PREDICTION OF FLARE ACTIVITY OF STELLAR AGGREGATES. I. THEORETICAL PART*

M. A. Mnatsakanyan and A. L. Mirzoyan

The problem is posed of predicting the number $n_k(t)$ of flare stars that have exhibited precisely k flares by the time t on the basis of data on these quantities known during the total time T of observations of the aggregate. The problem posed by Ambartsumyan [3] of determining the distribution function $f(\nu)$ of the true frequency of stellar flares from known chronology of these data is equivalent to the limiting form of our formulation — prediction in the future over an infinitely long time. An exact analytic solution of the problem obtained without any assumption about the function $f(\nu)$ is given. It permits prediction of the steady flare activity of the aggregate into both the future and the (known) past. It follows from this solution that prediction into the future is in principle impossible to times that exceed the doubled time 2T of the available observations (this means that the problem of determining the function $f(\nu)$ cannot be solved). Moreover, because of the unavoidable fluctuations in the observational data $n_k(T)$, such prediction is limited to even shorter times, and these are shorter the larger the value of k. Prediction into the past and into the future on the basis of the data $n_k(T)$ at the present time and its possible errors due to small fluctuations in these data are illustrated for the examples of the Pleiades and the Orion aggregate.

1. Introduction

Many-sided investigations of flare stars in stellar aggregates (Pleiades, Orion, etc.) have already made it possible to draw a number of important conclusions about the part they play in the evolution of stars and the stellar systems containing them (see [1]).

The principal conclusion is that almost all stars in young stellar aggregates fainter than a certain limiting magnitude (which varies from aggregate to aggregate) are flare stars. Intimately related to this is the problem of statistical estimation of the total number of flare stars in aggregates, including stars not yet discovered [2]. With the accumulation of observational data the estimates of the numbers for individual aggregates have increased systematically, and it was natural to attribute this to the idealization on which the estimates were based, namely, that of an approximately equal frequency of flares. The idea therefore took hold that there were two possible frequencies, and then three or four, around which the flare stars were grouped on the basis of their "true" flare frequencies.

The unsatisfactory description of the observational data by means of several discrete frequencies dictated, and the accumulation of sufficient statistical material enabled Ambartsumyan already in 1978 [3] to undertake it, the posing of a more general and complicated problem: the determination of the distribution function of the "true" frequencies of the flare stars, the problem being posed in this case on the basis of not only data relating to the end of the epoch of observation but also the "chronology" of these data, i.e., their behavior in time. The approximate representation of the distribution function of the flare stars with respect to frequencies in the form of an un-normalized gamma distribution for the Pleiades [3] and Orion [4] was the solution of this problem.

*Dedicated to V. A. Ambartsumyan on the occasion of his 80-th birthday.

Byurakan Astrophysical Observatory. Translated from Astrofizika, Vol. 29, No. 1, pp. 32-43, July-August, 1988. Original article submitted March 30, 1988; accepted for publication April 28, 1988.

0571-7256/88/2901-0423$12.50

The analysis that we have made of this inverse problem shows that it is improperly posed to a high degree, and therefore at the given stage of the observations the determination of the frequency distribution function $f(\nu)$ of stellar flares is extremely difficult. Instead of this, it appears to us, one should make a restriction to a more modest (but, at the same time, more general!) formulation of the problem: the prediction in time of the flare activity of the aggregate, the problem of determining $f(\nu)$ being avoided. Our formulation of the problem is as follows: from the known observational data $n_k(t)$, the numbers of stars of the aggregates that by the time t of observations have exhibited precisely k flares, to determine these same numbers for future times, i.e., predict the behavior of the numbers at future observation times.

Although the new formulation is not so improperly posed, it still has that shortcoming, and a rigorous analysis reveals a number of features that are also inherent in the original formulation. In the first place, this concerns the question of the correct theoretical description of the chronology of the observational data. In this paper we give theoretical expressions (see also [5]) that describe the analytical behavior of the numbers $n_r(T)$ in the past and in the future on the basis of the data $n_k(T)$ at the present time, and for the examples of the Pleiades and the Orion aggregate we illustrate the possible errors of prediction due to fluctuations of the observational data. Subsequently we shall compare them with the chronology of the observational data and attempt to take into account selection effects that distort the behavior of these quantities in the process of the observations.

2. Formulation of the Prediction Problem

Let N be the total number of flare stars in the considered aggregate, and $f(\nu)$ be the distribution of these flare stars with respect to the "true" flare frequencies, so that

$$N = \int_0^\infty f(\nu)\, d\nu.$$

Because of insufficient statistics of the observational data, the observed flare frequencies, i.e., the set of numbers $n_k(T)$, which indicate the numbers of stars that have exhibited precisely k flares during the complete epoch of observations, do not directly determine the function $f(\nu)$ and in their behavior (dependence on k) can deviate strongly from $f(\nu)$. If we had very good statistics in time, i.e., if we had observed the aggregate for such a long time T that all flare stars would have exhibited sufficiently large numbers of flares, then at large T the numbers $n_k(T)$ would approach the numbers of stars $f(\nu)d\nu$ having true flare frequencies in the interval $\Delta\nu \to \left(\frac{k}{T}, \frac{k+1}{T}\right)$, and in the limit we should have the exact correspondence

$$n_k(T) \underset{T\to\infty}{\to} f\left(\frac{k}{T}\right)\cdot\frac{1}{T}. \qquad (1)$$

But if the time of observations T is not so great, the numbers on the right- and left-hand sides of the relation (1) will differ appreciably. This in fact is the reason why the inverse problem of finding the distribution of the "true" flare frequencies arises.

For a variety of reasons, noted in [3], flares in different flare stars can be assumed to occur independently of each other, while the circumstance that the observations are made intermittently ensures to an adequate degree the condition of mutual independence of successively detected flares of a given star.

A second assumption concerns the condition of stationarity of the ensemble of flare stars of the aggregate as a whole during the complete epoch of observations (which is only a few decades); this is expressed by time-independence of the unknown distribution $f(\nu)$.

In the framework of such assumptions we can, following Ambartsumyan, assume that we are presented with a stationary Poisson process, and for the number of stars that exhibit by the time of observations t precisely r flares we can write

$$n_r(t) = \int_0^\infty f(\nu) \frac{(\nu t)^r}{r!} e^{-\nu t} d\nu, \qquad r = 0, 1, \ldots \tag{2}$$

As already adopted, we take as the current time t the sum of the exposures of the observations. If the flare activity of the aggregate is stationary, we can formally replace the time by the current number of the flare in the general chronological catalog of all observed flares of the aggregate, i.e., we can count the time by the number of detected flares. This assumption is very natural (during equal intervals of observational time the same number of flares is detected in the aggregate) and eliminates the technical difficulties of taking into account the real time of the observations of the aggregate.

We now turn to Ambartsumyan's formulation — the problem of determining the frequency distribution function $f(\nu)$. Suppose we were able to solve it and therefore knew this function. Then, using the relation (1), we could find the numbers $n_r(t)$ for all instants of time t, doing this, moreover, not only for observations relating to the past epoch but also for all times of future observations. That is, we should also know the behavior of $n_r(t)$ in the future, to infinitely long times $t \to \infty$, on the basis of data of observations available at the observation epoch T. Thus, the formulation of the problem — the prediction of all $n_r(t)$ to infinity — is completely equivalent to the problem of finding the function $f(\nu)$.

Conversely, ability to predict to the infinite future the value of $n_r(t)$ is equivalent to knowledge of the function $f(\nu)$, as follows from the clarifications of Eq. (1).

In view of the exceptional complexity of the formulation, we shall restrict ourselves below to a more modest formulation of the problem, namely, the prediction of $n_r(t)$ into the future, not over an infinitely long time interval, but only so far as is possible within the permitted errors. This formulation is in fact more general: If it is possible to predict $n_r(t)$ to infinity, then the problem of determining $f(\nu)$ will also be solved; but if that is not possible, determination of the function $f(\nu)$ will also be impossible. We shall establish below that such prediction is in principle possible no further than to times $t < 2T$, i.e., times that do not exceed twice the time of the existing observations.

It is clear that the problem of prediction can be solved only analytically, and for this we must above all know how to describe sufficiently well — and, moreover, analytically — the behavior of $n_r(t)$ in the past, $t \leq T$, and then continue such description to future times $t > T$.

4. Description of the Past

Suppose that at the present time T we know the numbers $n_k(T)$, the numbers of flare stars that during the complete epoch of observations T have exhibited precisely k flares. We pose the problem of determining these quantities $n_r(t)$ at times $t < T$, i.e., the numbers of flare stars that have exhibited by the current time t of observations precisely r flares. This problem can be solved relatively easily under the assumptions made above of stationarity of the aggregate and mutual independence of individual flares.

Now an already realized Poisson process leads to a uniform distribution of events. Therefore, if for one star precisely k flares have been observed during the time t, these k flares must be distributed uniformly (of course, randomly) over the interval (0, T).

We first derive an expression that describes the behavior of $n_0(t)$, the number of stars of the aggregate that up to the current time t have not exhibited a single flare. For this, we consider an individual star that during the time T has exhibited the given number k of flares.

If this star does not have a single flare in the interval (0, t), all of its k flares must occur during the time interval (t, T). The probability of an individual flare in this interval is $1 - t/T$, and for all k flares it is $(1 - t/T)^k$.

Therefore, the mathematical expectation of the number of stars that up to the time t have not exhibited a single flare as a proportion of the number $n_k(T)$ of stars that by

the time T have exhibited precisely k flares is $n_k(T)(1 - t/T)^k$, for each value of k = 0, 1, 2,... (for k = 0 this assertion is trivial).

Since we are interested in the number of stars $n_0(t)$ irrespective of the multiplicity of flares of the flare stars, i.e., among all flare stars known or unknown at the present time (T), we must sum the expression we have obtained over all values of k:

$$n_0(t) = \sum_{k=0}^{\infty} n_k(T)\left(1 - \frac{t}{T}\right)^k. \tag{3}$$

Similar arguments lead us to an expression that describes the behavior in time of the number $n_r(t)$ of stars that by the time t have exhibited precisely r flares. A flare that during the time T has precisely k flares exhibits during the interval (0, t) precisely r flares ($r \leq k$) with probability determined by the binomial distribution

$$C_k^r\left(\frac{t}{T}\right)^r\left(1 - \frac{t}{T}\right)^{k-r}.$$

The mathematical expectation of the corresponding number of stars is determined by

$$n_r(t) = \sum_{k=r}^{\infty} n_k(T)\, C_k^r\left(\frac{t}{T}\right)^r\left(1 - \frac{t}{T}\right)^{k-r}, \quad r = 0, 1, 2,... \tag{4}$$

For r = 0, this expression goes over into the earlier one.

The expression (4) can also be directly deduced from (3) by r-fold differentiation with allowance for the known relation

$$n_r(t) = (-1)^r \frac{t^r}{r!}\frac{d^r}{dt^r} n_0(t), \tag{5}$$

which expresses $n_r(t)$ in terms of $n_0(t)$, The relation (5), in its turn, can be directly verified by successive differentiation of Eq. (1).

The relation (3) cannot be used in practice, since its right-hand side contains, under the summation sign, the unobservable quantity $n_0(T)$, the number of flare stars of the aggregate that are not detected during the entire time of the observations. We therefore transform it to a different form that enables us to determine the behavior in time of the total number of flare stars found by the current time t, denoting this number by n(t). By virtue of the obvious relations

$$N = n_0(t) + n(t) = n_0(T) + n(T) = n(\infty)$$

it follows from (3) that

$$n(t) = n(T) - \sum_{k=1}^{\infty} n_k(T)\left(1 - \frac{t}{T}\right)^k.$$

Since $n(T) = \sum_{k=1}^{\infty} n_k(T)$, for the chronology of n(t) we finally have

$$n(t) = \sum_{k=1}^{\infty} n_k(T)\left[1 - \left(1 - \frac{t}{T}\right)^k\right], \tag{6}$$

The result (6) can also be obtained by summing the expression (4) over all values of r = 1, 2,... . For t = T, the expressions (4) and (6) become identities: $n_r(T) = n_r(T)$, since $\lim_{t \to T}\left(1 - \frac{t}{T}\right)^{k-r} = \delta_{kr}$.

The results (4) and (6) are what we require. They describe the theoretical behavior in time (the chronology) of all the numbers $n_r(t)$ in the past for $t \leq T$ on the basis of the given values $n_k(T)$ of these quantities at the present time. The analytic behavior in the time given by (4) and (6) is completely exact. Therefore, the problem reduces entirely to accurate determination of the values of the numerical coefficients $n_k(T)$, which are known from observations, with allowance for fluctuations and possible selection factors.

We make a remark concerning the limiting form of our (4), which describes the $n_r(t)$ chronology. As we already noted in Sec. 2, at long observation times T the set of numbers $n_k(T)$ approaches the continuous distribution $f(\nu)$ for values of ν in the interval $\left(\frac{k}{T}, \frac{k+1}{T}\right)$. With increasing T, the numbers $n_k(T)$ must tend to zero if N except for quantities with large values of k satisfying $k \gtrsim \nu \cdot T$. If in the expression (4) we go to the limit $T \to \infty$, $k \to \infty$ but in such a way that the ratio k/T keeps the definite value ν, we arrive at the expression (2). We here use the well-known limit relations

$$\left(1-\frac{t}{T}\right)^k = \left(1-\frac{t}{T}\right)^{\nu T} \underset{T\to\infty}{\to} e^{-\nu t}, \quad C_k^r = \frac{k!}{r!(k-r)!} \underset{k\to\infty}{\to} \frac{k^r}{r!}.$$

Thus, purely formally the expression (2) is the limiting form of our expression (4) as $T \to \infty$, i.e., for infinitely long intervals of time, this being an assumption that, strictly speaking, is also contained in the expression (2), since in practice the specification of the function $f(\nu)$ in accordance with the relation (1) is associated with an infinitely long time of observations of the aggregate. In this connection one might get the impression that (2) is approximate, but this is not the case. Both expressions are exact and are different representations: (2) is in terms of the unknown function $f(\nu)$, while (4) is in terms of the known $n_k(T)$.

5. Prediction Formulas

The relation (4) (the relations (3) and (6) are consequences of it) describes the behavior of the numbers $n_r(t)$ in the past on the basis of the $n_k(T)$ given at the present time. We pose this question: From the $n_k(T)$ given at the present time can we determine the behavior of $n_r(t)$ for future times $t > T$?

The answer to this question is in the affirmative. Moreover, and this is the most interesting result, the corresponding prediction formula is identical to the relation (4) that describes the past with the only difference that in it one must take $t > T$. In other words, formula (4) describes the behavior of $n_r(t)$ in both the past and the future.

This assertion can be proved in different ways. For example, one can use the Bayes approach to estimation of the probabilities of hypotheses. One can regard the expressions (4) as a system of linear equations for $n_k(T)$ for given $n_r(t)$ with fixed time $t < T$ with t taken to be the present time and T an arbitrary time in the future. The solution of this system is given by an expression like (4):

$$n_r(T) = \sum_{k=r}^{\infty} n_k(t)\, C_k^r \left(\frac{T}{t}\right)^r \left(1-\frac{T}{t}\right)^{k-r}, \quad r = 0, 1, 2, \ldots$$

as can be shown by directly substituting this solution in (4).

The simplest derivation is as follows. The number of stars that have precisely r flares during the time $t = (T + t_1)$ is determined by the expression (2). We consider first the case $r = 0$:

$$n_0(t) = n_0(T+t_1) = \int_0^{\infty} f(\nu)\, e^{-\nu(T+t_1)} d\nu.$$

Using the expansion for $\exp(-\nu t_1)$, we rewrite $n_0(T + t_1)$ in the form

$$n_0(T+t_1) = \sum_{j=0}^{\infty} (-1)^j \frac{t_1^j}{j!} \int_0^{\infty} e^{-\nu T} f(\nu)\, \nu^j d\nu =$$

$$\sum_{j=0}^{\infty} (-1)^j \left(\frac{t_1}{T}\right)^j \int_0^{\infty} f(\nu)\, e^{-\nu T} \frac{(\nu T)^j}{j!} d\nu = \sum_{j=0}^{\infty} n_j(T) (-1)^j \left(\frac{t_1}{T}\right)^j.$$

But $t_1 = t - T$, $t > T$, and therefore we finally have

$$n_0(t) = \sum_{k=0}^{\infty} n_k(T) \left(1-\frac{t}{T}\right)^k, \text{ or } n(t) = \sum_{k=1}^{\infty} n_k(T) \left[1-\left(1-\frac{t}{T}\right)^k\right]. \tag{7}$$

This formula, derived for $t > T$, is identical to (3), which is valid for $t \leq T$, i.e., it is valid for all $t \geq 0$. Therefore, the relation for $n_r(t)$ obtained from it by differentiation in accordance with (5) is also valid for all $t \geq 0$. This relation is

$$n_r(t) = \sum_{k=r}^{\infty} n_k(T) C_k^r \left(\frac{t}{T}\right)^r \left(1 - \frac{t}{T}\right)^{k-r}, \quad r = 0, 1, \ldots \tag{8}$$

Actually, this last relation can be derived from (2) similarly by a series expansion of $e^{-\nu t}$ (just like (3) and (4)).

The prediction formulas — they predict the future and the past solely on the basis of data at the present time — possess a number of interesting and important properties.

a) In practice we deal with a series $n_k(T)$ that terminates, and therefore for $t > 2T$ all the expressions $n_r(t)$ become very large in absolute magnitude (this behavior of $n_r(t)$ is due to the last nonvanishing term of the series, which in the limit $t \to \infty$ becomes the principal term). This mathematical remark means that in principle prediction of the behavior of $n_r(t)$ is impossible over times that exceed the present time of observations by two times or more.

b) If we substitute instead of $n_k(T)$ the Poisson distribution

$$n_k(T) = N \frac{(\nu T)^k}{k!} e^{-\nu T}, \quad k = 0, 1, \ldots,$$

then for the quantities $n_r(t)$ we again obtain the Poisson distribution

$$n_r(t) = N \frac{(\nu t)^r}{r!} e^{-\nu t}$$

for each instant of time t. If $n_k(T)$ is a superposition of Poisson distributions, then $n_t(t)$ is the same superposition of Poisson distributions. If in the limit

$$n_k(T) = \int_0^{\infty} f(\nu) \frac{(\nu T)^k}{k!} e^{-\nu T} d\nu,$$

then

$$n_r(t) = \int_0^{\infty} f(\nu) \frac{(\nu t)^r}{r!} e^{-\nu t} d\nu.$$

c) There is a group property (which is a consequence of the more general group property [6] of nonlinear prediction formulas), and it takes the following form. From given $\{n_r\}$ at the time t_1 it is possible to determine the values of n_r at the time t_2 and from these data at the time t_2 in turn determine n_r at the time t_3. This is equivalent to determination of $\{n_r\}$ at the time t_3 directly from the $\{n_r\}$ given at the time t_1. In other words, step by step advance in time in the prediction problem is as a result equivalent to prediction in one large step. Making the steps shorter does not lead to any improvement (in fact it can lead to an accumulation of computing errors).

d) A further distinctive property of the prediction formulas obtained above is their linearity in the quantities $n_k(T)$. The data at the present time (as at any time) contain natural fluctuations, and to improve the prediction it is desirable to eliminate them. By virtue of the linearity, it is possible, using data on $n_r(t)$ over the entire past interval (0, T), to determine by linear regression methods the best estimates for $n_k(T)$, and only then predict the future on their basis.

e) Such regression must also make it possible in principle to predict the series $n_k(T)$, i.e., extend it to larger values of k as yet unknown from observations, since prediction in time is clearly associated with more accurate determination of the terms of the series (4).

6. Conclusions

The main conclusion that can already be drawn is that prediction of the flare activity of an aggregate is impossible to times that exceed 2T. This, in its turn, means that

TABLE 1

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	$n(T)$
Pleiades	290	93	46	29	22	22	9	10	7	5	5	4	1	1	1	545
	290	93	46	29	22	18	13	10	7	5	5	4	1	1	1	545
Orion	379	76	23	7	1	1	2									489
	379	76	23	7	1	1	1	1								489

the inverse problem of determining the frequency distribution function cannot, in principle, be solved without additional assumptions about the form of this function (otherwise prediction to the infinite future would be possible). It is true that this assertion holds under the condition that the time of observations T is not so great as to enable one to assume that all the flare stars of the aggregate have exhibited sufficient numbers of flares and the expression (1) can be applied directly to them.

In Fig. 1 we show the curves n(t) and $n_r(t)$ calculated in accordance with the prediction formulas (7)–(8) for times t that describe both the past (t < T) and future (t > T) behavior of these quantities solely on the basis of observational data n(T), $n_k(T)$ known at the present time (t = T) for the Pleiades and the Orion aggregate (the data for time T are indicated by the black circles and are given in the corresponding upper rows of Table 1). The curves are indiated in Fig. 1 by the arrows.

To obtain an approximate idea of the possible errors of prediction due to the unavoidable fluctuations in the observational data $n_k(T)$, we changed the data slightly (the changed data are given by the bold numbers in the corresponding lower rows of Table 1). The corresponding prediction curves begin to deviate strongly from those calculated from the true observational data already at times less than 2T, doing this moreover more rapidly the larger the value of k (this can be easily understood, since the series (8) contains a "tail" of data $n_k(T)$ with $k \geq r$). The hatched regions in the figure show the corridor of possible errors. The deviation in the prediction curves for Orion is due to just a single, most probable (!) additional flare, that is, if one

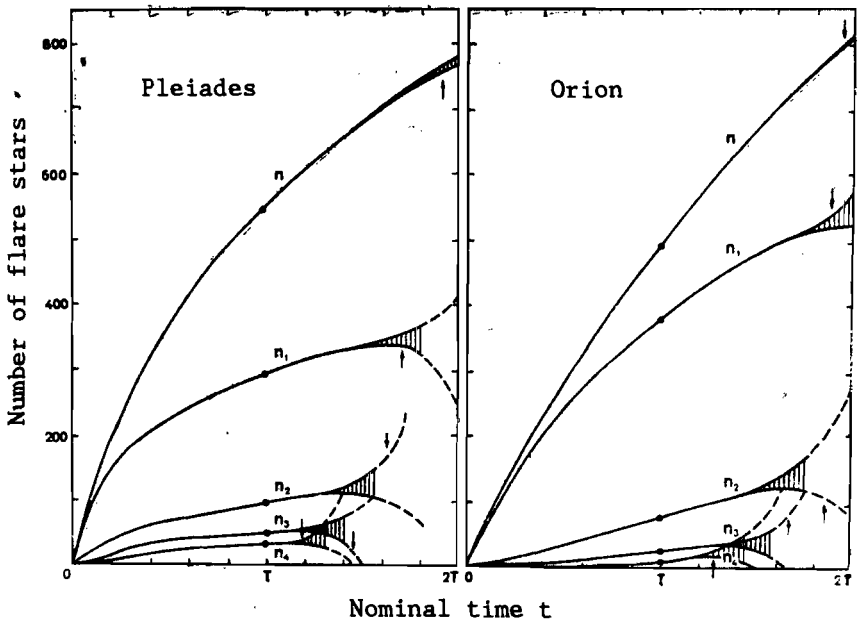

Fig. 1. Theoretical time dependence, calculated in accordance with Eqs. (8), of $n_r(t)$, the number of flare stars that have exhibited precisely r flares by the nominal time t the curve n(t) is the total number of flare stars found during the time t. The points corresponding to the value t = T represent the observational data at the present time.

of two stars that have exhibited seven flares each suddenly exhibits one more flare.

The most reliable prediction – almost to 2T – is realized for the number of flare stars, n(t), which statistically is the richest quantity and therefore least subject to fluctuations. Its behavior must asymptotically approach the estimate of the total number of flare stars in the aggregate.

In the following paper we shall make a comparison of the theoretical curves with the observational data $n_r(t)$ relating to the epoch of observations (0, T) with a view to the possible improvement of the numerical values of $n_k(T)$ by linear regression methods. Here we emphasize once more that as regards the form of the distribution function $f(\nu)$ of the flare stars with respect to the true frequencies no assumptions were made at all apart from the condition of its being independent of the time.

LITERATURE CITED

1. L. V. Mirzoyan, Nonstationarity and Evolution of Stars [in Russian], published by the Armenian Academy of Sciences, Erevan (1981).
2. V. A. Ambartsumyan, in: Stars, Nebulas, Galaxies [in Russian], Erevan (1969), p. 283.
3. V. A. Ambartsumyan, Astrofizika, 14, 367 (1978).
4. É. S. Parsamyan, Astrofizika, 16, 677 (1980).
5. M. A. Mnatsakanyan, Astrofizika, 24, 621 (1986).
6. M. A. Mnatsakanyan, in: Symposium "Renormalization-Group-86", D2-87-123 [in Russian], Dubna (1987), p. 376.