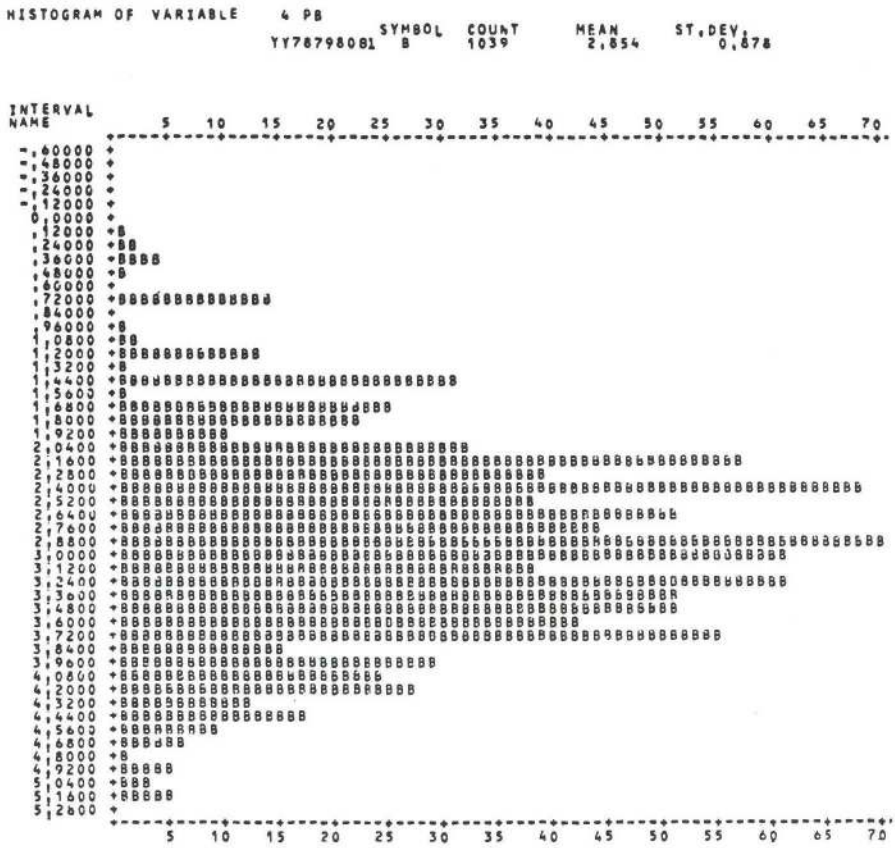

Fig. 1. Histogram of logarithmic concentrations of lead. Borovoe station, 1978-1981.

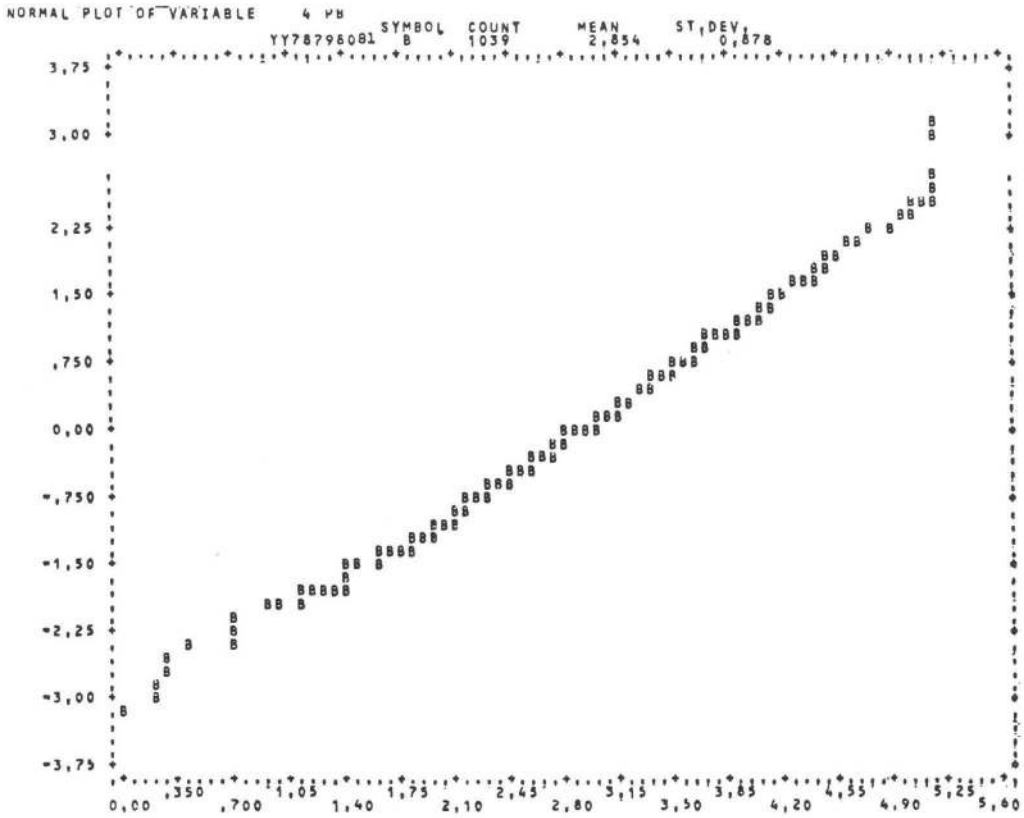

Fig. 2. Normal plot of cumulative logarithmic concentrations of lead. Borovoe station, 1978-81.

$$F(x_n) = \Phi^{-1}\left[\frac{3n-1}{3N+1}\right] \tag{2}$$

where n is the number of the variable x_n in the varational series, arranged in ascending order. The value of the $F(x_n)$ function corresponds to the probability

$$3n-1 \;/\; 3N+1$$

of the centered and normalized normal distribution

$$\Phi(t) = \int_{-\infty}^{t} N(x; 0,1)\mathrm{d}x,$$

where

$$N(x; 0,1) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\frac{x^2}{\sigma^2}\right]. \tag{3}$$

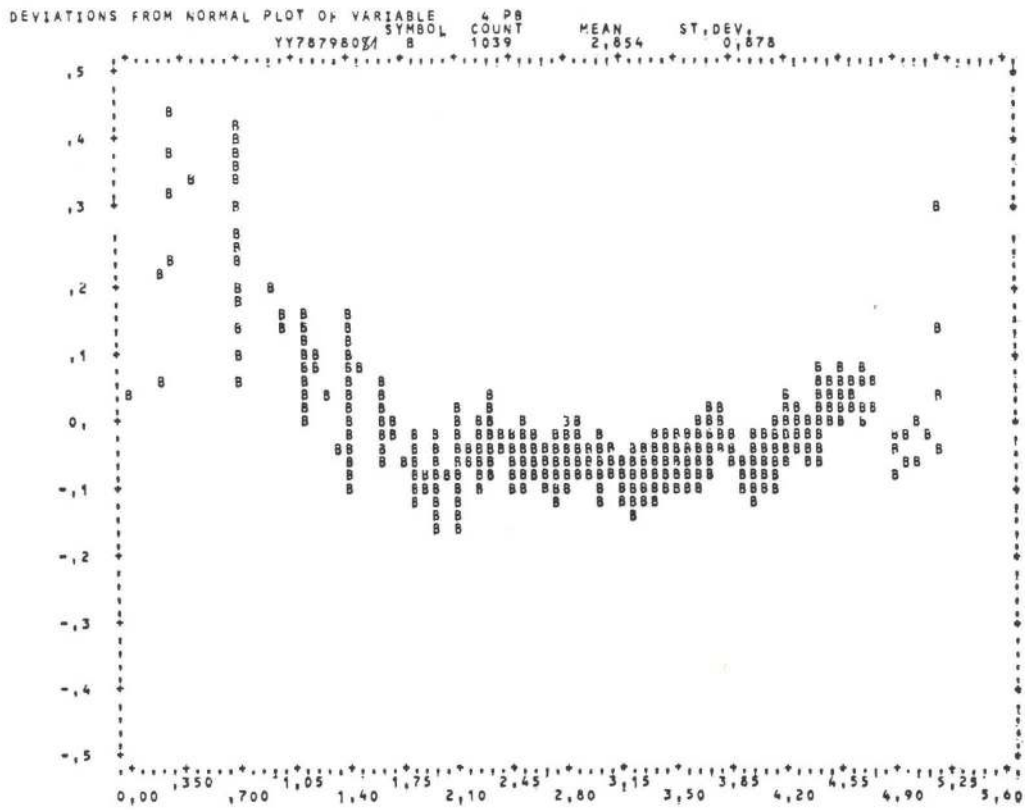

Fig. 3. Deviations from normal plot of logarithmic concentrations of lead. Borovoe station, 1978-1981.

Equation (4) represents Equation (2) with linear trend removed:

$$F(x_n) = \Phi^{-1}\left[\frac{3n-1}{3n+1}\right] - \frac{\bar{a}-x_n}{\bar{\sigma}} \tag{4}$$

where $\bar{a}$ and $\bar{\sigma}$ denote the sample average and variance, respectively. This equation shows the deviation from the straight line, specified by estimates of parameters $\bar{a}$ and $\bar{\sigma}$, and thereby gives a qualitative display of the degree of agreement between the event-data and a LN2 distribution, graphically revealing the nature of inconsistencies with the theoretical distribution.

Discussions of the problems concerned with plotting and evaluation of the distributions by employing graphs of this type can be found in Aivazyan (1983) and Kleiner and Gradel (1980).

As can be seen from Figures 1, 2, 3, the empirical density and distribution functions, as expected, differ from the theoretical ones. The question as to how to proceed in the case of such deviations is discussed at length in Aivazyan (1983). It is obvious that if we have available a sufficiently large class of model densities, for example the Pearson curves, we can find a density function that best approximates the behavior of the empirical density

under study, and, in the long run by expanding the number of hypothetical model densities, we can attain a very high degree of approximation, even in cases of 'crevices' in the model density frequency curves. However, the result has an essential shortcoming, which can be easily perceived when we attempt to apply the model law to the description of model density for any other sample from the same statistical population. In most cases the attempt is a failure. As a consequence, this approach cannot be used to solve the major modeling problem – expansion of the regularities perceived in the behavior of the sampling data over the general population. Thus, in analytical treatment of the data with the purpose of defining common statistical characteristics of air pollution, we shall use model laws and statistics that, perhaps, are less than optimal in terms of formal criteria, but have characteristics that are of much greater importance in our investigations, namely, degree of stability and invariance of the derived results with respect to methods of sample organization, different types of pollutants and geographic areas. Let us consider from this point of view the characteristics common to the distributions of the data plotted in (1), (2), (3).

Figures 4–6 show the empirical density distribution functions, the empirical distribution functions on normal probability paper, and the deviation from the normal distribution function (termed hereafter the histogram, normal graph and deviation from the normal graph) for logarithmic concentrations. For comparison with the model law, we can use the series generated by a random-number-generator.

The distributions are quite similar to the normal ones. However, when these multi-year data-series are divided into seasonal data series, i.e. from May to September, and from November to March, then the departure from normal becomes apparent. Generally, the data exhibit a lognormal distribution. This is due to the fact that the deviations from the straight line on the respective graphs, although causing distortions in the form of the line, are not so great as to obscure the normal distribution. It is obvious also that this lognormal distribution is formed under the effects of a large number of diverse factors, among which are yearly and seasonal variations. Probably a plausible explanation is offered also by the hypothesis of a similar influence of the factors reflecting the effects produced by the background constituents, anthropogenic local sources, As a matter of fact, if we compare Figure 3 with Figure 6, a number of common characteristics can be distinguished.

From comparison of Figures 3 and 6, some similarities and differences can be seen, from which we can get an idea of how the multi-year lognormal distributions are formed. These plots differ greatly in their form. We could hardly have expected it to be otherwise, since the second sample is a non-random sample taken from the first, and comprises less than 10% of its population (91 out of 1039 observations). However, a common feature is apparent in these graphs, reflecting the concentration distributions of lead in different areas. The type of deviation from the straight line clearly changes beyond the value 2.1 in both plots, which serves to indicate some common formation process, where the operating factors strongly affect the low concentration range, and their manifestations are common to all seasons and years of observations. The second consideration is that the logarithm of the upper concentration limit varies during the warm seasons around 3.9–4.0; this means

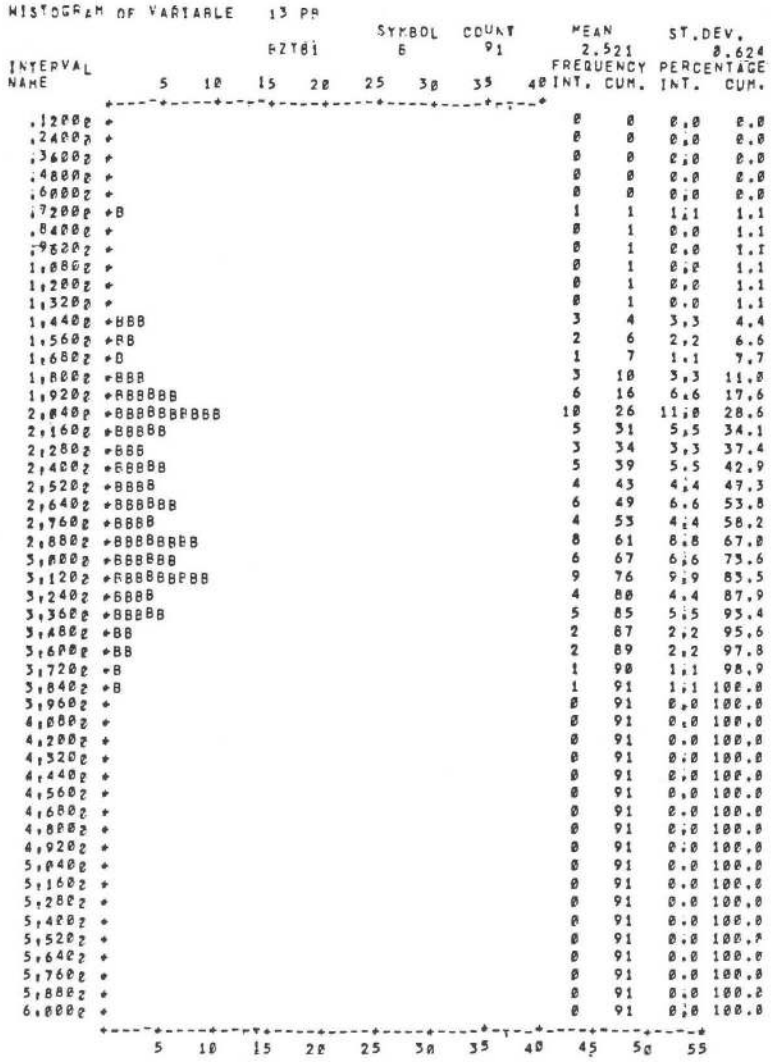

Fig. 4. Histogram of logarithmic concentrations of lead. Berezina, warm season, 1981.
1 – first mode; 2 – second mode.

that all the visible points in Figure 3 that lie beyond the boundary 3.9 reflect the influence of specific 'winter' factors, and from the form of the plot, it can be established that these effects do not coincide with the ordinary effects of the formation factors that we observe on the line-segment (2.1, 3.9). But then it becomes apparent that the lognormal distribution along this line-segment is a reflection of the effects of a very large number of formation factors, and the elimination of the effects of these factors results in manifestations of the operating mechanisms of other factors that are reflected in graphs of the (4) type as deviations from the straight line. That is, the subject-matter under study should be concerned not so much with the search of agreement between the observed

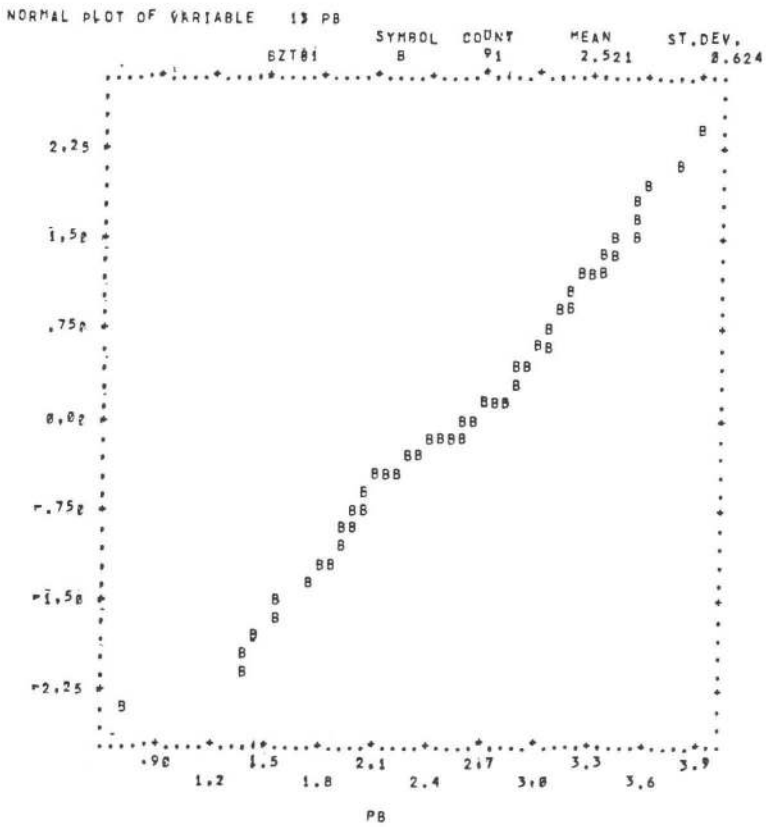

Fig. 5. Normal plot of logarithmic concentrations of lead. Berezina, warm season, 1981.

distribution and the LN2, as with the search for deviations from the normal and plausible explanations of their cause.

If similar plots for sulfur dioxide concentrations are compared, very great differences are perceived in the mechanisms governing the formation of the concentrations during the warm and cold seasons. The average values for the logarithms of the concentrations differ considerably – -0.29 for the warm seasons and 1.76 for the cold seasons. Differences are likewise reflected in the respective plots – most of the winter concentrations are located above a very small inteval (-0.5, 0.0) and about half of the concentrations for the warm season lie beneath it. The 'warm' concentrations terminate at about the logarithmic value 1.8, whereas most of the 'winter' concentrations lie within this range. That is, on the multi-year plot, zones can be distinguished that reflect the effects of cold and warm seasons. Even such a cursory examination makes it apparent that in order to determine natural background concentrations, it is necessary at least to get rid of the effects associated with the cold seasons, that are clearly contingent upon anthropogenic effects of the heating season.

The data series derived from observations on particulate matter display a similar

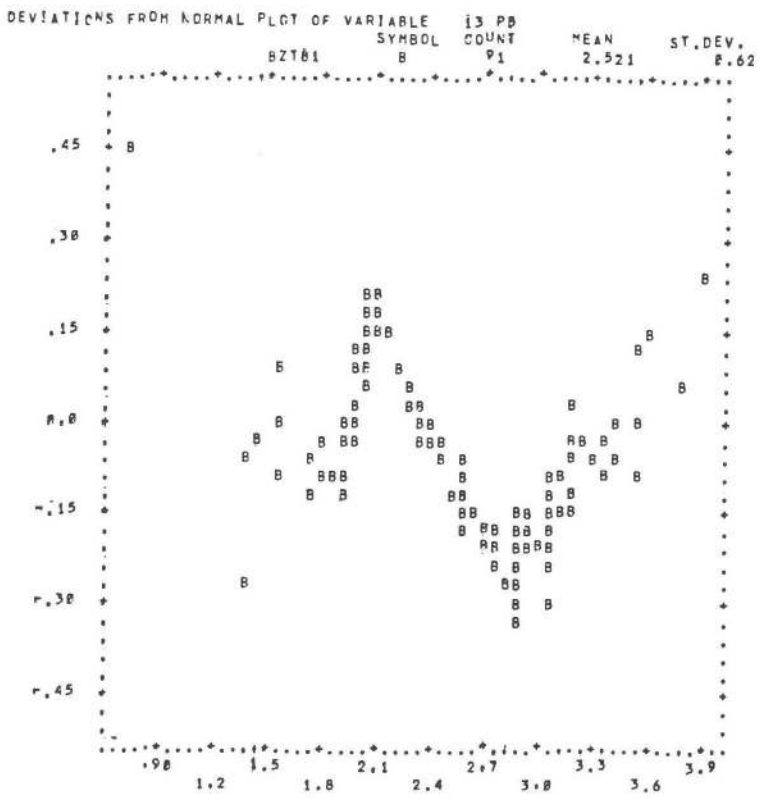

Fig. 6. Deviations from normal plot of logarithmic concentrations of lead. Berezina, warm season, 1981.

distribution pattern of the constituents, with a breakpoint on the curve, that depends upon the season.

Thus we conclude that it is necessary to design a statistical model that would enable the observed effects of various groups of factors to be taken into account, and to obtain quantitative estimates on the basis of statistical characteristics. Since the desired statistical model should describe the effects of different groups of factors, we are confronted with the problem of how to distinguish some typical samples from among the data. These samples should reflect quite fully the effects of different groups of factors and, at the same time, the regularities derived on their basis should be typical of a specific pollutant and area of observation. In order to determine such sampling characteristics, several hundreds of plots were analyzed, which show the logarithmic concentration distributions for periods between a decade to eight years. As a typical example, a data series of five-month duration was chosen, that characterized the warm or cold seasons. The period from May to September, inclusive, is regarded as the warm season; the period from November to March of the next year is regarded as the cold season. such a time-interval is, on the one hand, sufficiently long to show the effects of the major groups of factors and, on the other hand, sufficiently distinct from other observation series. Evidence that the seasonal

observational series are actually the major carriers of information on the effects of different groups of factors is found in the fact that in contrast to all data series, these data series include the highest percentage (over 80%) of deviations from the 'pure' lognormal distribution. Examination of Figure 6 makes it immediately apparant that the fluctuations are 'organized' into three line-segments, where each can be interpreted as manifestations of the effects of a group of factors controlling the formation of pollutant concentrations. The respective histogram (Figure 4) clearly displays a bimodal density function. These modes can be considered as central tendencies for each group of controlling factors. This implies that the model simulating the characteristics sample, which we have adopted for the seasonal observational series, should reflect the effects of different groups of factors, treated in the form of 'composite' distributions.

These results were derived only on the basis of graphical analysis and data presentation. For such an analysis, the authors used the package of applied statistical programs BMDP. The techniques used for analyzing and processing the meteorological data have been described by Zelenuk (1984) and Zelenuk, Zubenko *et al.* (1984) and a model has been proposed that describes the event-data series derived from background air pollution observations.

3.2. Construction of a statistical model for background air pollution monitoring data

Studies of seasonal observational series enable one to establish the specific multimodality of the density distributions, and the presence of characteristic deviations from the theoretical distribution function of the probability of logarithmic concentrations related to manifestations of different groups of causative factors. As a natural consequence, three problems arise. The first concerns the design of the statistical model for characteristic samples, with model parameters selected to reflect sample specifics (type of pollutant, area of observation, season, and mainly, the nature of the effects of the causative factors). The second problem refers to the method of analysis of the model with a view to determining the causative factors and the development of techniques for estimating model parameters. The third problem involves the derivation of statistical inferences concerning the entire data population from the seasonal sample population.

The next two chapters are devoted to the second and third problems. Here we shall describe the statistical model simulating background air pollution concentrations.

The following composite multimodel distribution model is used:

$$f(x) = p_1(x) f_1(x) + p_2(x) f_2(x), \tag{5}$$

where f_1 and f_2 are the density distributions, and p_1 and p_2 are the frequencies of realization, respectively. In contrast to classical composites (see, for example, Aivazyan *et al.* 1983) the frequencies are treated as a function of x, in order to distinguish the effects form the separate operation of the models within different intervals of the logarithmic concentration axis.

Let us take p_1 and p_2 as convolute functions of the 'switching' action of the laws, and let us impose normal noise with a zero mean value and variance σ. If we take

$$H(x; a) = \begin{cases} 1, x \leq a \\ 0, x > a \end{cases}, \quad H_1(x; a) = H(x; a), \quad H_2(x; a) = 1 - H(x: a) \tag{6}$$

then we can write

$$p_{iH,N}(x) = H_i(x: a) * N(x; 0, \sigma).$$

It will be recalled that

$$N(x; a, \sigma) = \frac{1}{\sqrt{\pi}\,\sigma} \exp\left[-\frac{(a-x)^2}{2\sigma^2} \right], \quad \Phi(x; a, \sigma) = \int_{-\infty}^{x} N(t: a, \sigma)\, \mathrm{d}t .$$

Then

$$p_{iH,N}(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} H_i(y; a) \exp\left[-\frac{1}{2} \frac{(y-x)^2}{\sigma^2} \right] \mathrm{d}y$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(a-x)/\sigma} \exp -\frac{z^2}{2}\, \mathrm{d}z = \Phi\left(\frac{a-x}{\sigma} \right). \tag{7}$$

By introducing the constants π_1 and π_2 characterizing the frequency of $x \leq a$ and $x > a$, respectively, we get

$$f(x) = \pi_1 \Phi\left(\frac{a-x}{\sigma} \right) f_1(x) + \pi_2 \Phi\left(\frac{x-a}{\sigma} \right) f_2(x). \tag{8}$$

It is obvious that π_1 and π_2 should be connected through normalization,

$$\int_{-\infty}^{\infty} f(x)\, \mathrm{d}x = 1.$$

In the case under discussion, when $f_k = N(x; b_k, s_k)$ then for $k = 1.2$ we get

$$\int_{-\infty}^{\infty} \left[\pi_1 \Phi\left(\frac{a-x}{\sigma} \right) N(x; b_1, s_1) + \pi_2 \Phi\left(\frac{x-a}{\sigma} \right) N(x; b_2, s_2) \right] \mathrm{d}x = 1 .$$

Using the identity

$$\int_{-\infty}^{\infty} \Phi\left(\frac{a-x}{\sigma} \right) N(x; b, s) = \Phi\left[\frac{a-b}{\sqrt{\sigma^2+s_1^2}} \right],$$

we derive the normalization conditions for two distributions

$$\pi_1 \Phi\left(\frac{a-b_1}{\sqrt{\sigma^2+s_1^2}} \right) + \pi_2 \Phi\left(\frac{b_2-a}{\sqrt{\sigma^2+s_2^2}} \right) = 1 .$$

The proposed method of compositing is derived from a qualitative analysis of the mechanisms governing the seperate distribution models. The 'switching' mechanism is

presumed to operate in such a manner that the zone of concentration grouping, defined by one of the formation mechanisms, should include more of its 'own' concentrations than those of a neighbor, which is possibly an even stronger mechanisms (i.e., including a larger number of observations). In practice, upon examination of the plots of density and distribution functions, the major zones of central tendencies are distinguished under the assumption that each one reflects certain regularities in the behavior of formation factors. In those cases, when the mutual influence of groups of factors is very great, their action can be regarded as a single formation mechanism. Such an approach to the analysis of the distribution, as was shown in the preceding section, is justified above all by the fact the composite action of different groups of seasonal factors forms a distribution closely similar to the LN2. Thus, each component of the distribution $f_1(x)$ may be regarded as the result of the operation of relatively independent causative factors.

Then, the parameters a and σ acquire a meaningful interpretation. From the attributes of the function it follows that

$$f(x) = \begin{cases} f_1(x), & x \leq a - 3 \\ f_2(x), & x > a + 3. \end{cases} \tag{9}$$

That is, the parameter 'a' plays the role of a 'switching point' for each of the models f_1. Parameter σ defines the extent of the 'switching zone' where composite factors do not reveal sufficient manifestations to form individual concentration groups. When $\sigma \to \infty$, the composition under study shows a tendency towards the well-known type of compositing.

This statistical model can represent data series of air-pollutant concentrations. These samples can be described by a set of quantities

$$\left\{ a_{i-1},\ \sigma_{i-1},\ \left\{ b_i,\ S_i,\ \pi_i \right\},\ a_i,\ \sigma_i \right\}_{i=1}^{k}. \tag{10}$$

Each set enables one to reproduce the information that we derive from examination of the plots, representing empirical density and distribution functions. The value π can be substituted by the percent of observations within the respective distribution, as related to the total number of observations.

4. Assessment of Background Air-Pollution Model Parameters

The problem of assessing model parameters discussed in the former section, is closely associated with the possibilities of non-statistical meaningful interpretation of the model. Such an interpretation, defining the possible use of model parameter estimates, may allow one to formulate certain requirements for the estimation itself. At any rate, it should be useful in ascertaining what types of statistics proposed by the model are consistent with our knowledge of the processes studied, leading to the solution of practical problems of evaluation of background air-pollution.

In de Nevers *et al.* (1979), a suggestion is made that for the description of air-pollutant distributions, 'composite' distributions should be used, ones that on lognormal paper look

like a graphical combination of two LN2 distributions. de Nevers *et al.* (1979) offer an example illustrating how two quite different types of meteorological conditions in the area of observation (with air-mass transfer across a canyon to the area of observation during particular time periods) lead to the formation of the distribution.

From the analysis presented in Zelenuk and Cherkhanov (1985), certain common features in the behavior of different substances in different areas can be perceived. There is a considerable coincidence – between 60 to 90% – on days when the sulfur dioxide and lead contents exceed the 'average' level (concentrations exceeding the 'average' level were graphically determined). These pollutants differ considerably, according to their origin, input rates, transport and dispersion, from which the inference can be drawn that the factors that control the formation of these two major components are characteristic of large-scale meteorological processes. These are associated with two major groups of factors. The first, which causes high pollution concentration levels, is related to the stable anticyclonic state of the atmosphere, when the concentrations of pollutants are determined mainly by diffusion. The second group of factors is associated with the unstable cyclonic state that is favorable for the dispersion of pollutants by turbulent processes, and, accordingly, with low concentration levels. Analysis of these relationships revealed that the coefficients of cross-correlation amount to 65-75%. Thus graphical estimation of the components of composite distributions, specified by Equations (2) and (4), enables one to interpret components of the background air-pollution process. In order to show this, cluster analysis of the composite distribution was performed with a view to ascertaining whether the two major components comprise 'natural' clusters in the event-data. As was demonstrated in Perone *et al.* (1975), the application of such techniques to solving problems bearing on air-pollutant concentrations, is permissible even when there is a weak assumption of homogeneity and uniformity of measurement scales. As in Perone *et al.* (1975), the present authorsd employed a cluster-analysis to distinguish the compositing components. They showed that an analysis of the sampled data series made it possible to establish the essential relationship between the graphically defined compositing components and the general characteristics of the meteorological processes.

The problem of obtaining estimates for the parameters of the statistical model of background air-pollution frequency distributions incurs difficulties due to the impossibility of verifying the estimates. As a matter of fact, it is not possible to evaluate any of the model parameters, using values obtained from direct measurements or reproduced from physical considerations; from the data series description proposed by model (10), we cannot choose the most informative set of parameters. In order to test the model function, to develop methods for graphical evaluation of the parameters, to elaborate concepts on the precision of such estimates and to use them, models simulating the observational data series were therefore designed.

To perform a simulation, it is necessary to establish what attributes of the model should be simulated and investigated. Taking into account the main problem of our investigation, the determination of statistics that provide a description of the time-series, ensuring comparison with observational data and the possibility of distinguishing several components in the simulated composite distribution, we shall restrict our attention to the

composite performance of two distributions, for which we shall attempt to find an answer to the questions: what statistics fit the description of different composite distributions, and what is the accuracy of such a description?

Figures 7, 8, and 9 respectively depict a histogram, a distribution function plotted on normal probability paper, and the same function with the linear trend removed, for a composite distribution, conforming to normal laws, with parameters $N(x; 25, 1)$ and $N(x; 27, 1)$ respectively. The number of observations in the first case is 35 and in the is -85. (Henceforth, we shall speak of composite distributions 35N/25,1/ and 85N/27,1/). For evaluation and graphical analysis, the same data-presentation is used as for ordinary observation series. Simulation is performed employing the normal distribution.

The plots presented in Figures 7, 8, and 9 enable us to demonstrate the techniques for assessment of the composite distribution parameters. At first, the major components are singled out in Figure 9, they lie, obviously, on opposite sides of value 26.5. Then the intervals for the grouping of both composite components can be pointed out: intervals [21.9, 26.5] and [26.5, 28.95]. Within the intervals thus distinguished, we can find the respective central tendencies. In determining these, the influence of the 'transitional process' should be considered, which consists in the existence of observations of different

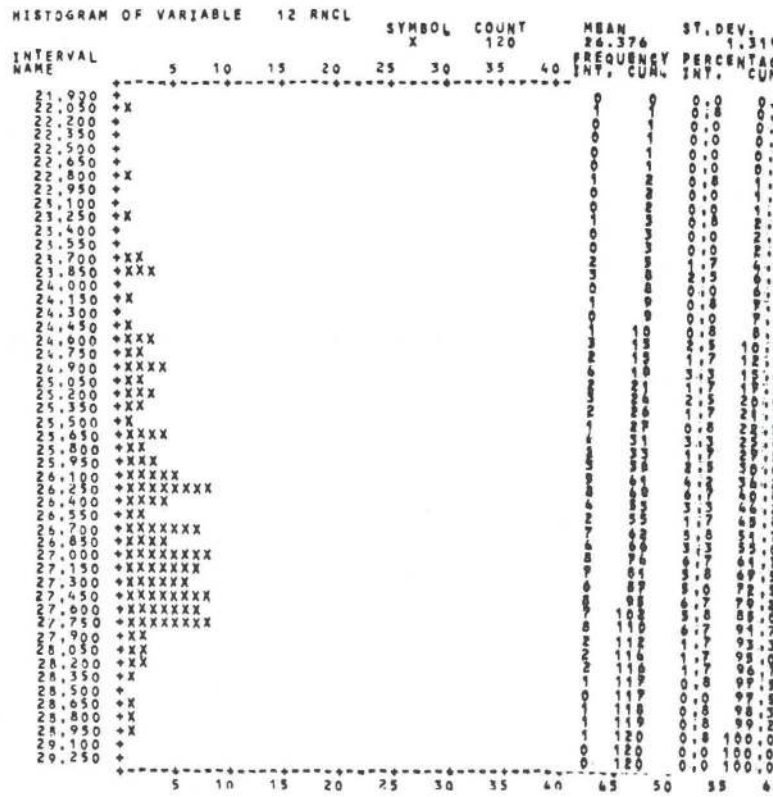

Fig. 7. Histogram of composite distributions 85N/27,1/ and 35N/25,1/ (simulation).

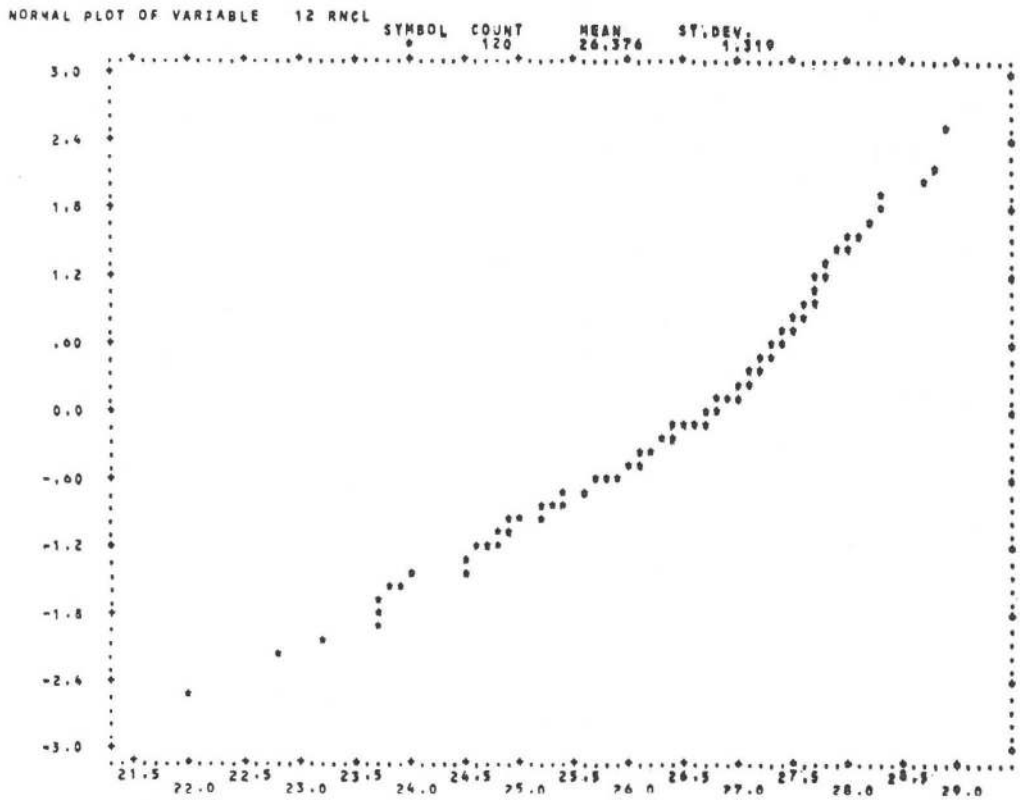

Fig. 8. Normal plot of composite distributions 85N/27,1/ and 35N/25,1/ (simulation).

components near the 'switching point' 26.5, i.e., it is necessary to exclude from consideration the region of the 'switching point'; for practical purposes it is sufficient to exclude three intervals to the right and to the left of it. In the remaining intervals, the points located at the centers of the intervals give the maximum total contribution to the set of obserations. This contribution is estimated from the histogram in Figure 7, where the respective intervals are denoted by stars. In practical estimates, three successive intervals are considered, choosing that triplet that provides the maximum contribution; the right boundary of the middle interval of the triplet is taken as the estimate of the central tendency. The estimates for the composite distribution, derived in this manner, amount to 24.9 and 27.45, the true values being 25.0 and 27.0, respectively. This method of estimation underwent multiple evaluation in regard to different kinds of composites, and the estimates for the central tendencies, derived from them, differ by no more than 10% from the known centers. Estimates of the variance were not performed: as such an estimate for the switching points, we adopted the length of $\pm 3\sigma$. Using the histogram, it is possible also to evaluate the relative weight-percent of the composite components. This is determined as the percent of observation points lying on opposite sides of the 'switching point'. This etimate is less accurate, revealing in our case proportions of 44 and 56% for the left and

Fig. 9. Deviations from normal plot of composite distributions 85N/27,1/ and 35N/25,1/ (simulation).

right sides of the distribution, respectively. Nevertheless, it enables both sides of the distribution to be evaluated qualitatively.

Thus, returning to model (10)

$$\{a_{i-1},\ \sigma_{i-1},\ (b_i,\ S_i,\ \pi_i),\ a_i,\ \sigma_{i=1}^{k}\}$$

we can represent the distribution under study as

$$\{23.0,\ 0.15,\ (24.9,\ 44\%),\ 26.55,\ 0.15,\ (27.45,\ 56\%),\ 28.95,\ 0.15\}.$$

As is evident from comparison of Figures 7, 8, 9, this description precisely reflects the observed distribution pattern. The present author analyzed over one hundred plots of model functions describing the composite distributions. In general, the estimation techniques and results are quite satisfactory and can be used for estimates of the model parameters and for the logical design of the model. However, a question of great interest arises: which of the parameters are of the greatest statistical stability in relation to changes in the formation of the composition? In order to answer this question, a further analysis was performed, varying the compositing parameters. Let us examine some of the results.

Figures 10, 11, and 12 present plots describing the distribution of the same components, i.e., of the composite distribution $N(x; 25.1)$ and $N(x; 27.1)$. These distributions differ only

in the amount of implemented components: 35–85 in the first case already discussed, and 55–70 in the second. This example is typical from the point of view of the extent to which it is possible to rely on the configuration of the distribution pattern in any hypotheses offered on the nature of the data. It is obvious that with real concentration distributions, accepting the hypothesis validated above that two major types of meteorological factors influence the formation of the composites, situations may occur when days with different meteorological conditions exhibit ratios of 35/85, as well as 55/70. Notwithstanding the differences in these plots, they obviously represent similar processes, which sould be reflected in the respective estimates. What estimates for the second group of plots can serve to disclose the similarity with the first group?

In Figure 12 three parts of the composite distribution can be clearly distinguished. These are represented by the intervals [22.50, 25.05], [25.20, 27.15], [27.15, 28.95]. As a matter of fact, we are dealing with three distributions, defining the composite components; the middle distribution characterizes the combined influence of factors under which the unmixed effects of two extreme components of the distribution were formed. It is obvious that the newly derived estimates for the central tendencies, and for the weights of the components, bear no resemblance to the respective parameters of the first distribution depicted in Figures 3, 5. The fact that they describe processes that have some similarities can be established only upon comparison of the intervals. The similarity is obvious between intervals [21.9, 26.55] and [22.50, 25.05], also between intervals [26.55, 28.95] and [27.15, 28.95]. This example is a reflection of the regularity repeatedly observed by the authors: when the parameters are subject to variations, then of greatest stability are the interval estimates of the composite components. Taking into account the fact that for our purposes they quite satisfactorily describe the distribution, they may be adopted as the major statistics characterizing the component distributions.

The statistics that we have selected in the form of estimates of the grouping intervals can, most likely, reveal the major regularities. However, the question remains: how does one obtain estimates of such grouping intervals. The application of computer-programmed discrimination analysis is formulated in Buchstaber *et al.* (1983) and the possibility of introducing *a priori* information into the classification algorithms is discussed. The present authors, participating in the development of the algorithms and programs for the work (Buchstaber *et al.* 1983) performed a large number of experimental computations on the application of different algorithms. Experience gained from this work (Buchstaber *et al.* 1983), tends to indicate that each of the studied problems is independent and intricate, to which the application of monotypic discrimination methods is impossible.

This calls for visual, graphical, discrimination of the compositing components. Such a method, by the way, is widely used in routine practice, when the data can be represented in the form of two-dimensional patterns.

Figures 10, 11 and 12 present plots of composite distributions that cannot be unambiguously classified. For instance, the distribution can be represented as the sum of three components, specified by the intervals [5, 15], [17, 30], [27, 37], and specified by two intervals [5, 17] and [17, 37]. Notwithstanding the outward dissimilarity of these

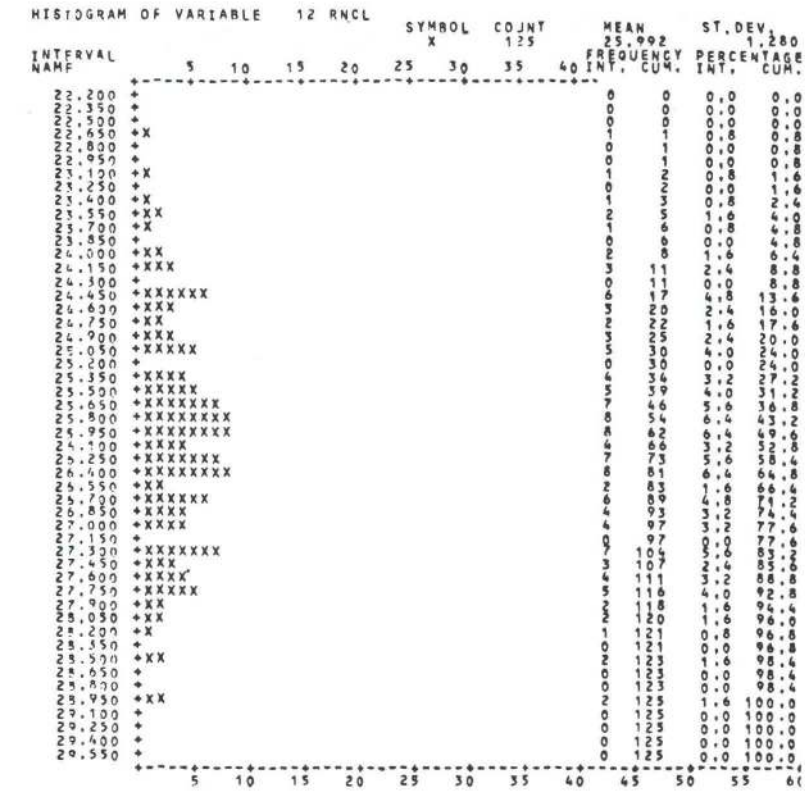

Fig. 10. Histogram of mixed distributions 70N/27,1/ and 55N/25,1/ (modeling).

classifications, not less than 50% of the respective intervals reveal a trajectory cross-over, which enables one to judge the similarity in the studied distributions represented by such intervals.

Thus, it can be considered an established fact that the most statistically stable distribution, in the sense of the preservation of the major characteristics of different components of the composite distribution, represented by the statistical model of background air-pollution herein proposed, are the intervals of the observational central tendencies. Evaluation of the intervals is performed graphically, which ensures quite a reliable distinguishing of the effects caused by the components, this process being supplemented with the percentage of the number of observations falling within the band. That is, analysis of the simulation data and of the possibilities of graphical evaluation and interpretation of the parameters, makes it possible to reduce the informative description of the time-series from (10) to

$$\{a_{i-1}, (\mathrm{II}_i\%), a_i\}_{i=1}^{k} . \tag{11}$$

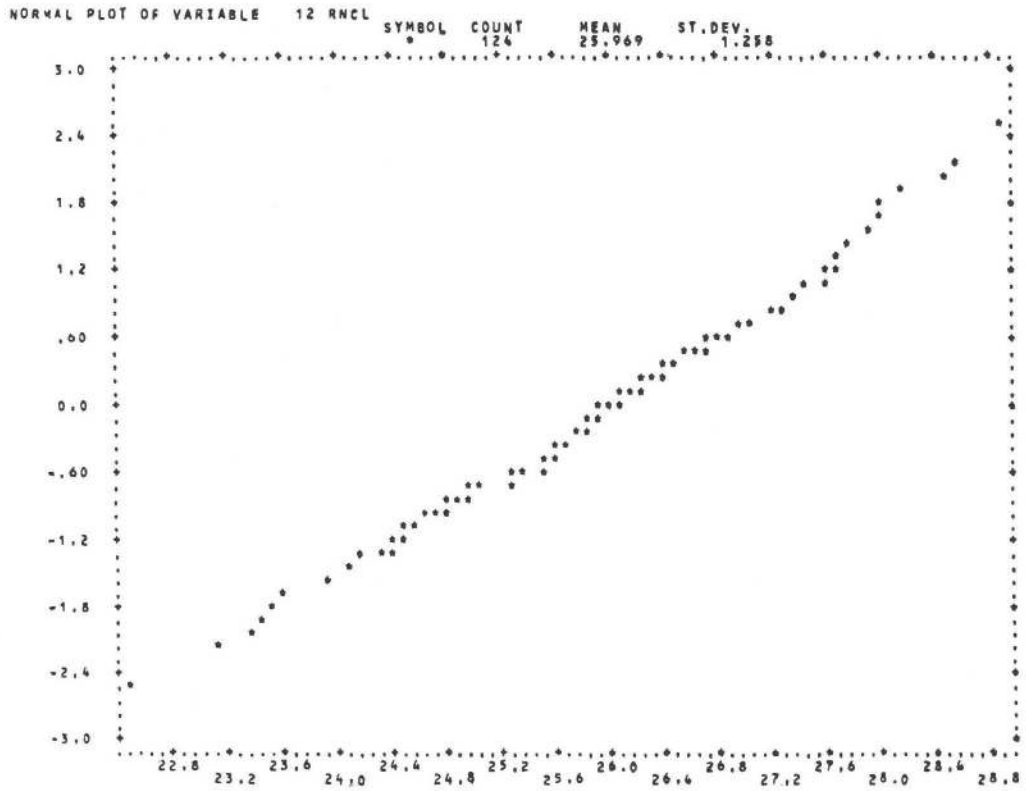

Fig. 11. Normal plot of mixed distributions 70N/27,1/ and 55N/25,1/ (modeling).

5. Discrimination of the Components of Background Air Pollution

5.1. Estimation of the Central Tendencies of Multi-Modal Frequency Distributions of Seasonal Data Series

The proposed statistical model of background air pollution was used in Izrael *et al.* (1985) to estimate the components of an observational series, sampled at the 'Borovoe' background monitoring station. Discrimination of the components was performed by using a simplified model that assumes that the data series is the result of two major processes that affect the concentration frequency distribution. It is clear that estimates of the central tendencies in the lower part of the distributions can provide meaningful inferences concerning the model and the background concentration levels themselves. With respect to the model, it can be assumed that in the case of agreement between the graphically defined distribution characteristics and the processes occurring in the atmosphere, on the one hand, and the established concepts on the nature of these processes, on the other hand, it would be possible to predict the behavior of the estimates of the central tendencies in the lower parts of the distribution. The hypothesis can be offered that such estimates should possess essentially greater statistical stability, as

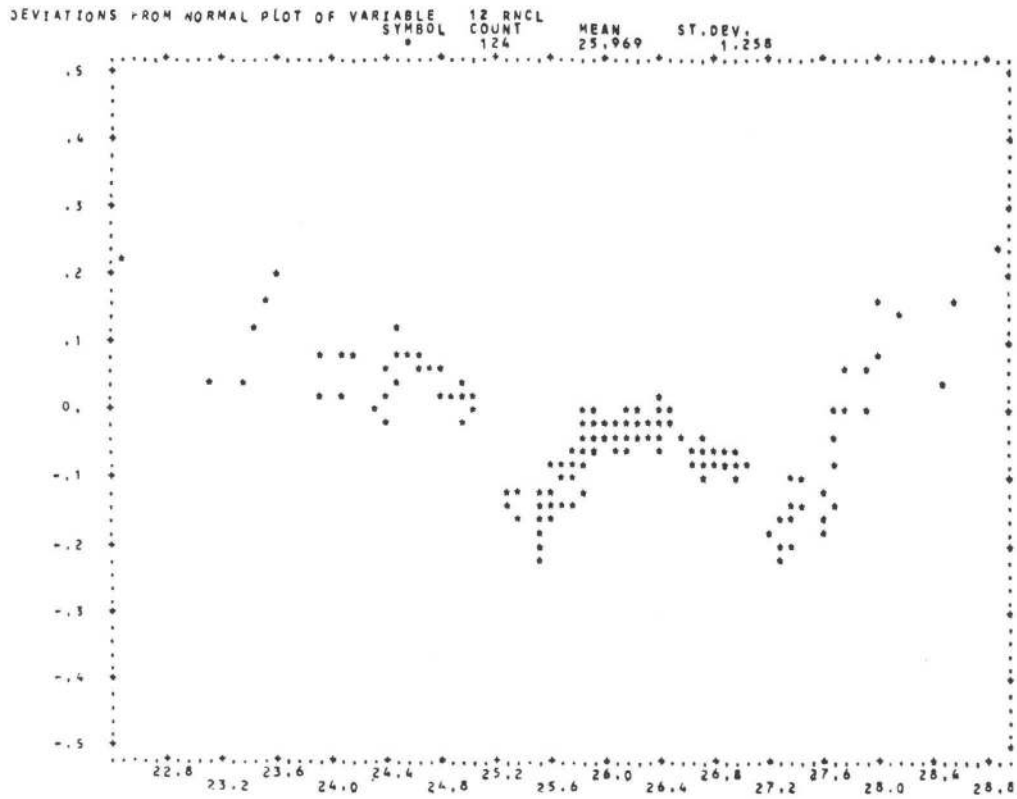

Fig. 12. Deviations from normal plot of distributions 70N/27,1/ and 55N/25,1/ (modeling).

compared to estimates for the upper parts of the distributions, considering that the physical processes forming the lower part lead to averaging of the concentration levels over the entire lower troposphere, thus 'smoothing' the pollution concentrations, as a consequence of which the processes are characteristically defined as being of a greater scale and, naturally, of greater statistical stability. Evidence that lends support to this hypothesis is found in Izrael *et al.* (1985). The plots in Figure 13 demonstrate the difference in the behavior of two major components. This implies that the application of the proposed model to background air pollution concentrations, and the techniques used to derive estimates for the model parameters is fully justified.

The statistical stability of the estimates of the central tendency of low concentration frequently suggests its possible use for assessment of background concentration levels of air-pollutants. As a matter of fact, upon comparison with ordinarily employed mean values, it becomes apparent that the derived estimates have a number of advantages. They enable one to validate statistically, and to calculate the characteristics of, regional background levels as an average over minimal concentration values for a given time period (Rovinskii *et al.*, 1982).

Zelenuk and Cherkhanov (1985) discuss the possible use of the estimates for the central

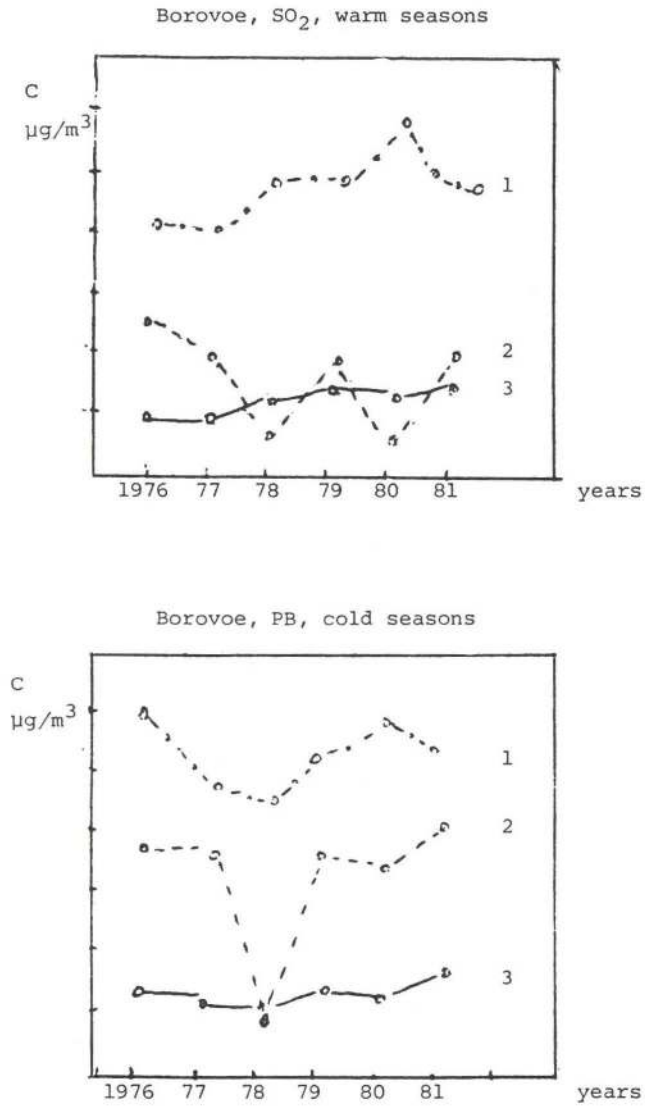

Fig. 13. Results of estimation of two concentration central tendencies. (1) lower component; (2) upper component; (3) composite series.

tendencies, derived graphically, in solving problems related to the evaluation of background pollution concentration levels in the atmosphere. For instance, pollutant deposition should be calculated taking account of the entire set of factors controlling the formation of concentration levels, which can be feasibly performed by using the estimates of the central tendencies that are common to multiyear seasonal data series. It may be found necessary to make such estimates for harmful pollutants that cause deleterious effects on vegetation. Analysis of the upper parts of the distribution may be necessary for

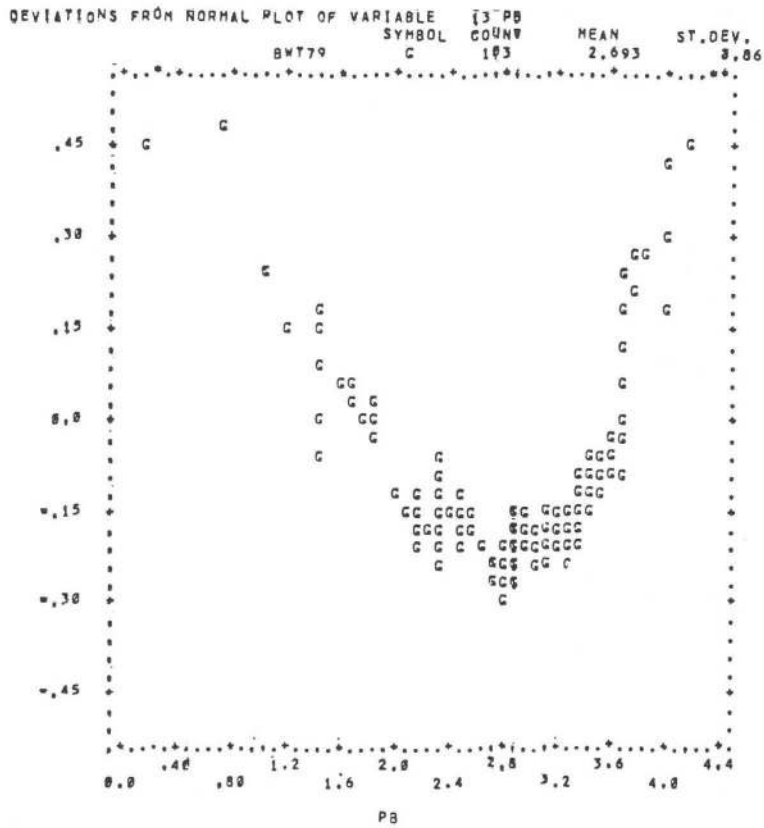

Fig. 14. Deviations from normal plot of logarithmic concentrations of lead. Borovoe, warm season, 1979.

estimates of the probable pollution level from anthropogenic effects in impact areas. In problems connected with short-time variations of background concentrations levels or, to be more exact, of the concentration level of pollutants in impact areas, estimates can occur that characterize data series of one-two decades or one-two months duration. From parctice (Zelenuk and Cherhanov, 1985), it is apparent that use of such data series for discrimination of components is difficult due to the fact that the formation mechanisms are not sufficiently understood. A further serious handicap is the inadequacy of the results of averaging meteorological characteristics over such short time-periods, which leads to strong scattering of the derived estimates and makes it impossible to compare estimates for different periods. The use of observational series for a period of more than a year does not allow one to obtain the necessary estimates, namely due to the mixing of several distributions. This is in good agreement with the statistical design of the investigations, from which it can be inferred (see section 3.2) that data-series of less than one season and more than one-year duration demonstrate much less deviation from the lognormal distribution than seasonal data-series. In the first and second case, it can be assumed that the distributions are unimodal, if this is required for deriving estimates of the central tendencies. In the case of short-time series, these centers describe the average effects of

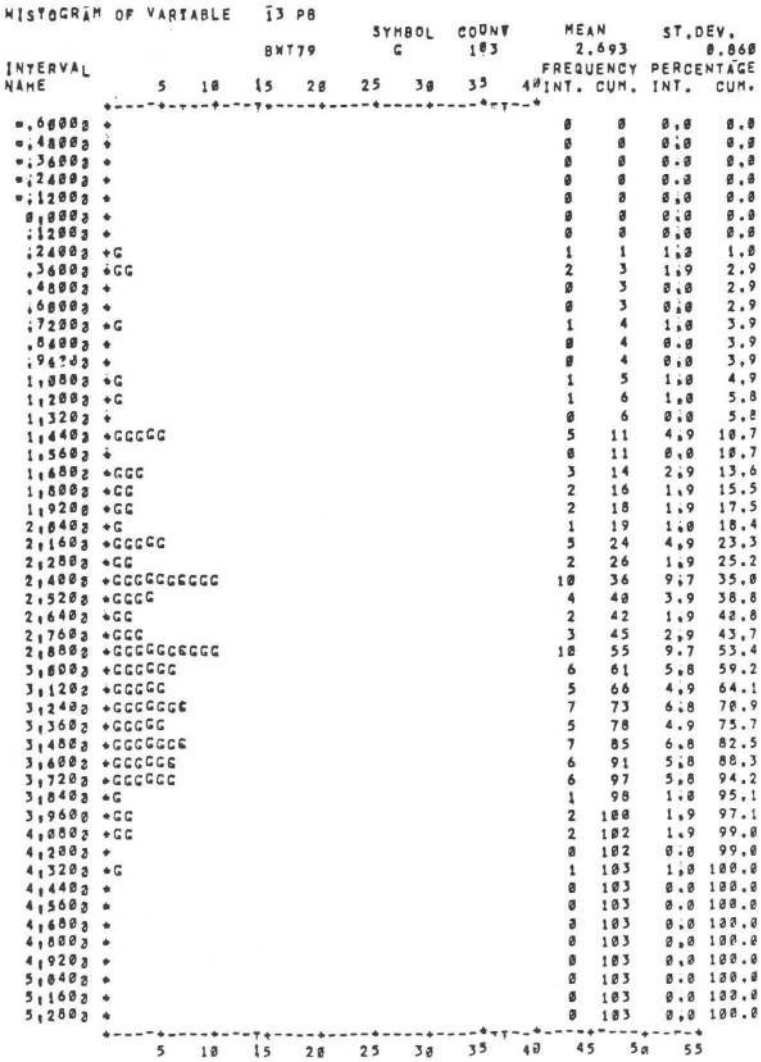

Fig. 15. Histogram of logarithmic concentrations of lead. Borovoe, warm season, 1979.

different types of pollution distributions, typical of this short period and small areas of observation. For data series of one-year duration and more, such centers describe the average effects of all the probable pollution formation mechanisms operating within a sufficiently large region under different synoptic conditions.

5.2. Estimates of Selective Grouping Intervals

A method for estimating the grouping intervals in time-series was discussed in the previous chapter. Now let us see whether this method can be used to discriminate the components in seasonal data series.

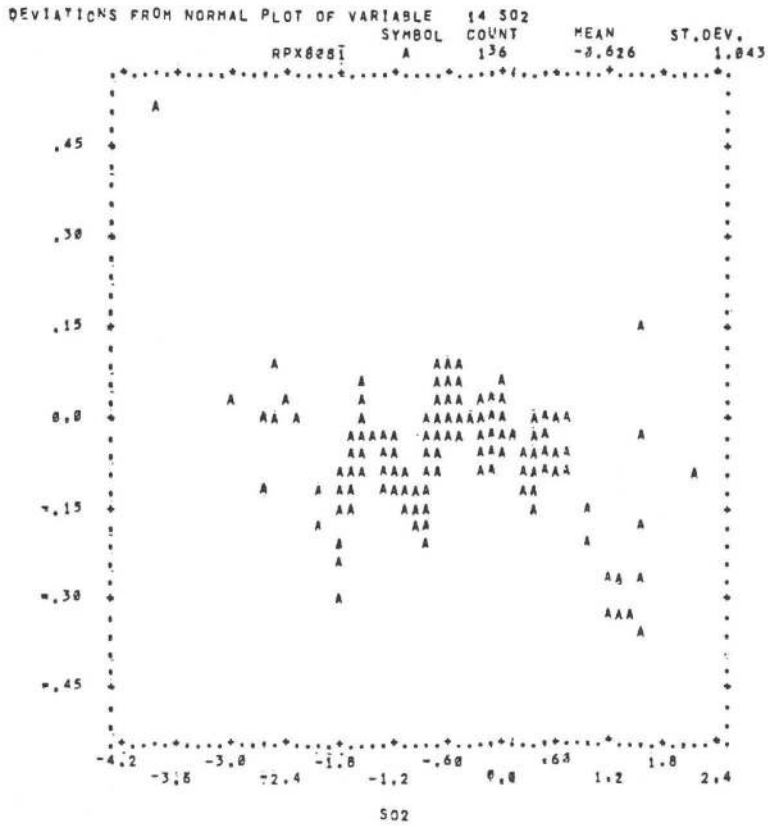

Fig. 16. Deviations from normal plot of logarithmic concentrations of sulfur dioxide. Repetek, cold season, 1980-1981.

Figures 14–22 present several characteristic data series, derived for different pollutants during different seasons and at different stations. As is apparent from the graphs, the components of background air pollution can be reflected in various types of deviations from the normal pattern. Analysis of such deviations does not enable one to judge which types of deviations are more typical for a particular station, pollutant or season. At the same time, the presence of distinctly pronounced components in most of the graphs, again confirms the proposed model. In order to obtain evidence demonstrating the general nature of the regularities governing the formation of concentrations, we included data available on measurements of sulfur dioxide concentrations, that were performed in accordance with the international program on long-range transport of air pollutants. The data were taken from observational series obtained at stations located in the impact areas of Jergul, Norway and Abisko, Sweden. These data, differing according to their sampling and analytical techniques (which in practical work led to a need to recalculate the values; for comparison with the rest of the data array, the logarithmic concentrations should be reduced by 1.6), demonstrate the same type of distribution pattern as the data from the 'Borovoe', 'Berezina B.R.', and 'Repetek B.R.' stations. From this, the inference can be

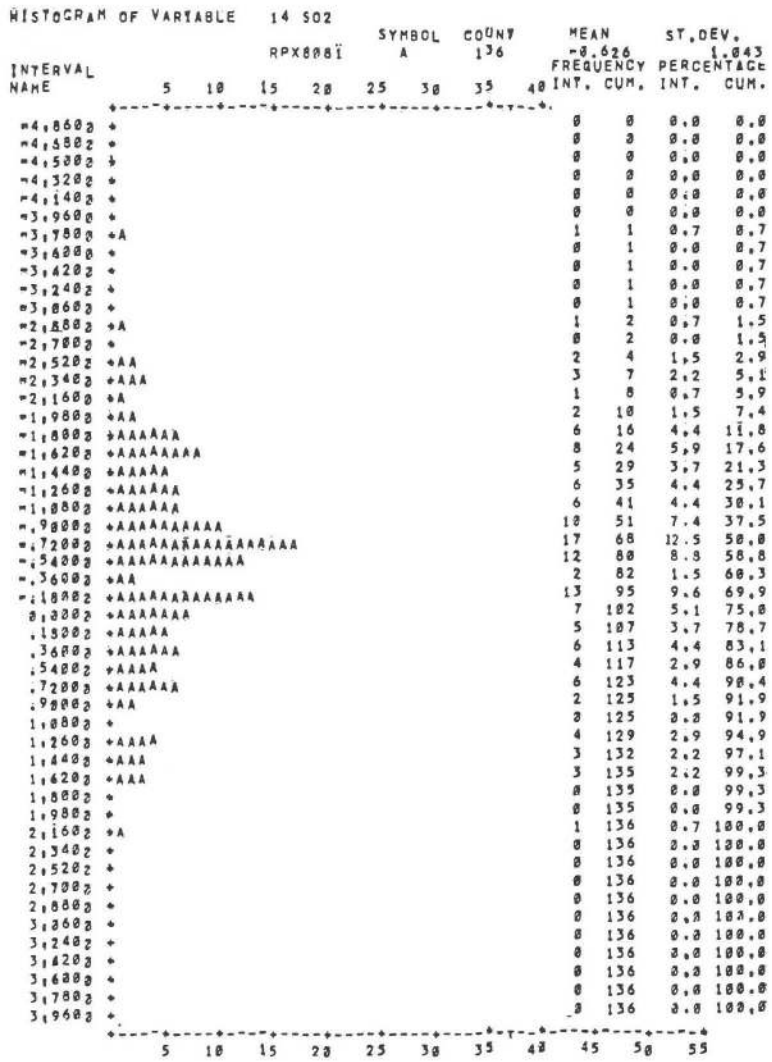

Fig. 17. Histogram of logarithmic concentrations of sulfur dioxide. Repetek, cold season, 1980-1981.

drawn that pollution concentration distributions are controlled by processes having common features, at least on the scale of continents. This implies that the statistical model enables one to describe observational data from background monitoring stations over entire continents. Such a description for each separately taken data series represents a set of intervals and corresponding weighted quantities, specified in the form of Equation 11.

The graphical estimate allows for informal interpretation of the compositing components, using two classes, 'well' and 'poorly' defined. As was shown in Section 3.3., such deviations do not influence essentially the principal 'recognition' of the existing components. Studies in which experts analyzed the plots and distinguished the grouping

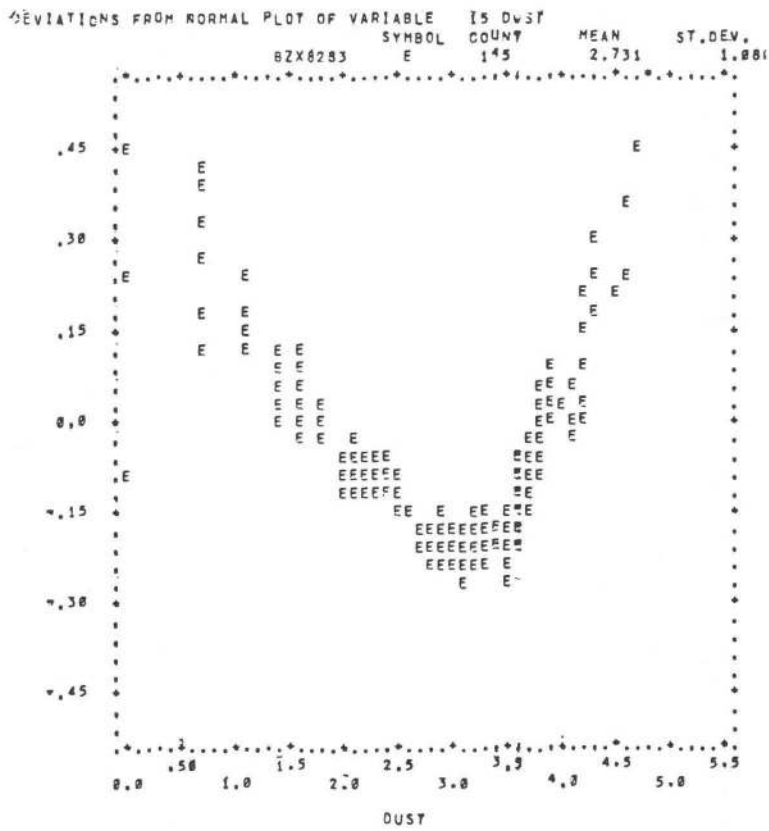

Fig. 18. Deviations from normal plot of logarithmic concentrations of suspended particulate matter. Berezina, cold season, 1982-1983.

intervals, served to indicate that the discrepancies in the estimates are small and, at any rate, much less than the probable variations in the estimates associated with the redistribution of the relative weighted quantities of the components, the case to be considered in Section 5.3. Estimates performed for each interval enable one to identify some rather general and statistically stable characteristics, so that we can judge the effects caused by different concentration formation mechanisms. The plots depicted in Figures 14–22 give a rather fair idea of the techniques used to distinguish the central tendencies on the basis of graphical discrimination.

Each of the series analyzed specified its own series of gruping intervals, that henceforth are termed the series of central tendencies. These intervals and series, tabulated in Tables I–XI, in included practically all the statistical information available for analysis. In order to derive general conclusions, characterizing seasonal data series that are large in time and space, certain new statistics should be designed that generalize the random central tendencies. The following section is devoted to the design philosophy of such statistics.

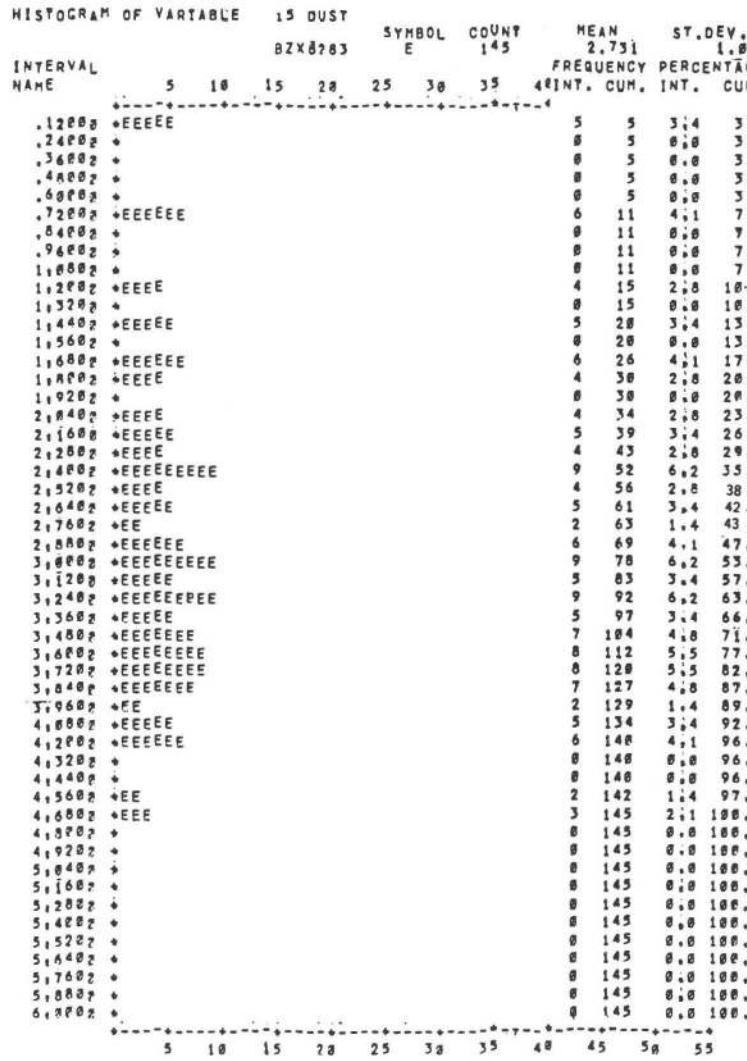

Fig. 19. Histogram of logarithmic concentrations of suspended particulate matter. Berezina, cold season, 1982-1983.

5.3. METHODS OF CONSTRUCTION OF STATISTICALLY STABLE ESTIMATES OF POLLUTION COMPONENTS

Each series specifies, in fact, the subdivision of the logarithmic concentration axis into intevals. Thereby, the search for statistically stable characteristics of the series proceeds within the region over which they are defined, i.e., over the range of all intervals of the logarithmic concentration axis, henceforth denoted as I. Each of Tables I–XI exemplifies a sample in space I. Let us denote the sample specified by the array of central tendencies as W. In order to compare the intervals and the subdivisions of the axis with each other, a

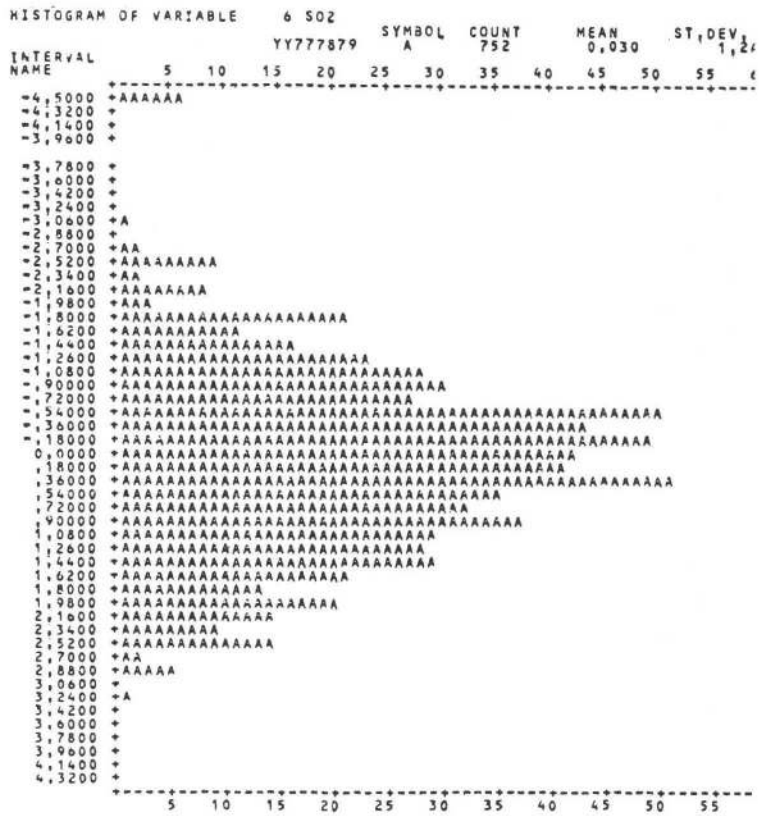

Fig. 20. Deviations from normal plot of logarithmic concentrations of sulfur dioxide. Jergul, Norway, cold season, 1981.

measure of similitude is needed. Let us consider the intervals $I_1 = [\alpha_1, \beta_1], I_2 = [\alpha_2, \beta_2]$. As a measure of their similitude, it is natural to take the quantity characterizing the cross-over of the intervals in relation to their sizes, i.e., the ratio of their cross-over to their unification:

$$\mu^1 = \frac{|I_1 \bigcap I_2|}{|I_1 \bigcap I_2|}, \tag{12}$$

where $|I|$ is the length of interval I.

Such a definition of the measure of similitude, although convenient, has two shortcomings: firstly, it is inconvenient for computations and, secondly, it performs well only for central tendencies of common ground points. The next proximity measure eliminates such shortcomings:

$$\mu(I_1,I_2) = \frac{\min(\beta_1,\beta_2) - \max(\alpha_1,\alpha_2)}{\min(\beta_1,\beta_2) - \min(\alpha_1,\alpha_2)} \tag{13}$$

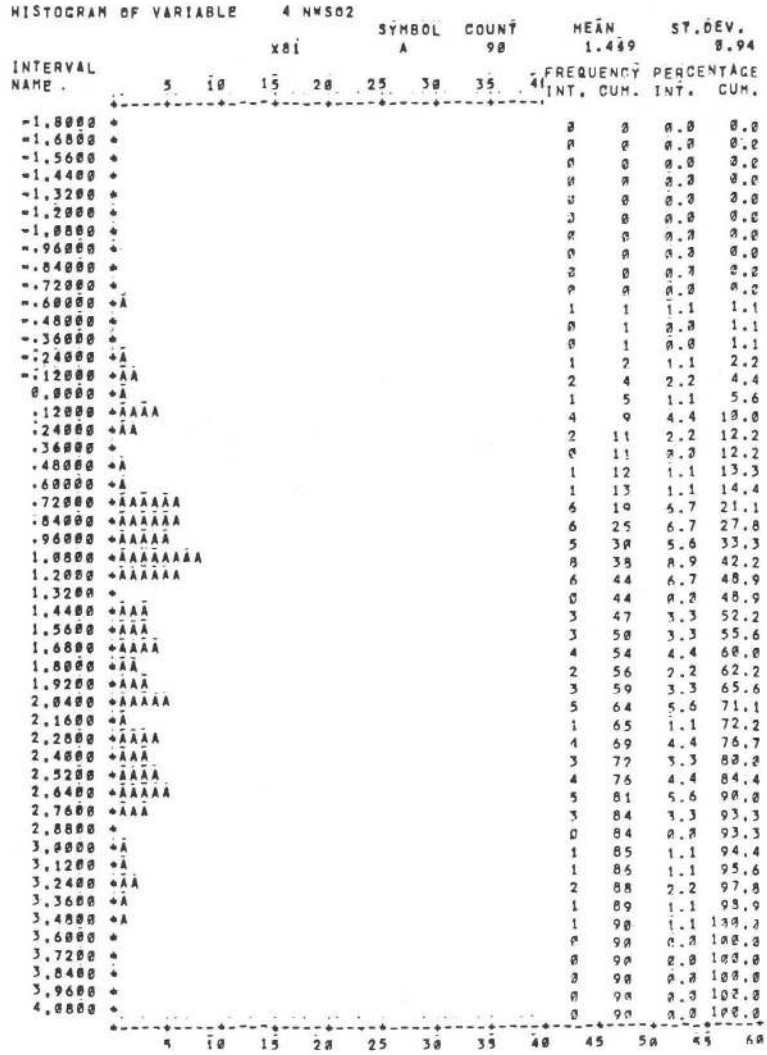

Fig. 21. Histogram of logarithmic concentrations of sulfur dioxide. Jergul, Norway, cold season, 1981.

The measure μ assumes the value $-1 < \mu \leq 1$. It can be easily seen that if $\mu > 0$, $\mu'(I_1, I_2) = \mu(I_1, I_2)$.

Now if we take a certain interval, for example I_0, we immediately obtain a set of numbers where each number corresponds to one of the intervals of the sample W, and characterizes the measure of its similitude to I_0. Having chosen a certain scalar measure for this set, for example, the average value for the measure of proximity, we get the functional number

$$F: I \to R^1 \tag{14}$$

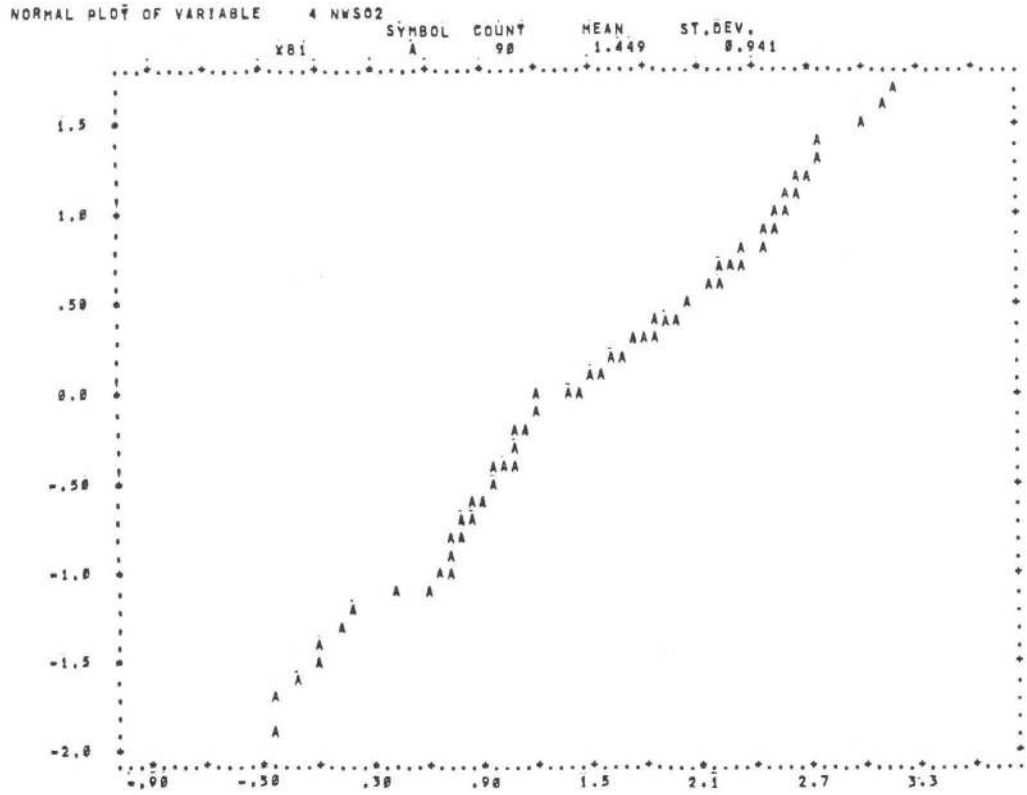

Fig. 22. Normal plot of logarithmic concentrations of sulfur dioxide. Jergul, Norway, cold season, 1981.

estimated over the sample. The fact that such a representation is possible reflects the objective of the investigation, the search for stable statistics to describe the sample.

From examination of the Tables, where random grouping intervals are represented, it is apparent that many intervals are similar from one season to another. The postulate can therefore be offered that within the entire data series, a certain relationship exists which is differently manifested during different seasons. For a description of this relationship, we shall use the method of perinterval estimates: if the relationship can be defined by the intervals, then it itself can be described by the interval that represents the integrated effects of the grouping factors. Such an intergral we shall term the statistical grouping interval. Now we can state the problem concerning the algorithm for discrimination of a statistically stable grouping interval. Using μ, let us synthesize an algorithm for estimation over intervals of sample W of the measure as to how much the given interval I_0 can be regarded as a statistical concentration grouping interval:

(1) Let us specify I_0 and $\mu_0 \in (0,1)$.

(2) Let us design the section $S(I_0)$ in the form of a set of intervals, ordered in accordance with the series to which they belong, and for which the measure of similarity with I_0 exceeds μ_0.

TABLE I

Random grouping intervals of logarithmic concentrations of sulfur dioxide. Borovoe, 1976-1983. (T) warm season, 1978; (X) cold season, 1979-1980, (%) percentage of points falling within the interval

T76	%	X7677	%	T77	%	X7778	%	T78	%
-2.0, -0.7	18			-3.0, -1.8	13	-2.4, -0.4	16		
				-1.8, -1.2	30				
-0.7, +0.1	17	-1.6, 0.2	26	-1.2, 0.2	57			-1.4, -0.1	40
0.1, 0.8	60	0.2, 1.0	41			-0.4, 1.0	32	-0.1, 0.7	52
0.8, 1.5	5							0.7, 1.2	8
		1.0, 2.5	33			1.0, 1.6	22		
						1.6, 2.6	30		

X7879	%	T79	%	X7980	%	T80	%	X8182	%
		-2.0, -1.2	10			-2.2, -1.2	30		
		-1.2, -0.8	30			-1.2, -0.2	20		
		-0.8, -0.4	37						
0.0, 1.1	32	-0.4, 0.0	13	-0.9, 0.5	14	-0.2, 0.5	15	-1.0, 0.5	17
		0.0, 0.7	10	0.5, 1.2	16	0.5, 1.5	34	0.5, 1.9	42
1.1, 1.4	23			1.2, 3.3	70			1.9, 2.7	18
1.4, 2.2	29								
2.2, 2.9	15								
								2.7, 4.5	13

T81	%	X8182	%	T82	%	X8283	%	T83	%
								-2.0, -0.4	14
-1.5, -0.6	12			-1.4, -0.2	22				
-0.6, 0.2	25			-0.2, 1.0	72	-0.5, 1.1	12	-0.4, 1.6	86
0.2, 1.2	46			1.0, 1.8	6				
1.2, 2.2	17	1.5, 2.1	15			1.1, 2.6	70		
		2.1, 3.1	46						
		3.1, 3.9	39			2.6, 4.0	18		

X8384	%
-0.5, 1.7	52
1.7, 3.2	48

TABLE II

Random grouping intervals of logarithmic concentrations of sulfur dioxide. Berezina B.Z., 1980-1983. (T) warm season, 1981; (X) cold season, 1981-1982. (%) percentage of points falling within the interval

X8081	%	T81	%	X8182	%	T82	%	X8283	%
		−1.5, −1.5	8			−1.4, −0.3	14		
−0.5, 0.6	10	−0.5, 0.3	42	0.3, 1.2	5	0.1, 1.0	63	0.0, 0.8	7
0.7, 1.5	20	0.3, 1.5	37			1.0, 1.6	15	0.8, 1.5	17
1.5, 2.1	14	1.5, 2.5	13	1.2, 2.8	50	1.6, 2.0	5	1.5, 2.5	53
2.1, 2.7	37								
2.8, 3.3	10			3.0, 4.0	45			2.5, 3.8	23
3.3, 4.5	10								

T83	%	X8384	%
−1.7, −0.2	20		
−0.2, 1.0	68	0.2, 1.3	16
1.0, 2.2	12	1.4, 1.8	32
		1.8, 2.8	25

TABLE III

Random grouping intervals of logarithmic concentrations of sulfur dioxide. Repetek B.Z., 1980-1983. (T) warm season, 1981; (X) cold season, 1981-1982. (%) percentage of points falling within the interval

X8081	%	T81	%	X8182	%	T82	%	X8283	%
−3.0, −2.4	6	−3.0, −2.3	18	−3.0, −2.5	2				
−2.0, −1.5	15	−2.3, −1.2	62	−2.3, −1.2	48	−2.3, −1.2	13	−2.2, −1.3	19
−1.5, −1.0	15	−1.2, −0.6	10	−1.2, 0.0	30	−1.2, 0.5	75	−1.2, 1.0	69
−1.0, −0.4	30								
−0.4, 1.0	25	−0.66, 0.9	10	0.2, 1.5	20	0.5, 1.0	12		
1.0, 2.0	10							1.0, 2.3	12

T83	%	X8384	%
−2.0, −1.3	4	−1.8, −0.8	23
−1.2, −0.5	24	−0.8, 0.1	43
−0.5, 0.4	47		
0.4, 1.0	10		
1.0, 2.3	10	0.1, 2.3	32

TABLE IV

Random grouping intervals of logarithmic concentrations of sulfur dioxide. Repetek B.Z., 1980-1983. (T) warm season, 1981; (X) cold season, 1981-1982. (%) percentage of points falling within the interval

X81	%	T81	%
-2.5, -1.5	12	-2.6, -1.6	8
-1.5, -0.4	36	-1.6, -1.1	14
		-1.1, -0.2	23
-0.4, 0.8	32	-0.2, 0.8	38
0.8, 1.2	13	0.8, 1.3	14
1.3, 2.0	7	1.3, 1.9	3

TABLE V

Random grouping intervals of logarithmic concentrations of sulfur dioxide. Abisko, Sweden, 1981. (T, X) warm and cold seasons, respectively. (%) percentage of points falling within the interval

X81	%	T81	%
-2.2, -1.6	7	-3.3, -2.5	5
		-2.5, -1.6	13
-1.6, -0.7	26	-1.6, -0.8	23
-0.7, 0.0	27	-0.8, 0.7	40
0.0, 1.6	35	0.7, 1.9	19
1.6, 2.3	5		

(3) For the section $S(I_0) = \{I_l\}_1^L$, we shall design a matrix M of $L \times L$ sizes, where the element (k, l) equals $\mu(I_k, I_l)$.

(4) For the matrix M, we calculate the value

$$F_{W,\mu_0}(I_0) = \frac{1}{L^2} \sum \gamma_{kl} \cdot \mu(I_k, I_l) . \tag{15}$$

(In this case, the weighted quantities $\gamma_{k,l}$ are taken equal to 1).

Now the functional F can serve for description of the statistical characteristics of sample W, that can be expressed as extremal statistics.

For instance, by using the method of exhausitive search, we can find an interval I^*, on which F_{W,μ_0} attains its maximum. This interval should be naturally regarded as the statistically stable grouping interval μ_0. The threshold value for μ_0 is statistically stable for the defined interval I^*, if

$$F_{W,\mu}(I^*) = F_{W,\mu_0 \pm \delta}(I^*) \tag{16}$$

i.e., small variations of μ_0 do not cause changes in the functional number.

TABLE VI

Random grouping intervals of logarithmic concentrations of lead. Borovoe, 1976-1983. (T) warm season, 1979; (X) cold season, 1979-1980. (%) percentage of points falling within the interval

T76	%	X7677	%	T77	%	X7778	%	T78	%
0.3, 1.5	26								
1.5, 2.3	34	1.2, 2.6	20	1.0, 2.0	37	1.2, 2.5	17	0.7, 2.5	45
				2.0, 2.7	33				
2.3, 3.7	40	2.6, 3.2	36	2.7, 3.7	30	2.5, 2.9	15	2.5, 3.0	41
						2.9, 4.2	68	3.0, 4.2	16
		3.2, 5.0	44						

X7879	%	T79	%	X7980	%	T80	%	X8081	%
		0.2, 1.6	11						
1.5, 2.5	40	1.7, 2.6	30			1.3, 2.4	27	1.2, 2.4	21
2.5, 3.1	16	2.6, 3.7	53	2.5, 3.7	40	2.4, 3.5	66	2.4, 3.1	21
3.1, 3.8	24							3.1, 4.0	49
3.8, 4.7	20	3.7, 4.3	6	3.7, 4.5	35	3.5, 4.1	7		
				4.5, 5.2	25			4.0, 5.2	9

T81	%	X8182	%	T82	%	X8283	%	T83	%
0.7, .25	7					0.7, 1.6	10	0.0, 1.5	25
1.5, 2.7	50	1.4, 2.8	32	1.0, 2.8	67	1.7, 2.3	24	1.5, 2.7	52
2.7, 3.5	42	2.8, 3.4	31	2.8, 3.6	27	2.4, 3.2	44	2.7, 3.7	23
						3.2, 3.6	8		
		3.4, 4.6	37	3.6, 4.3	6	3.7, 4.7	12		

Let us assume that a statistically stable grouping interval has been found for the given sample W, if for μ^*-stable, a statistically stable interval I^* is found such that

$$F_{W,\mu^*}(I^*) > -\epsilon.$$

As is evident from examination of the Tables of random central tendencies, the intervals are already ordered in accordance with the concentration values and their position within the 'chain' of intervals, which ultimately designates a single statistical grouping interval. We shall demonstrate how such an interval can be found from the data tabulated in Table XII, including random central tendencies of logarithmic concentrations of sulfur dioxide for the 'Borovoe' background monitoring station.

Let us examine the last four series from the Table corresponding to seasons T 82, X

TABLE VII

Random grouping intervals of logarithmic concentrations of lead. Berezina B.Z., 1980-1983. (T) war season, 1980; (X) cold season, 1981-1982. (%) percentage of points falling within the interval

X8081	%	T81	%	X8182	%	T82	%	X8283	%
		1.1, 1.7	6	0.6, 1.5	8	0.6, 1.1	15	1.0, 1.7	20
1.5, 3.1	31	1.7, 2.1	22	1.6, 2.0	12	1.1, 2.1	40	1.7, 2.9	43
		2.1, 2.8	36	2.1, 3.2	65	2.1, 2.9	30		
3.2, 3.5	22	2.8, 3.8	49	3.2, 3.8	15	3.0, 3.9	12	3.0, 3.7	27
3.6, 4.1	20							3.9, 4.3	10
4.2, 5.2	10								

T83	%	X8384	%
1.0, 2.1	29	1.0, 1.7	12
2.1, 2.5	16	1.8, 2.6	42
2.5, 3.3	39		
3.3, 3.8	7	2.5, 3.5	29
		3.5, 3.7	17
3.8, 4.3	7		

8283, T 83, X 8384. Let us assume that we are interested in identifying the interval containing point 0.0. As a trial value of I_0, let us take $I_0 = [-1.0, 1.0]$. The measures of proximity of the trial intervals to all other intervals of the series can be calculated from the sequence of intervals with maximal μ values. These intervals are [-0.2,1.0], [-0.5,1.1], [-0.4,1.6], [-0.5,1.7]. Besides, the interval [-0.2,1.0] is joined to the following interval [1.0, 1.8] – this does not follow from the requirements for a maximal value of the proximity measure, but from the fact that this interval includes only 6% of the total number of observations and cannot be used for independent analysis, since in this case it represents some weakly expressed process. The minimal value for the proximity measure is estimated in this case over the unified interval [-0.2, 1.8], and is 0.42. That is, it can be said that the interval [-0.1,1.0] is of 0.4 statistical stability in relation to the intervals of the sample under study describing the four seasons. It is obvious that for the four intervals taken from the sequence, an interval can be specified that ensures maximal μ_8 value and functional F. This interval, derived by averaging the limits, is [-0.7,1.5]. Now, considering the interval as I^*, we shall obtain the following μ-values for four of the sampled intervals: 0.77, 0.75, 0.95, 0.86. The chosen statistical grouping interval is of 0.75 statistical stability. The corresponding value for the functional, calculated from the matrix of mutual proximity of the intervals, is 0.8. This procedure represents a method for practical realization of the proposed algorithm, that enables the major conditions to be fulfilled, and ensures attainment of the functional extremum. It is obvious that such an exhaustive search is

TABLE VIII

Random grouping intervals of logarithmic concentrations of lead. Repetek B.Z., 1980-1983. (T) warm season, 1981; (X) cold season, 1981-1982. (%) percentage of points falling within the interval

X8081	%	T81	%	X8182	%	T82	%	X8283	%
				0.7, 2.2	42	0.6, 1.6	21	0.6, 1.4	25
1.0, 1.8	10	1.3, 2.0	32					1.5, 2.0	18
2.0, 2.7	45	2.0, 2.3	26			1.7, 2.3	36		
		2.3, 2.6	24						
2.8, 3.5	35	2.6, 3.2	18	2.2, 3.0	15	2.4, 3.0	49	2.0, 3.2	49
						3.0, 3.5	5		
3.5, 4.0	8			3.0, 4.2	13			3.3, 4.3	8

T83	%	X8384	%
0.6, 1.3	7	0.8, 1.7	12
1.3, 1.7	27		
1.8, 2.4	50	1.7, 2.2	88
		2.3, 2.7	12
2.4, 3.2	15		
		3.2, 4.3	9

possible only because of the small sizes of the samples used. When the available intervals amount to several hundreds, then it becomes necessary to optimize the procedures.

This procedure for the search for statistically stable grouping intervals was applied to all of the tables including series of central tendencies. The derived estimates are the optimal ones, from the point of view of μ^*, i.e., maximal μ^*-stability is attained. Estimates for the central tendencies are statistics to be used to derive estimates of the next level – of statistical grouping intervals. With the aid of these estimates, statistical inferences can be drawn concerning multi-year processes and processes of large scales. The possibility of drawing such inferences is discussed in the following section.

5.4. Analysis of Components of Background Air Pollution

The statistical grouping intervals for different pollutants and stations are listed in Tables XII–XIV. In addition to information on the limits of the intervals, the Tables include the mean weighted values of the components, averaged over the intervals involved in the formation of central tendencies, also the number of warm and cold seasons, during which the components, distinguished by the respective interval, were evident.

The selection of statistical grouping intervals is quite a statistically stable procedure in relation to seasonal changes or variations in the interval-estimations. This is supported by the fact that the estimate derived in the previous chapter of the interval [-0.7, 1.5] for the

TABLE IX

Random grouping intervals of logarithmic concentrations of dust. Borovoe, 1976-1983. (T) warm season, 1978; (X) cold season, 1979-1980. (%) percentage of points falling within the interval

T76	%	X7677	%	T77	%	X7778	%
0.3, 1.6	16	0.8, 2.0	20			0.8, 1.8	28
1.6, 3.8	60	2.0, 2.4	10	1.5, 3.3	39	1.8, 2.7	23
		2.4, 3.1	20			2.7, 3.3	39
3.8, 4.5	24	3.1, 4.0	50	3.3, 4.8	61	3.3, 4.2	10

T78	%	X7879	%	T79	%	X7980	%
		0.5, 2.0	34	0.9, 2.2	17		
1.3, 2.8	23	2.0, 2.7	30	2.2, 3.2	23	1.0, 2.5	20
						2.5, 3.2	42
2.8, 4.3	77	2.7, 4.0	36	3.2, 4.8	60	3.2, 4.7	35

T80	%	X8081	%	T81	%	X8182	%
2.2, 3.1	15	1.6, 2.7	21	1.5, 3.3	26	1.6, 3.3	85
		2.7, 3.7	42				
3.1, 3.8	59	3.7, 4.3	37	3.3, 5.0	74	3.3, 4.0	10
3.8, 4.6	26					4.0, 4.4	5

T82	%	X8283	%	T83	%
		0.7, 1.6	12		
1.2, 3.2	20	1.6, 3.1	57	1.3, 3.2	53
3.2, 4.7	80	3.1, 3.8	31	3.2, 4.2	47

four seasons is close to the estimate for the entire data series [-0.2, 1.4) – their proximity measure is 0.7. Concerning variations in estimates of the intervals, it should be noted that the applied method of evaluation (averaging) strongly reduces their influence, and the possibility of unification of the central tendencies in the formation of a statistically stable grouping interval ensures stability of a number of intervals. In order to obtain series of central tendencies, the help of several experts was enlisted. Although differences in the estimates reached 50% at times, variations in the resultant estimates for the statistical grouping intervals did not exceed 10%. Of importance here is the choice of the μ_0 value which to a great extent determines the selection of the intervals. In our case, $\mu_0 = 3$ was used. For practical purposes this is quite sufficient, geometrically it is a measure of proximity, for example, of two equal intervals that overlap each other by half their length.

TABLE X

Random grouping intervals of logarithmic concentrations of dust. Berezina B.Z., 1980-1983. (T) warm season, 1981 (X) cold season, 1981-1982. (%) percentage of points falling within the interval

X8081	%	T81	%	X8182	%	T82	%	X8283	%
				0.1, 1.5	13			0.1, 1.0	8
		1.5, 2.5	11			1.0, 1.8	8	1.0, 2.6	35
2.0, 2.6	15	2.6, 3.3	19	1.8, 3.1	29	2.0, 2.8	17	2.6, 3.3	22
2.8, 3.6	42					2.8, 3.3	20		
3.6, 4.0	30	3.6, 4.2	50	3.2, 4.3	67	3.3, 4.0	53	3.3, 4.0	25
4.2, 4.5	12	4.2, 4.8	20					4.0, 4.8	11

T83	%	X8384	%
1.3, 1.9	3		
2.0, 3.3	33	1.8, 2.7	27
		2.8, 3.5	29
3.3, 4.8	60	3.5, 3.9	24
		4.0, 4.5	10

Thus, statistically stable characteristics have been obtained that reflect the relationships in background air pollution monitoring data series. These relationships are caused, presumably, by the presence of several mechanisms governing the concentration formation processes that differ in their effects, on account of which the statistical grouping intervals are interpreted as estimates of the areas of action of different components of background air pollution. The formulation of the problem of component discrimination implies detection of components inherent to the given region, e.g., components associated with long-range transport, local and global anthropogenic effects, etc. It should be mentioned that such estimates can be derived only if additional information and parameters are introduced into the statistical model. The areas of action distinguished for the components by the proposed method, reflect certain general processes occurring in the impact regions and represent statistically derived relationships that, consequently, may be used for knowledgeable interpretation of the data. Such an analysis is outside the scope of the present study.

Let us examine the results of estimates of the effects of compositing of concentration formation factors, and the statistical relationships reflected in these estimates. From analysis of Table XII, it is apparent that the two lower intervals, which include the highest concentrations for all stations, result from the effects of factors that occur mainly during

TABLE XI

Random grouping intervals of logarithmic concentrations of dust. Repetek B.Z., 1980-1983. (T) warm season, 1981; (X) cold season, 1981-1982. (%) percentage of points falling within the interval

X8081	%	T81	%	X8182	%	T82	%	X8283	%
2.4, 2.9	8	2.6, 3.8	53	2.7, 3.3	11	2.9, 3.6	16	2.6, 3.2	20
2.9, 3.6	20								
3.6, 4.7	53	3.8, 4.2	32	3.3, 4.3	50	3.6, 4.0	23	3.2, 3.9	51
		4.2, 5.4	15	4.4, 5.0	31	4.0, 4.3	23	3.9, 5.0	29
						4.3, 5.0	38		
4.7, 6.4	19			5.0, 6.0	8				

T83	%	X8384	%
2.9, 3.7	27	2.5, 3.5	20
3.7, 4.3	37	3.5, 4.3	49
4.3, 5.1	34	4.3, 5.0	31

cold seasons. Thus, it becomes possible to distinguish these intervals for sulfur dioxide as effects of the heating season. In this case, the levels characterized by lower concentrations should be considered as background values; namely below 4.0 $\mu g\ m^{-3}$ for Berezin B.Z. and Borovoe, and below 2.7 $\mu g\ m^{-3}$ for Repetek B.Z. For the impact regions of Norway and Sweden, these background levels are 3.3 $\mu g\ m^{-3}$ and 5.0 $\mu g\ m^{-3}$. 70% of the observations for all stations, including those in Norway and Sweden, fall within these limits. As can be seen, for all areas of observation, the derived estimates are close to each other, that is, they may be used as estimates of background concentration level for continents.

From analysis of the lead component, based on the data listed in Table XIII, the following estimates were derived for the upper level of background values: 18 ng m^{-3} for Berezin B.R. and 12 ng m^{-3} for Repetek B.R. and Borovoe. Over 60% of the observations are below this limit.

On the basis of the data presented in Table XIV, similar estimates can be made for total suspended particulates. According to the available data, the background concentration levels lie below 57 $\mu g\ m^{-3}$ for Berezin B.R., 56 $\mu g\ m^{-3}$ for Borovoe, and 59 $\mu g\ m^{-3}$ for Repetek B.R. Over 50% of the observations lie below these limits.

The estimates derived in this manner for background concentration levels of atmospheric pollutants are in agreement with concepts on the background value as a statistically stable concentration level typical of the area. However, this is not the only consideration on which the determination of the background concentration evel should be based. In Section 5.1 we demonstrated the practical use of another intuitive method for determination of the background value (Rovinskii *et al.*, 1982).

The intervals represented in Table XII can serve as examples illustrating the foregoing.

TABLE XII

Statistically stable grouping intervals of sulfur dioxide concentrations, $\mu g\ m^{-3}$. Column 1 shows intervals; columns 2 and 3 – number of season during which this interval was observed, and the average percentage of observation falling within the given intervals (2 – warm, 3 – cold seasons, respectively), 4 – seasonal effects. (A) Berezina B.Z.; (B) Borovoe; (C) Repetek B.Z., (D) Jergul, Norway; (E) Abisko, Sweden

	A			B	
1	2	3	1	2	3
			0.1 , 0.33	5 (23%)	1 (16%)
0.22, 0.8	3 (14%)	0	0.25, 0.8	7 (27%)	1 (20%)
0.9 , 4.0	3 (79%)	4 (21%)	0.8 , 4.0	7 (68%)	7 (38%)
4.0 , 12.0	3 (10%)	4 (53%)	4.0 , 16.0	1 (17%)	8 (54%)
16.0 , 54.0	0	4 (29%)	16.0 , 56.0	0	3 (23%)

	C	
1	2	3
0.05, 0.1	1 (18%)	2 (4%)
0.12, 0.3	3 (26%)	4 (21%)
0.3 , 1.0	3 (36%)	4 (44%)
1.0 , 2.7	3 (27%)	2 (22%)
2.2 , 9.0	1 (10%)	3 (18%)

	D			E	
1	2	3	1	2	3
			0.04, 0.08	1 (15%)	0
0.08, 0.22	1 (8%)	1 (12%)	0.08, 0.2	1 (13%)	1 (7%)
0.22, 0.8	1 (37%)	1 (36%)	0.2 , 0.5	1 (23%)	1 (26%)
0.8 , 2.2	1 (38%)	1 (32%)	0.5 , 5.0	1 (59%)	1 (63%)
2.2 , 3.3	1 (14%)	1 (13%)			
3.3 , 6.7	1 (3%)	1 (7%)			
			5.0, 10.0		1 (5%)

TABLE XIII

Statistical grouping intervals of lead, ng m^{-3}. Columns 1, 2, 3 – see Table XII for legends. 4 – seasonal effects. (A) Berezina B.Z.: (B) Borovoe; (C) Repetek B.Z.

A			B		
1	2	3	1	2	3
2.7 , 5.5	3 (16%)	3 (13%)	1.5 , 4.5	4 (17%)	1 (10%)
5.5 , 18.0	3 (60%)	4 (50%)	4.5 , 12.0	8 (47%)	6 (23%)
18.0 , 45.0	3 (23%)	4 (27%)	12.0 , 40.0	8 (39%)	7 (59%)
45.0 , 59.0	1 (7%)	2 (15%)	40.0 , 56.0	3 (6%)	5 (30%)
57.0 , 100.0	0	1 (10%)	59.0 , 100.0	0	2 (17%)

C		
1	2	3
1.8 , 4.5	2 (14%)	3 (26%)
4.5 , 6.7	2 (30%)	2 (14%)
6.7 , 12.0	3 (45%)	2 (62%)
24.0 , 57.0		4 (9%)

It is apparent that the interval 0.2–0.8 $\mu g\ m^{-3}$ is a statistically stable interval in relation to the series of data from the other stations. The same refers to the interval 5–30 ng m^{-3} for lead shown in Table XIII. These examples illustrate that statistical grouping intervals can be derived on these intervals themselves. It is obvious that the series of statistical grouping intervals, determining the subdivision of the concentration axis, can be used for the identification of intervals of the 'second order of statistical stability'. In this case, statistical stability should be understood in relation to the influence of specific regional factors, which leads to concepts of processes developing over continents and, accordingly, to the concept of a background air pollution level for continents. At least it can be stated that the estimates derived for sulfur dioxide and suspended particulate matter are better than those available from the literature. The use of lower concentration intervals as background estimates incurs difficulties due to their weak expression in terms of the weighted quantities of the components and their frequency of occurence during different seasons. In respect to the estimates presented herein, it can be said that they are manifested at Borovoe, Repetek and at the stations in Jergul, Norway and Abisko, Sweden with a frequency of about 30% for sulfur dioxide, and with a similar frequency for suspended

TABLE XIV

Statistically stable grouping intervals of concentrations of suspended particulate matter, $\mu g\ m^{-3}$. Columns 1, 2, 3 – see Table XII for legends. 4 – seasonal effects. (A) Berezina B.Z.; (B) Borovoe; (C) Repetek B.Z.

A			B		
1	2	3	1	2	3
1.1 , 4.0	0	3 (9%)			
3.3 , 9.0	3 (7%)	1 (35%)	2.2 , 7.0	2 (16%)	4 (21%)
7.0 , 27.0	3 (30%)	4 (41%)	4.5 , 24.0	8 (32%)	7 (55%)
27.0 , 57.0	3 (54%)	4 (45%)	24.0 , 56.0	8 (67%)	7 (29%)
54.0 , 75.0	1 (20%)	3 (11%)			

C		
1	2	3
13.0 , 36.0	3 (32%)	4 (15%)
36.0 , 59.0	3 (31%)	4 (50%)
59.0 , 100.0	3 (37%)	3 (30%)
90.0 , 160.0		2 (14%)

particulate matter at Borovoe, Berezin and Repetek.

The components so identified may be used not only for the development of concepts on the background levels of pollution, but also to analyze the dynamics of the effects of air pollution at the background level. An example can be offered illustrating such effects for sulfur dioxide at the Borovoe station. Consider the concentration interval 4.0–16.0 $\mu g\ m^{-3}$. This interval is present in all cold seasons, the frequency distribution revealing that 50% of the data fall within this band. There is reason to believe that the chosen interval reflects the winter effects of anthropogenic factors, particularly of the heating season. If we examine this interval during different seasons, and estimate the corresponding central tendencies, an interesting fact emerges – the centres, appearing relatively stable, exhibit a distinct trend. They can be represented by the following row of numbers: 5.5, 6.0, 7.3, 9.0, 10.0, 10.0, 6.0, 12.0, showing a more than two-fold increase over an 8-yr period. Thus our method of analysis may provide estimates of trends in the background air pollution components. However, such an investigation demands greater knowledge of the process of formation of background air pollution components, which requires, first of all, a feasible analytical treatment of the estimates derived with the aid of the statistical model of background air pollution.

In concluding this chapter, let us discuss another statistical characteristic, relying on the use of the components of pollution distinguished in this study. Since, as has been shown above, different components make unequal contributions to the general level of pollution, it is natural to ask – what is the share of one or another component? An answer cannot be obtained from analysis of the data presented in the Tables, reflecting only the level of pollution typical of the components, and the frequency of their occurence in the data array. The 'weight' of the component can be characterized by the total concentration of the components during the entire period of their occurrence, as related to the sum of all concentrations for the period of time under study. In practice, such a summation is performed separately for the observations that fall within different grouping intervals, i.e., the sum C_i can be represented as:

$$\sum_{i=1}^{N} C_i = \sum_{C_i>a_0}^{C_i<a_1} C_i + \sum_{C_i>a_1}^{C_i<a_1} + \ldots + \sum_{C_i>a_{k-1}}^{C_i<a_k} C_i .$$

The weights of the components, λ_i are defined as

$$\lambda_i = \sum_{C_j \in [a_{i-1}, a_i]} C_j \Big/ \sum_{1}^{N} C_i , \quad \sum_{1}^{k} \lambda_i = 1 .$$

Let us now present some results derived from analysis of the weights of pollution components.

From an analytical treatment of the data series of sulfur dioxide, the total weight of the components, clustered above the 1 μg m^{-3}level, was found to be: at Berezina, B.R. for the warm periods – over 95%, for the cold periods – 100%; at Borovoe for the cold period – over 90%, for the warm period – over 85%; at Repetek B.R. for the cold periods – over 70%, for the warm periods – 70%. From the data series on lead concentrations, the total weight of the components exceeding 5 ng m^{-3} were over 95%, both for the warm and cold seasons, for the data from the Berezina, Borovoe and Repetek B.R. stations. At the same time, in nearly all cases, concentration intervals can be distinguished that are responsible for the major part of the pollution. For instance, for sulfur dioxide such intervals are: at the Berezina B.R. station for the warm periods – [0.9, 4.0] (73%); at Borovoe for the warm periods – [0.8, 4.0] (80%); at Repetek B.R. for the warm period – [1.0, 2.7] (53%).

Analysis of the weights and dynamics of the components is a subject area of special interest.

It can be seen that the two lower components from intervals [0.1, 0.3] and [0.25, 0.8] make an insignificant contribution. Statistically most stable is the contribution of the component [4.0, 16.0]. The highest variability is typical of the components from intervals [0.8, 4.0] and [16.0, 56.0]: during the period under discussion they changed 9–11 times, the changes being mutually interrelated. This pattern shows that in the area of the Borovoe station, the effects of the heating season have increased and shows the specific component responsible for this increase.

On the basis of the foregoing, the weighted values of the pollution components can be

recommended as statistical characteristics, reflecting the nature of the background air pollution in impact areas, and the suggestion is advanced that they should be used as criteria for tracking the dynamics of background pollutants, which is an essential monitoring problem.

6. Conclusions

The design philosophy employed in this study of a statistical model of background air pollution had the following objectives:

(1) Evaluation of the information content of background monitoring data for the description of the behavior, in space and time, of atmospheric pollutants arriving from impact areas, and in order to distinguish background air-pollution characteristics common to all time-periods and different monitoring stations.

(2) Elaboration of methods for the derivation of background air-pollution characteristics of temporal and spatial statistical stability. The major results are as follows:

(1) It has been demonstrated that the data obtained from measurements of different pollutants at different background monitoring stations, serve to define the subdivisions of the concentration scale into zones of action of several major pollution components. The problem of distinguishing the pollution characteristics common to different time periods and stations is therefore reduced to the problem of comparing the subdivisions of different concentration scales. This result has been derived on the basis of the design, analysis and interpretation of the statistical model, herein proposed.

(2) Mathematical techniques have been designed for the discrimination of concentration grouping intervals in the vent-data series, that are statistically stable in time and space, and can be interpreted as manifestations of background air pollution. This result was derived by employing methods specially developed for estimation of statistically stable grouping intervals, and for their interpretation as characteristics of different components of background air pollution. A number of inferences have been drawn concerning the nature of the data and analytical methods. They can be formulated as follows:

- Notwithstanding the high variability of the event-data, the body of available information and its accuracy allow one to distinguish in the data-series the effects of the same probabilistic processes.
- The information embodied in the observational series can be retrieved with the aid of the proposed statistical model, and represents different subdivisions of concentration scales, typical of the monitoring station and period of observation.
- Background air pollution in each area is controlled by several groups of factors that are manifested in the forms of different levels of pollution, i.e., the physical effects of different components of pollution are a statistical manifestation of the different zones of the concentration scales.
- A typical example, representing a typical time interval, that best reflects the action of different pollution components, is a seasonal data series.
- The subdivisions of the concentration scales, characterizing the effects of pollution by different pollutants measured at one observing station during several seasons, reveal

common features that enable one, with the use of specially designed techniques, to distinguish the components of pollution that are statistically stable and typical of the given area of observation.

- On the basis of the statistically stable pollution components, inferences can be drawn concerning the processes of air pollution in impact areas, and normal pollutant concentration levels in these areas.
- Analysis of the frequency of occurence of the components during different seasons, and of the share of different components to the general pollution of the atmosphere, enables one to describe the seasonal concentration variations; to identify the components that experience distinct anthropogenic effects, and to distinguish them from the components that define the background air-pollution level proper and to draw conclusions concerning the time variations for different components. Employing these techniques, estimates were derived for background air-pollution concentration levels for different pollutants and different monitoring stations, and the inference was drawn that sulfur dioxide has undergone a considerable increase in the area of the Borovoe station during winter periods, due to anthropogenic effects.

Hence, the statistical model herein proposed is a device designed to derive statistical information that can be interpreted explicitly. The problem concerning the comparison of the statistics, reflecting local and regional effects, and incidents of global effects (Dege, 1982; Zelenuk, 1984) can be solved within the framework of the proposed model by comparison of different statistical characteristics, using them as source-material for the designing 'statistics from statistics' that in terms of the model can be used for the description of the effects of events of large magnitudes. For instance, the use of the method of statistically stable interval estimation in application to intervals describing manifestations of pollution components, typical of different stations, enables one to pinpoint the intervals that are not only statistically stable in space, but also in time, i.e., to distinguish concentration intervals revealing certain common features within the continents.

These techniques require computational facilities and advanced computer programs. Further development of the background monitoring network, and expansion of the proposed methods over a broad class of background monitoring problems, call for the creation of effective man-machine systems in order to obtain relevant inferences from the accumulated data.

Use of the proposed model and application of related statistical characteristics for estimation of background air pollution, enable valid and statistically stable estimates to be derived concerning the actual state of the natural environment. This is regarded as one of the most essential problem areas to be solved using the background monitoring system.

Acknowledgments

The authors wish to thank Professor R. E. Munn who assisted greatly in clarifying some of the technical details and in editing, and M. Weinreich for the organization and final preparation of this paper.

References

Aivazyan, S. A., Enyukov, I. S., and Meshalkin L. D.: 1983, *Prikladnaya statistika (Applied statistics)*, Moscow: Nauka (in Russian).

Antonovsky, M. Ya., Buchstaber, V. M., and Zelenuk, E. A.: 1985, Background atmospheric pollution analysis on the basis of multi-mode distributions. Proceedings of the Internal Symposium on Integrated Global Monitoring of the State of the Biosphere, II Tashkent, U.S.S.R., 14-19 October, 1985. Tech. Doc. WMO/TD No. 151, Feb. 1987. *Tezisy dokl. III Mezhdunarodnovo simpoziuma 'Kompleksnyi global'nyi monitoring sostoyaniya biosfery'*, (Abstract of Report to III International Symposium 'Complex Global Monitoring of the Biosphere State') Leningrad: Gidrometeoizdat, pp. 53-54 (in Russian).

Augustinyak, S. and Sventz, S.: 1982, 'Determination of Environmental Changes on the Basis of the Generalized Signal Theory', in *Problemy fonovov monitoringa sostoyaniya prirodnoi sredy* (Problems of natural environment background monitoring). Vip. 2, Leningrad: Gidrometeoizdat, pp. 205-213 (in Russian).

Benarie, M. M.: 1982, 'Air-Pollution Modeling Operations and Their Limits', in *Mathematical Models for Planning and Controlling Air Quality*. IIASA Proceedings Series, v. 17, pp. 109-117.

Berlyand, M. E.: 1975, 'Sovremennye problemy atmosfernoi diffizii i zagryanzneniya atmosfery (Modern Problems of Atmospheric Diffusion and Air-Pollution)', Leningrad: Gidrometeoizdat (in Russian).

Berlyand, M. E.: 1984, 'Bearing on the Fundamental Principles for Air-Pollution Prediction', in *Sb. dokl. na Mezhdunarodnom soveshchanii VMO PA VI* (Coll. of Reports to the International Conference VMO PA VI), Leningrad: Gidrometeoizdat, pp. 9-15 (in Russian).

Burtseva, L. V., Lapensko, L. A., Volosneva, T. A., and Vas'kovskii, A. T.: 1982, 'Lead, Cadmium, Arsenic and Mercury Concentrations in the Atmosphere According to the Results Derived at the Borovoe Background Monitoring Station During the Period 1977-80', in *Monitoring fonovovo zagryazheniya prirodnoi sredy (Background Pollution Monitoring of the Natural Environment)*, Vip. 1, Leningrad: Gidrometeoizdat, pp. 101-11 (in Russian).

Burtseva, L. V., Volosneva, T. A., Lapenki, L. A., and Pastukhov, B. V.: 1982, 'Comparison of Methods for Monitoring Heavy Metals, Sulfir Dioxide and Sulfates', in *Monitoring fonovovozagryazneniya prirodnoi sredy (Background Pollution Monitoring of the Natural Environment)*, Vip. 1, Leningrad: Gidrometeoizdat, pp. 212-224 (in Russian).

Buchstaber, V. M., Zelenuk, E. A., and Maslov, V. K.: 1983, 'Methods of Analysis and Development of Automatic Classification Algorithms on the Basis of Mathematical Models', in *Prikladnaya statistika. Uchenye zapiski po statistike. (Applied Statistics. Scientific Notes on Statistics)*, T. 45, Moscow: Nauka, pp. 126-144 (in Russian).

Dege, S.: 1982, 'Simulation of Intraregional and Interregional Transfer of Pollutants and Evaluation of the State of the Natural Environment', in *Problemy fonovo monitoringa sostoyaniya okruzhayushchei prirodnoi sredy (Problems of Natural Environment Background Monitoring)*, Vip. 1, Leningrad: Gidrometeoizdat, pp. 141-147 (in Russian). Warsaw, v. 82, No. 3-5, pp. 23-37.

de Nevers, N., Lee, K. W. and Franc, N. H.: 1979, 'Patterns in TSP Distribution Functions', *Journal of APCA*.

Gruza, R. V. and Reitenbach, R. G.: 1982, *Statistika i analiz gidrometeorologischeskikh dannykk (Statistics and analysis of hydrometeorological data)*. Leningrad: Gidrometeoizdat (in Russian).

Harris, E. D. and Tabor, E. D.: 1956, 'Statistical Considerations Related to the Planning and Operation of National Air Sampling Network', *Proceedings of the 49th Annual Meeting of APCA*, Buffalo, pp. 7-9.

Izrael, Yu. A.: 1984, *Ekologiya i kontrol' sostoyaniya prirodnoi sredy (Ecology and Environmental Control)*. Moscow: Gidrometeoizdat (in Russian).

Izrael, Yu. A., Rovinskii, F. A., Antonovski, M. Ya., Buchstaber, V. M., Zelenuk, E. A., and Cherkhanov, Yu. P.: 1985, *K statisticheskomu obosnovaniyu komponent zagryazneniya atmosfery v fonovom raione. (On the Statistical Validation of Air-pollution Components in Normative Areas)*. Moscow: Doklady Academia of Sci., 276, No. 2, pp. 334-337.

Izrael, Yu. A., Antonovski, M. Ya., Buchstaber, V. M., and Zelenyuk, E.A.: 1987, *A Problem of Finding Out and Estimating Background Levels of the Components of Pollution of the Atmosphere*, Moscow: Doklady Academia of Sci., V. 292, N. 2.

Kleiner, B. and Gradel, T. E.: 1980, 'Exploratory Data Analysis in Geophysical Sciences', *Reviews of Geophysics and Space Physics*. V. 18, No. 3, pp. 699-717.

Larsen, R. J.: 1961, 'A Method of Determining Source Reduction Required to Meet Air Quality Standards', *Journal of APCA*, No. 11, p. 71.

Lynn, D. A.: 1976, 'Air Pollution', *Treat and Response*, N.Y., pp. 179-196.

Mage, M. D.: 1980, 'An Explicit Solution for S_B Parameters Using Four Percentile Points, *Technometrics*, No. 22, pp. 247.

Mage, D. T.: 1981, 'A Review of Applications of Probability Models for Describing Aerometric Data', *Environmetrics-81: Selected Papers*, SIAM, Philadelphia, pp. 42-52.

Mage, D. T. and Ott, W. R.: 1975, 'An Improved Model for Analysis of Air and Water Pollution Demands', *International Conference on Environmental Sensing and Assignment*, IEEE-ICESA, v. 1, pp. 20-5.

Pastukhov, B. V., Popova, E. V., and Syroegina, O. A.: 1982, 'Background Monitoring of Sulfur Compounds', in *Monitoring fonovovo zagryazneniya prirodnoi sredy. (Background Pollution Monitoring of the Natural Environment)*, Vip. 1, Leningrad: Gidrometeoizdat, pp. 83-95 (in Russian).

Perone, S. P., Pichler, M., Gaarenstrom, P., and Mayers, G. H.: 1975, 'The Application of Pattern Recognition Techniques to the Characterization of Atmospheric Aerosols', *International Conference on Environmental Sensing and Assignment*, IEEE-ICESA, v. 1, p. 5-4.

Rovinskii, F. Ya. and Buyanova, L. D.: 1982, 'Background Monitoring of the Natural Environment', in *Problemy fonovovo monitoringa sostoyaniya prirodnoi sredy Problems of Natural Environment Background Monitoring)*, Vip. 1, Leningrad: Gidrometeoizdat, pp. 5-11 (in Russian).

Rovinskii, F. Ya. and Wiersma, G. B.: 1987, 'Procedures and Methods for Integrated Global Background Monitoring of Environmental Pollution', *WMO Tech. Doc.* No. 178, *GEMS Info. Series* No. 5.

Zelenuk, E. A.: 1984, 'Statistical Analysis of Metrological Data in the System of Background Air-Pollution Monitoring', in: *Tez. dokl. pyatoi Vsesoyuznoi konferentsii 'Problemy metrologicheskovo obespecjeniya sistem obrabotki izmeritel'noi informatsii – SO11 – 5' (Abstr. Reports 5th All-Union Conference 'Problems of metrological provision for Aerometric Data Processing Systems - SO11 - 5')*. Moscow, pp. 38-41 (in Russian).

Zelenuk, E. A., Zubenko, A. A., and Cherkhanov, Yu. P.: 1984, 'Sampling and Data Analysis in the Background Monitoring System', in *Tez. dokl. pyatoi Vsesoyuznoi kunferentsii 'Problemy metrologicheskovo obespecheniya sistem obrabotki izmeritel'noi informatsii'* (Abstr. Reports 5th All-Union Conference 'Problems of meteorological provision for aerometric data processing systems – SO11 – 5'), Moscow, pp. 41-44 (in Russian).

Zelenuk, E. A. and Cherkhanov, Yu. P.: 1985, 'Investigations of Background Air-Pollution in Terms of the Statistical Model', in *Monitoring fonovo zagryazneniya prirodnoi sredy (Backgrouna Pollution Monitoring of the Natural Environment)*, Vip. 2,