IL NUOVO CIMENTO VOL. 14 C, N. 6 Novembre-Dicembre 1991

On the Possibility of Investigation of the Mass Composition and Energy Spectra of Primary Cosmic Ray (PCR) in the Energy Range from 10^{15} to 10^{17} eV Using EAS Data.

A. A. CHILINGARIAN and G. Z. ZAZIAN
Yerevan Physics Institute
Alikhanian Brother St. 2, 375036 Yerevan 36, Armenia, USSR

(ricevuto il 19 Aprile 1990; approvato il 2 Gennaio 1991)

Summary. — A method allowing one to determine the mass composition of primary cosmic radiation by means of simultaneous analysis of model and experimental data is presented in this paper. The most important part of this work is the quantitative comparison of multivariate distributions and the use of methods of nonparametric statistics for probability density estimation in a multivariate space of features. To check the method offered, events with $E_0 > 500$ TeV were generated by the Monte Carlo method. The showers generated were preliminarily processed by algorithms used in experimental data handling. The apparatus-induced distortions of the measured EAS characteristics have been taken into account. The method allows one to select an experimental event initiated by primary protons and iron nuclei with an efficiency of 70 – 80%. Also a new multivariate method of incident particle energy estimation based on the nonparametric regression is described. The method proposed, together with the above-mentioned multivariate EAS classification, allows one to determine the energy spectrum of incident protons and nuclei. Detection and investigation of the products of interaction of these particles with the atmosphere will allow us to study proton-nuclei and nuclei-nuclei interactions at energies from 10^{15} to 10^{17} eV.

PACS 94.40. – Cosmic-ray interactions with the Earth.

1. – Introduction.

The ambiguity of interpretation of the results of experiments with cosmic rays is connected with both significant gaps in our knowledge of the characteristics of hadron-nuclear interactions at superaccelerator energies and indefiniteness of the PCR composition. The extra difficulties are due to indirect experiments and hence, due to the use of Monte Carlo simulations of development and detection of different components of nuclear electromagnetic cascades.

To research into hadron-nuclear interactions in CR, one should know the type of cosmic-initiating incident particle. Besides, the investigation of the mass composi-

tion of PCR is of particular interest is connection with the problem of the CR origin.

At present the data available on the mass composition of PCR in the energy range $(10^{15} \div 10^{17})$ eV are obtained by detecting and investigating the different components of EAS and γ-families detected by X-ray emulsion chambers (X-REC). And if the first data set states a «normal composition»—extrapolation of PCR composition (40% protons and 20% iron nuclei) measured by direct methods in the energy range $(10^{11} \div 10^{14})$ eV [1, 5], then the data on γ-family fluxes testify to a decrease of the protons fraction in PCR at $E_0 > 10^{16}$ eV down to (15 ÷ 20)% and hence to dominance of iron nuclei [2].

This contradiction, yet inexplicable, may be due to different experimental data handling methods. Besides, the γ-family characteristics are more sensitive to the model of strong interactions than the EAS ones, hence, variation of model parameters can change the estimate of the proton and iron nuclei fractions. The problem of relative dominance of iron nuclei is very important, because the interpretation of the experimental data obtained in UHE CR is based on the mass composition of PCR [3].

The PCR energy spectrum has been investigated up to energy 10^{15} eV in sateilite and bailoon experiments. The measurements based on detection of EAS mainly fit to the sateilite and bailoon data in the energy range up to 10^{15} eV [4]. The integral spectrum index changes from 1.6 to 2 in the energy range $(5 \cdot 10^{14} \div 5 \cdot 10^{15})$ eV—the so-called spectrum «knee».

A possible step to the description of spectrum breaking is the study of the ener spectra of separate groups of nuclei and protons in the region of the knee, since by them one can judge about the validity of different models of the origin and propagation of CR. The proton energy spectra are studied up to 10^{15} eV by the JACEE collaboration [5] and in sateilite experiments up to 10^{14} eV [6]. Selecting the proton showers via the presence of the high-energy hadrons in the calorimeter ($E_h/E_0 \geq 0.25$), the energy spectrum of protons with energy 10^{15} eV has been obtained in ref. [7]. This method is based on the fact that the incident particle energy dissipation is more intense in the cascades initiated by incident nuclei. But due to large fluctations in the portion of energy transferred to hadrons at a fixed initial energy, such selection may considerably reduce the number of proton showers. In ref. [8] the fractions of different groups of nuclei in PCR were estimated and the energy spectra of the corresponding nuclei in the energy range $(10^{15} - 10^{16})$ eV were obtained by the method of solving the inverse problem. The results of ref. [7, 8] mainly coincide with the direct experiment data.

In this paper an approach is presented, which allows one to determine the mass composition of PCR by means of simultaneous analysis of model and experimental data. The most important part of the method is the quantitative comparison of multivariate distributions and use of a nonparametric technique to estimate the probability density in a multidimensional feature space. As compared to the earlier used method of inverse problem solution, with the help of which the mass composition of PCR was first determined in the energy range $E_0 > 10^{15}$ eV with sufficient quantitative certainty [8], in the proposed method the object of analysis is each particular event (a point in the multivariate space of EAS parameters) rather than alternative distributions of model and experimental data. That is why, along with the averaged characteristics, the belonging of each event to a certain class is determined.

This approach was first used to estimate the upper limit of the iron nuclei fraction

according to the γ-family characteristics [9, 10]. As opposed to ref. [10], where events were classified into two classes ($A > 50$ and $A < 50$), now it is possible to classify events into an unlimited number of classes [11].

After particle identification the new multivariate method of estimation of incident particle energy, used also to estimate the hadron energy according to X-ray emulsion chamber data [12], was applied to recover the energy characteristics of fluxes of different nuclei incident on the atmosphere.

Events with $E_0 > 500$ TeV have been simulated to check the methods proposed. The showers were registered at a depth of 700 g/cm². The model data were preliminarily handled according to algorithms used in the data handling at the Tien-Shan station [13]. The finite resolution of the installation measuring EAS characteristics has been taken into account.

2. – Simulation of a nuclear-electromagnetic cascade in the atmosphere.

The incident particle energy was drawn according to the energy spectrum given in ref. [4]. A normal CR composition was simulated (40% protons and 20% iron nuclei). Five groups of nuclei were considered: protons (p), α-particles (α), nuclei with $A = (7 - 16)$ CNO, $A = (24 \div 27)$ H and $A = (50 \div 56)$ VH. Protons and α-particles further were unified into one group.

Strong interactions were simulated by algorithms which reproduce the quark-gluon string model (QGSM) spectra for hA interactions [14]. The algorithm used allows one to simulate nucleon, pion, kaon and Λ-hyperon interactions with N^{14} at $0.03 < E < 10^6$ TeV. Production of $N\bar{N}$ pairs, π, k, Λ, γ was taken into account. In the frame of the Regge theory there were simulated processes of single and double reaction as well as inelastic recharging of $\pi^+ \to \pi^0$, $p \to n$.

The incident nucleus fragmentation was taken into account when simulating AN^{14} interactions. The hN^{14} interaction cross-sections were approximated as

$$(2.1)\qquad \begin{cases} \sigma_{\text{prod}}^{pN^{14}} = 295 - 23.94 \log(E) - 3.55 \log^2(E), \\ \sigma_{\text{prod}}^{\pi N^{14}} = 226 - 24.511 \log(E) - 2.31 \log^2(E), \\ \sigma_{\text{prod}}^{kN^{14}} = 198 - 27.021 \log(E) - 1.81 \log^2(E), \end{cases}$$

where E is the incident particle energy in the laboratory system of coordinates (E is in TeV). Energy dependence of the mean multiplicity of charged particles in hN^{14} interactions was approximated as

$$(2.2)\qquad \langle N \rangle_{ch} = 0.817 \ln^2 E - 3.127 \ln E + 10.27 \qquad (E \text{ is in TeV}).$$

The number of secondaries was drawn according to KNO distribution. It was taken that the mean transverse momentum of secondaries increased with the energy according to

$$(2.3)\qquad \langle P_\perp \rangle = 0.36(1 - 0.023 \ln(E/0.1)) \qquad (E \text{ is in TeV}).$$

For different secondaries $\langle P_\perp \rangle$ and the shape of distribution over $P_\perp$ differed and corresponded to the existing accelerator data.

Electromagnetic interactions were simulated as in ref. [15]. Pair production,

bremsstrahlung and multiple Coulomb scattering were taken into account. At the same time, it was assumed that transversal development of electron-photon cascades is due to only multiple Coulomb scattering. To calculate the average EAS characteristics, we have used the approximated formulae obtained in ref. [16].

3. – Comparison of EAS single characteristics and choice of optimal features.

To choose features most sensitive to the PCR composition, the single characteristics of EAS initiated by primary protons and iron nuclei were compared. The following EAS characteristics have been considered: the total number of electrons N_e, the total number of muons with $E_\mu > 5$ GeV, the EAS age parameter S, the total number, energy, mean energy, average distance to EAS cores and dispersion of spatial and energy distributions of muons with $E_\mu > 200$ GeV and hadrons with $E_h > 200$ GeV in EAS and the linear regression coefficients of spatial and energy distributions of muons and hadrons ($E = C1 + C2R$).

A quantitative comparison of various features is presented in table I, where the P-quantiles of statistical tests of comparison of samples from univariate distributions as well as the Bhattacharya probabilistic distance [17] between the samples are given. The Student, Kolmogorov, Mann-Whitney tests have been used. It follows from these data that the most appropriate feature to determine the EAS composition are the high-energy muon characteristics. The hadron component characteristics are less sensitive to the primary particle type (the higher the P-quantile of the test, the stronger the difference between the corresponding distributions). In this paper we have used EAS characteristics only (N_e, N_μ ($E > 5$ GeV), S). Though the sensitivity of N_e and S to the primary particle type is low, however, due to a different degree of correlation between N_e, N_μ and S for events initiated by primary protons and iron nuclei, as is seen from table II and III, the use of all the three EAS characteristics essentially improves the event classification reliability. The choice of these characteris-

TABLE I. – *P-quantiles of statistical tests of comparison of univariate distribution of different characteristics of* EAS *initiated by primary protons and iron group nuclei at* $1 \cdot 10^5 < N_e < 2 \cdot 10^6$.

	Student	Kolmogorov	Mann-Whitney	Bhattacharya distance
N_e	0.835	1.021	1.17	0.002
$N_\mu (E_\mu > 5\,\text{GeV})$	23.647	7.329	16.12	0.433
S	7.046	3.173	6.60	0.074
$N_\mu (E_\mu > 200\,\text{GeV})$	25.435	7.553	16.34	0.451
$\sum E_\mu (E_\mu > 200\,\text{GeV})$	15.217	5.552	12.45	0.189
$\langle E_\mu \rangle (E_\mu > 200\,\text{GeV})$	6.792	5.528	11.13	0.458
$\langle R_\mu \rangle (E_\mu > 200\,\text{GeV})$	18.341	6.533	14.36	0.295
$N_h (E_h > 200\,\text{GeV})$	4.717	2.400	4.88	0.248
$\sum E_h (E_h > 200\,\text{GeV})$	4.015	2.609	5.30	0.495
$\langle E_h \rangle (E_h > 200\,\text{GeV})$	3.503	1.656	3.44	0.128
$\langle R_h \rangle (R_h > 200\,\text{GeV})$	5.903	3.395	6.73	0.033

TABLE II. – *The coefficients of correlation between the characteristics of the electron-photon and the muon components of EAS with $1 \cdot 10^5 < N_e < 2 \cdot 10^5$ from primary protons (the quantities marked with * correspond to muons with $E_\mu > 200$ GeV).*

	N_e	N_μ	S	N_μ^*	$\sum E_\mu^*$	$\langle E_\mu^* \rangle$	$\langle R_\mu^* \rangle$
N_e	****	0.396	−0.132	0.145	0.108	0.01	−0.180
N_μ	0.393	*****	0.443	0.838	0.717	0.044	0.516
S	−0.132	0.443	*****	0.468	0.441	0.039	0.469
N_μ^*	0.145	0.838	0.468	*****	0.896	0.104	0.726
$\sum E_\mu^*$	0.108	0.717	0.411	0.896	*****	0.476	0.658
$\langle E_\mu^* \rangle$	0.010	0.044	0.039	0.104	0.476	*****	0.093
$\langle R_\mu^* \rangle$	−0.18	0.516	0.469	0.726	0.658	0.093	*****

TABLE III. – *The coefficients of correlation between the characteristics of the electron-photon and the muon components of EAS with $1 \cdot 10^5 < N_e < 2 \cdot 10^5$ from primary iron nuclei (the quantities marked with * correspond to muons with $E_\mu > 200$ GeV).*

	N_e	N_μ	S	N_μ^*	$\sum E_\mu^*$	$\langle E_\mu^* \rangle$	$\langle R_\mu^* \rangle$
N_e	****	0.555	−0.205	0.353	0.366	0.252	−0.026
N_μ	0.555	*****	0.195	0.830	0.820	0.375	0.345
S	−0.205	0.195	*****	0.238	0.224	0.075	0.224
N_μ^*	0.353	0.830	0.238	*****	0.978	0.391	0.633
$\sum E_\mu^*$	0.366	0.820	0.224	0.978	*****	0.556	0.611
$\langle E_\mu^* \rangle$	0.252	0.375	0.075	0.391	0.566	*****	0.213
$\langle R_\mu^* \rangle$	−0.026	0.345	0.224	0.633	0.611	0.213	*****

tics is also due to their relatively low sensitivity to strong interaction characteristics, which allows one to hope for obtaining a model-independent inference on the PCR mass composition.

4. – Classification of the distribution mixture.

Let us consider the stochastic mechanism $(A, \mathcal{P})$ which generates the observation v in a multivariate feature space (v is a d-dimensional vector of values measured in experiment, d is dimensionality of the feature space). The basic states space A is a unification of events from different primary nuclei. We know no law of nature like $(A, \mathcal{P})$, that is why, to determine a probability measure on A, the total Monte Carlo simulation of the phenomenon under investigation is performed, including experimental data registration and handling.

The set of d-dimensional U vectors obtained in simulations is the simular analog of the experimentally measured values of V. But, as opposed to experimental data, it is known to which of the alternative classes each of the events belongs. These «labeled» events include *a priori* information about dynamics of the process under investigation, which is given in a nonparametric form, as finite-size samples. The sequence

$\{U, t_j\}$, where $i = 1, M_{TS}$, $j = 1, L$. t_j is the class index, we usually call a training series or sample (TS) which is also denoted by $(A, \mathcal{P})$.

Since both physical processes of particle production and those of registration are stochastic and the information about the phenomena under investigation is smeared out, the data analysis is uncertain in the sense that one need not wait for event separation into compact nonoverlapping groups corresponding to different primary nuclei. The only thing we can require when classifying experimental data by various primary nuclei is to minimize the losses due to an incorrect classification to some degree and to ensure the use of *a priori* information completely. Such a procedure is the Bayes decision rule with nonparametric estimation of the multivariate probability density function, which, when using a simple loss function (the loss is zero in case of correct classification and is the same at any error), takes the form

$$\hat{A} = r(v, A, \mathcal{P}) = \operatorname{argmax}\{\hat{P}(A_i/v)\}, \qquad i = 1, L, \tag{4.1}$$

where $\hat{P}(A_i/v) \sim P_i \hat{P}(v/A_i)$ are *a posteriori* densities. $\hat{P}(v/A_i)$ are conditional densities which are estimated by TS $(A, \mathcal{P})$ using one of many nonparametric methods available [18]. L is the number of groups of nuclei.

Initial (*a priori*) values of P_i are taken equal. The monograph [19] is devoted to the interplay of *a priori* and experimental information in fraction estimation problems. Here we shall not go into discussion of competence of the choice of a uniform *a priori* distribution, but only mention that at such a choice the *a posteriori* probability and hence, the results of classification, will be totally defined by experimental information, which seems reasonable to us in the given physical task.

To estimate conditional densities, we used Parzen's method with automatic kernel width adaptation. In this method some probability density values are calculated which correspond to different values of method parameters. Then the sequence obtained is ordered and the median of this sequence is chosen as final estimate. Depending on the value of the probability density in the vicinity of V, due to stabilizing properties of the median, each time we will choose an estimate with a width most fitting for that region [20]. The probability density is estimated by

$$\hat{P}(V/A_i) = 1/(2\pi^{d/2} h^d) \sum_{j=1}^{M_i} \exp[-r_j^2/h^2] W_j, \qquad i = 1, L, \tag{4.2}$$

where d is the feature space dimensionality, M_i is the number of vectors of the i-th TS class, r_j is the distance to the j-th neighbour in the Mahalanobis metric:

$$r_j = (V - U_j)^T R^{-1} (V - U_j), \tag{4.3}$$

where R is a sampling covariance matrix of class to which U_j belongs, W_j is the event weight, h is the kernel width.

The classification methods, like all the statistical ones, include the procedure quality test as a necessary element. This stage beside all the others is also necessary for the determination of the primary composition. The most natural procedure quality estimate is the error probability which depends on both the degree of overlapping of alternative multivariate distributions and the decision rule being used (Bayes decision rules provide minimum error probability as compared to any other one):

$$R_N^B = E\{\theta[r(V, A, \mathcal{P})]\}, \tag{4.4}$$

where

$$\vartheta\{t_j, r(U, A, \mathcal{L})\} = \begin{cases} 0, & \text{at correct classification}, \\ 1, & \text{otherwise}. \end{cases}$$

and E stands for mathematical expectation. The expectation is taken over all possible samples of volume M and over the whole d-dimensional space of measured values.

Since we do not exactly know to what class the experimental vectors belong, the estimate of R_{ij}^{ϑ} we obtain via TS is

$$R_{ij} = 1/M_{TS} \sum_{i=1}^{M_{TS}} \vartheta\{t_j, r(U_i, A, \mathcal{L})\}, \tag{4.5}$$

i.e. we classify the $\{U_i\}$ TS and check the correctness of classification over the index of the class t_j, $j = 1, L$. However, as numerous investigations have shown (*e.g.*, [21]), this estimate is systematically biased and hence, a cross-validation estimation is preferable:

$$R_{ij}^{\vartheta} = 1/M_{TS} \sum_{i=1}^{M_{TS}} \vartheta\{t_j, r(U_i, A, \mathcal{L}_{(i)})\}, \tag{4.6}$$

where A, $\mathcal{L}_{(i)}$ is a TS with a removed i-th element, which is classified. This estimate is unbiased and has an essentially smaller r.m.s. deviation. The advantage of R_{ij}^{ϑ} is especially notable when the feature space has a higher dimensionality [22].

Note that we have the possibility to estimate the errors probability of various types by imposing to classification various TS classes, $\{U_i, t_j\}$, $j = 1, L$. L is the number of classes.

By R_{ij} we denote the probability for the classification of the i-th class events as belonging to the j-th class (misclassification).

Now let us estimate the *a posteriori* fraction of various kernel types in the incident flux.

It is known [23] that the best estimate of *a posteriori* fraction (in case of uniform *a priori* information and absence of classification errors) is the empirical fraction

$$P_i^* = M_i / M_{\Sigma t}, \tag{4.7}$$

where M_i is the number of events classified as initiated by the kernel group A_i, $M_{\Sigma t}$ is the total number of events registered during experiment. It can be shown (see [24], where a formula for the case with $L = 2$ is derived) that with account of classification errors the fraction of various kernels can be obtained as the solution of the following set of linear equations:

$$\left(1 - \sum_{j \neq i}^{L} R_{ij}\right) P_i - \sum_{k \neq i}^{L} P_k R_{ki} = P_i^*, \qquad i = 1, L. \tag{4.8}$$

In the first sum summation goes over j, in the second over k. All estimates of R_{ij} and P_i^* are obtained over one and the same TS using the same decision rules.

The accuracy of estimates is defined by the TS size and number of experimental data as well as by the value of the classification errors, which present the «quality» of

discrimination in the chosen feature subset. Note that the set (4.3) is a poorly defined system and at large values of classification errors the solutions of the set are unpredictable and hence, the choice of a feature combination providing a high percentage ($\geqslant 60\%$) of correct classification is a necessary preliminary stage.

5. – The method of primary flux determination.

As we have shown in the previous section, to estimate the fraction of various nuclei in an incident flux of cosmic radiation, beside classification of an experimental sample by a TS, it is also necessary to calculate any misclassification coefficients, R_{ij}. The error in the determination of the fraction of various kernels is a function of the errors both from classification and in determination of R_{ij}.

The possibility to decrease the bias and variance of misclassification rates estimates was discussed in ref. [24], where it was mentioned that it is possible to improve the accuracy of R_{ij} estimates, if the TS size is large enough to separate the TS into independent subsamples.

Unfortunately, time consumption per model event generation increases abruptly with energy and we have not to expect much model information in the energy range $E > 10^{15}$ eV.

Thus, the problem of an efficient use of the information contained in simulation results is as never actual for cosmic ray and accelerator physics, since the classical sampling models do not allow us to extract the whole information carried by a sample.

The methods of sample control during handling are widely used in the last few years. One of these is the leave-one-out-for-a-time test considered in the previous section, which allows us to decrease the sample bias.

A more efficient procedure actively developing in both applied and theoretical respects in the last decade is the bootstrap which lies in replication of the initial sample very many times by means of random sampling with replacement [25]. The thus obtained conditionally independent bootstrap-replicas in many respects stand for independent samples from the general population (under the condition of sufficiently large size of the initial sample), In fact, the bootstrap substitutes the unknown general population by a single sample, i.e. the ideology described in sect. 4 of this paper is followed.

There are two ways of distribution mixture coefficient estimation: i) to obtain the bootstrap estimate of the misclassification coefficients R_{ij}^*, then classify and estimate the fraction or ii) carry out fraction estimation over each bootstrap replica, then obtain the fraction and the standard deviation bootstrap expectation. The second way is preferable, because obtaining the standard deviation in the first case is time-consuming. It is enough to say that the errors propagation formulae obtained by the REDUCE symbolic manipulation program occupy several standard sheets in case of classification into four classes.

The bootstrap method of the distribution mixture coefficient estimation takes following form:

$$P \equiv P(P_1, \ldots, P_k) = f(V, \mathscr{P}, c(V, A, \mathscr{P})), \tag{5.1}$$

This solution is a complex function of experimental data and the TS as well as the decision rule c being used. By several TS bootstrap replicas we calculate the bootstrap expectation and the bootstrap standard deviation of the mixture coefficients P_i.

TABLE IV. – *The Bayes error matrix obtained by the leave-one-out method, by* TS *within the range* $1 \cdot 10^5 < N_e < 2 \cdot 10^5$.

	P	CNO	H	VH
P	0.798	0.102	0.067	0.033
CNO	0.127	0.688	0.105	0.080
H	0.072	0.113	0.691	0.124
VH	0.034	0.090	0.150	0.726

which are used as estimates of the fraction of different kernel groups in the primary flux.

To test the method, the generated events were grouped in two. The first were used to create a TS and the second as pseudo-experimental events. The EAS characteristics (N_e, N_μ, S) were used in the events classification, where events in different fixed intervals over N_e were selected. The TS consisted of four classes in accordance with the primary kernel type (p-protons and α-particles, CNO-kerneis with $A = 7 \div 16$, H with $A = 24 \div 27$ and VH with $A = 50 \div 56$).

Table IV presents the Bayes error matrix obtained as a result of a leave-one-out test over TS. The diagonal elements of this matrix show the probability for a correct events classification and the nondiagonal elements the probability for misclassifications. It is seen from table IV that the correct classifications make about $(70 \div 80)\%$ (classification of «boundary» groups (protons and iron group nuclei) is essentially better than that of the intermediate groups). Note that the accuracy of classification can be improved by selecting events at narrow zenith angles ϑ (ϑ varies between 0 and 45°).

Table V shows the recovered nuclei fractions obtained by classification of model events for one interval over N_e. The errors presented are obtained by the bootstrap procedure. Fractions of kernel groups given in EAS simulations (true fractions) are presented *ibid*. As is seen from this table, the proposed method allows us to determine the fraction of protons and iron nuclei in the incident flux with quite a good accuracy. To improve the accuracy of determination of the fraction of intermediate nuclei it is necessary to increase the size of TS.

In this work events obtained by the same model are used as control (pseudo-experimental) and training samples. During the experimental data handling the model adequacy test is a necessary stage. The difficulty lies in the fact that the changes both in the strong interaction model and in the mass composition can lead to the same change of the observed values. To overcome this ambiguity one can use the «self-consistency» method developed in ref. [10].

TABLE V. – *Recovered fractions of four groups of nuclei within* $1 \cdot 10^5 < N_e < 2 \cdot 10^5$ (W_{in} *is a «true» fraction,* W_{out} *a recovered one*).

	N_{ev}	W_{in}	$E_\alpha(W_{out})$	$\tau_\alpha(W_{out})$
P	200	0.370	0.345	0.038
CNO	188	0.272	0.229	0.067
H	194	0.168	0.222	0.057
VH	163	0.189	0.194	0.019

6. – The nonparametric regression method.

As ground for estimation of the primary particle energy serves the fact of its correlation with the measured EAS parameters. Table VI presents the coefficients of the primary particle energy correlation with various shower parameters. It is seen that though the total number of electrons in a shower (N_e) is the main parameter used to estimate the primary energy, the characteristics of the muon component of EAS correlate with E_0 somewhat better. That is why our purpose was to investigate the possibility of improvement of the accuracy of estimation of the primary particle energy via the characteristics of the electron-photon and the muon components of EAS.

First, some words about general formulation of the regression problem (we will mainly follow ref. [26]). Suppose a flux of particles is sporadically and independently incident on the atmosphere in accordance with some spectrum $f(E)$. Then these particles, undergoing random collisions and interactions with air atom nuclei, initiate an extensive air shower, the parameters of which are registered by the experimental set-up, *i.e.* each value of E is put into coincidence with some random vector of measurements, X, according to some conditional probability density $P(X/E)$.

The peculiarity of solution of the regression problem in the cosmic-ray physics is the fact that neither the true spectrum $f(E)$ nor the conditional density $P(X/E)$ are known in the general case, but there is a training sequence $\{E_i, X_i\}$, $i = 1..M_{TS}$ (obtained by simulation) and it is required to «recover» the regression $E = E(X)$ by this sequence (M_{TS} is the number of events in the training sample).

In the absence of systematic errors (the mathematical expectation of random vector measurement at a fixed independent variable (energy) is equal to the regression function value in that point) this problem is reduced to one of minimization of the average risk:

$$I(\alpha) = \int (E - F(X, \alpha))^2 P(X, E)\, dX\, dE\,, \tag{6.1}$$

where $F(X, \alpha)$ is some functional family depending on the parameter α. $P(X, E) \sim P(X)P(X/E)$ is the probability density function. If there is available *a priori* information about the form of probability function and the chosen functional family $F(X, \alpha)$ is not too complex, then the regression problem can be solved by the least mean squares or the maximal likelihood standard methods.

Due to the complicated stochastic picture of particles and nuclei passing through the atmosphere and the detectors, we have not to expect a standard probability interpretation of all random processes, that is why we have chosen a method

TABLE VI. – *The coefficients of correlation of the characteristics of the electron-photon and the muon components of* EAS *with initial energy* $1 \cdot 10^5 < N_e < 2 \cdot 10^5$.

	N_μ	N_μ ($E_\mu > 5$ GeV)	S	N_μ ($E_\mu > 200$ GeV)	$\sum E_\mu$ ($E_\mu > 200$ GeV)
P	0.355	0.730	0.33	0.678	0.584
Fe	0.495	0.953	0.23	0.899	0.892

based on a nonparametric way of treatment of *a priori* information, which does not impose any structure and totally uses the information carried by TS.

The method is based on the obvious fact that the events close to some metric (usually the Mahalanobis metric [27] is used) in the feature space have similar energy: the compactness hypothesis. The method based on the consideration of the «nearest neighbours» is first analysed in ref. [28]. In this work it was shown that when the number of the nearest neighbours, K, and the total number of events in TS, M, tend to infinity so that $K/M \to 0$, then the risk of the procedure tends to the minimum achievable Bayes risk and even the use of one neighbour increases the risk only twice as compared to the Bayes risk. The uniform consistency of the following estimate is shown in ref. [29]:

$$\hat{E}(X) = \sum_{i=1}^{k} C_i E_{(i)}(X), \qquad \sum_{i=1}^{k} C_i = 1, \tag{6.2}$$

where $E_{(i)}(X)$ is the value of the independent variable (energy) of the i-th nearest neighbour of the event X in the feature space.

The weight coefficients C_i are optimized by TS so that some quality function, *e.g.*, the mean-square error (MSE) of estimation, is minimized.

$$\mathrm{MSE} = \sqrt{\sum_{i=1}^{M_{TS}} (E_i - \hat{E}_{(i)}(X))^2 / M_{TS}}, \tag{6.3}$$

where the index (i) means that the i-th event, the energy of which is estimated, is temporarily removed from TS (leave-one-out test). Despite the fact that the nonparametric procedures are optimal under unlimited sampling, for the case of finite samples there are practically no theoretical and practical recommendations on the choice of the method parameters (*e.g.*, the number of nearest neighbours). That is why we apply the estimate adaptation ideology to the regression analysis, which was developed for multivariate nonparametric estimation of density function [20]. In this approach there are simultaneously calculated several estimates corresponding to different method parameters. The median of the ordered sequence is taken as final estimate.

7. – The method of particle energy determination.

To estimate the accuracy of primary particle energy determination by the method of nonparametric regression, there were generated showers with initial energy $E_0 > 500$ TeV. The preprocessing of showers was carried out by the data handling algorithms used in the Tien-Shan experiment [13]. The detector-induced fluctuations were taken into account when determining the characteristics of the electron-photon and the muon components. After preprocessing of showers, part of them were used as TS and another as «experimental» data. Table VII presents the results of estimation of the energy of «pseudo-experimental» events in various ranges of N_e initiated by incident protons and nuclei with $A > 24$. The relative mean-square errors (RMSE) are presented *ibid.*

$$\mathrm{RMSE}_i = (E_i - \hat{E}_i)/E_i, \qquad i = 1, M, \tag{7.1}$$

TABLE VII. – *Mean-square errors of estimation of the energy of protons and nuclei with* $A > 24$.

$\langle N_e \rangle / 10^5$	0.66 ± 0.196	1.404 ± 0.281	2.716 ± 0.534	9.758 ± 1.079
$\langle N_\mu \rangle / 10^3$	2.74 ± 0.258	3.839 ± 0.335	5.881 ± 0.667	14.545 ± 1.01
N^p_{TS}	864	913	484	402
N^p_{exp}	439	465	256	216
$RMSE^p$ (%)	20	24.3	25	25
N^A_{TS}	377	357	184	128
N^A_{exp}	184	225	102	73
$RMSE^A$ (%)	10.1	10.6	9.7	10.6

where E_i is the true energy of the event estimated and $\hat{E}_i$ is its nonparametric estimate. The features used are N_e and N_μ; it is seen from the table that in case of events initiated by incident nuclei, the mean-square error of estimation is 2.5 times smaller than that for proton-induced events, which is due to a stronger correlation of EAS parameters with the initial energy in the case of nuclear events as compared to the proton-initiated events (see table VI).

The «true» integral energy spectra of protons and nuclei with $A > 24$ (the true spectrum corresponds to 100% of correct classification of protons and nuclei and to zero error in the determination of primary particle energy) and the «experiment» spectra obtained as a result of Bayesian classification of «pseudo-experimental» events and then by nonparametric estimation of energy are compared.

There is a satisfactory agreement between the «true» and estimated energy spectra. The energy of events is somewhat overestimated for the «proton» events (the measured EAS characteristics of misclassified nuclei are attributed to protons with high energy) and are underestimated in the case of selected events attributed to heavy nuclei (vice versa). The presence of such distortions leads to some change in the index of the integral energy spectrum of incident protons and nuclei. To obtain a quantitative estimate of the degree of distortion of the index of the integral energy spectrum of protons and heavy nuclei, the corresponding distributions were approximated by power law, with the help of the minimization program «FUMILI» from the CERN program-library. For true proton events the spectrum is approximated by

$$I_p^{tr}(E > 500) = (4.75 \pm 0.36)\, E^{-1.75 \pm 0.013},$$

for selected proton events it is approximated by

$$I_p^{est}(E > 500) = (4.59 \pm 0.061)\, E^{-1.73 \pm 0.029}.$$

For the events initiated by heavy nuclei it is approximated by

$$I_A^{tr}(E > 500) = (4.56 \pm 0.075)\, E^{-1.69 \pm 0.030},$$

$$I_A^{est}(E > 500) = (4.84 \pm 0.099)\, E^{-1.77 \pm 0.037},$$

respectively.

It follows from these formulae that the relative error at the determination of the

integral energy spectrum index is 3% for incident protons and 5% for incident nuclei with $A > 24$.

8. – Conclusion.

The classification method allows one to select experimental events initiated by incident protons and nuclei with an efficiency of $(70 \div 80)\%$ and determine the mass composition of PCR at energies from 10^{15} to 10^{17} eV. The main advantages of the method proposed are:

i) its being a multivariate one, *i.e.* inclusion of additional EAS parameters in the analysis meet no difficulties;

ii) individual analysis (event by event)—each experimental event is an object of analysis—their belonging to a certain class and the error of statistical solution are determined;

iii) *a priori* chosen probability family is not imposed on data—the results of simulation are used directly during the process of statistical solutions.

We hope that the use of the proposed method when handling the experimental data obtained at complex arrays will allow us to get unambiguous information about the character of strong interactions at superaccelerator energies.

The nonparametric regression method for studying the energy spectra of PCR in the energy range from 10^{15} to 10^{17} eV via EAS data is based on

i) high reliability of multivariate classification of EAS;

ii) high accuracy of primary particle energy determination by the nonparametric regression method (relative mean-square error is $(10 \div 25)\%$).

By the characteristics of the electron-photon and the muon components of EAS we can determine the parameters of proton and nuclear «beams» incident on the atmosphere.

Detection and investigation of the products of interaction of these particles with the atmosphere (target) will allow us to study PA and AA interactions at energies $(10^{15} \div 10^{17})$ eV.

* * *

We are grateful to A. M. Dunaevsky and N. Stamenov for useful discussions and to E. A. Mamidjanian for stimulating interest in the work.

One of the authors (ZGZ) also thanks A. M. Dunaevsky for provision of EAS simulation algorithms.

REFERENCES

[1] J. N. STAMENOV: *High energy cosmic ray mass composition on the basis of* EAS *studies at mountain altitudes*, in *V International Symposium on Very High Energy Cosmic Rays Interactions, Lodz, Poland, 1988.*

[2] MT. FUJI EMULSION CHAMBER COLLABORATION: *Apparent decrease of proton fraction*

around the «Knee» observed in mt. Fuji emulsion chamber experiment, in *Proceedings of the XX ICRC*, Vol. 1 (Moscow, 1987), p. 382.
[3] J. RICH, D. L. OWEN and M. SPIRO: *Phys. Rep.*, **151**, 1 (1987).
[4] S. I. NIKOLSKY: *Energy Spectrum and Mass Composition of* PCR (Nauka, Moscow, 1987).
[5] JACEE COLLABORATION: *Energy spectra of primary protons and helium nuclei in the energy range* ($10^{12} - 10^{15}$) eV *from* JACEE, in *Proceedings of the XX ICRC*, Vol. 1 (Moscow, 1986), p. 371.
[6] I. P. IVANENKO *et al.*: *Spectrum and charge composition of cosmic rays with energy from* 2 *to* 100 TeV, in *V International Symposium on Very High Energy Cosmic Rays Interaction* (Lodz, Poland, 1988), p. 197.
[7] A. P. CHUBENKO, N. P. KRUTIKOVA *et al.*: *Primary protons, inclusive cross-section in P-A interaction and their relation to energy spectra of hadrons in the atmosphere*, in *Proceedings of the XX ICRC*, Vol. 1 (Moscow, 1986), p. 373.
[8] I. N. STAMENOV: PCR *mass composition investigation by* EAS, Ph. doctor dissertation, Lebedev Physics Institute (1981).
[9] A. A. CHILINGARIAN: *The CR high energy particles investigation with nonparametric statistics*, Ph. doctor dissertation, Yerevan Physics Institute (1984).
[10] A. A. CHILINGARIAN, S. KH. GALFAYAN *et al.*: *Upper boundary of iron nuclei fraction in* PCR, Preprint FIAN 75, 1988.
[11] A. A. CHILINGARIAN and H. Z. ZAZYAN: *A classification method of* PCR *mass composition determination*, Preprint EPI 1204(81), 1989.
[12] A. A. CHILINGARIAN and H. Z. ZAZYAN: *Nonparametric method of hadron energy estimation*, Preprint EPI 1217(3), 1990.
[13] N. M. NIKOLSKAYA and I. N. STAMENOV: *The Tien-Shan* EAS *installation data han g algorithm investigation*, Preprint FIAN 125, 1975.
[14] YU. M. SHABELSKY: *Cross-sections and spectrums of secondary particles in hadron-nuclei collisions*, Preprint LIAF 1221, 1986.
[15] A. M. DUNAEVSKY and A. V. URISON: *Scaling, cross-section increase and simulation of electron-nuclear cascades in the athmosphere*, Preprint FIAN 150, 1975.
[16] A. V. PLYASHESHNIKOV, A. K. KONOPELKO and K. V. VOROBYEV: *The three-dimensional development of high energy electromagnetic cascades in atmosphere*, Preprint FIAN 92, 1988.
[17] S. R. RAO: *Cluster analysis as applied to study of race mixing in human populations*, in *Classification and Clustering* (Academic Press, New York, N.Y., London, 1977).
[18] L. DEVROY and L. GYORFI: *Nonparametric Density Estimation. The* L1 *View* (John Wiley & Sons, New York, N.Y., 1985).
[19] E. E. LEAMER: *Ad Hoc Inference with Nonexperimental Data* (John Wiley & Sons, New York, N.Y., Chichester, Brisbane, Toronto, 1978).
[20] A. A. CHILINGARIAN: *Comp. Phys. Commun.*, 54, 381 (1988).
[21] C. T. TOUSSAINT: *IEEE Trans. Inf.*, IT-20, 472 (1974).
[22] A. A. CHILINGARIAN and S. CH. GALFAYAN: *Calculation of Bayes Risk by KNN Method. Stat. Problem of Control* (Vilnius, 1984), p. 66.
[23] J. D. HEY: *An Introduction to Bayesian Statistical Inference* (Martin Robertson, 1983).
[24] A. A. CHILINGARIAN, S. KH. GALFAYAN *et al.*: *Multidimensional analysis of* EAS *and Roentgen-emulsion data*, Preprint FIAN 332, 1986.
[25] B. EFRON: *The Jackknife Bootstrap and other Resempling Plans* (Society for Industrial and Applied Mathematics, Philadelphia, Penn., 1982).
[26] V. N. VAPNIK: *Dependence Reconstruction by Empiric Data* (Nauka, Moscow, 1974).
[27] P. C. MAHANALOBIS: *Proc. Natl. Inst. Sci. India*, 12, 49 (1963).
[28] T. M. COVER: *IEEE Trans. Inf. Theory*, IT-14, 50 (1968).
[29] A. P. DEVROYE: *IEEE Trans. Inf. Theory*, IT-24, 142 (1978).