Development of a statistical forecasting method using the example of hit parades

Kornilov Vladimir Sergeevich
Engineer
Closed Joint Stock Company "Prochnost"

Abstract. The article describes a forecasting method that allows you to determine the position of an object in the rating. The method is based on the theory of mathematical statistics without the use of artificial intelligence (neural networks). The proposed method is illustrated by a representative sample of Russian charts collected by the author from 1999-2020.

Keywords: data array, rating units, compression ratio.

Introduction

All existing forecasting methods are usually divided into intuitive and formalized [1, 2]. Formalized methods are divided into domain models and time series models. In the modern world, neural networks related to time series models are increasingly used for forecasting [3]. But neural networks have a number of disadvantages: arbitrary assignment of weight coefficients, the absence of a theory for choosing a network architecture, the requirement to obtain an adequate result of a large amount of data, close to the general population of the phenomenon. Therefore, despite attempts at widespread use and aggressive advertising, neural networks are not an uncontested forecasting method [4, 5]. The article describes the forecasting method proposed by the author and is illustrated with examples based on the statistics of hit parades.

Purpose of the study - to develop a forecasting method that allows to determine the position of an object in the rating based on a statistical sample of hit parades.

Materials and methods

Statistical data of Russian hit parades from 1999 to 2020 collected by the author were used as a representative sample required for the statistical model (research material). The theory of mathematical statistics is a research method.

Results and discussion

Let's take the following designations: "song" is a set of places that a song has visited during its time in the hit parade. From the point of view of mathematics, this is a set (array) of positive integers ranging from 1 to *m*, where *m* is the number of places in the hit parade. A song being in the hit parade can change its positions within the specified range. An example of representing a song in the form of an array in general form $[m_1\ m_2\ m_i]$, where m_i – are the places visited by the song while in the hit parade, with $1 \leq m_i \leq m$. A number of parameters can be obtained from the recording of a song as an array of places it has visited. External parameters of the song: *n* - the number of weeks spent by the song in the hit parade. In terms of mathematics - the number of numbers in the array, peak (*p*) is the highest place the song has climbed to (the smallest number from the array). When recording the

external result, the number of weeks spent by the song in the hit parade is indicated in brackets, as well as the peak of the song and the number of weeks spent by the song at the peak, thus *np(n)* - weeks in the hit parade, peak (the number of weeks at the peak).

Song rating.

A song's rating (*S*) is a function of the number of weeks a song has been on the charts and where the song has held: $S = f(n, m_n)$.

The rating is calculated using the main formula of the hit parade.

The main formula of the hit parade.

Let's assume that 1 point is given for one week spent by a song in the last place of the charts, and *m* points are given for a week spent in the first place. The dependence of the rating on the place $S_n = S_n(m_n)$ – is linear, its graph will be a straight line passing through points (1;*m*) and (*m*;1) (see fig. 1).

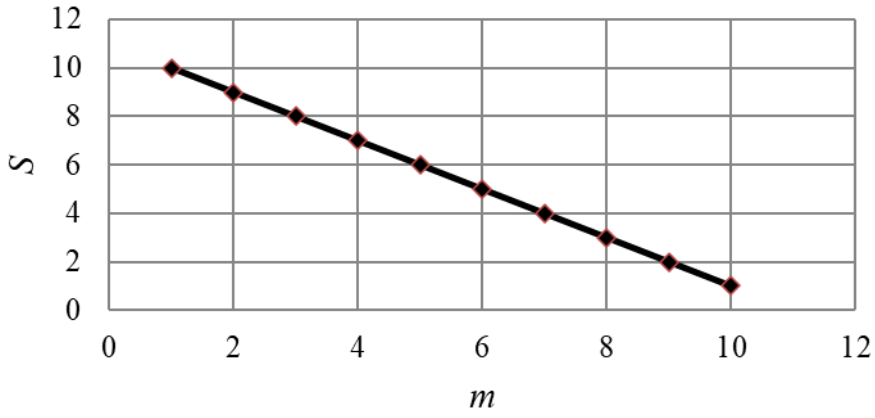

Fig. 1. Dependence of the rating on the place in the hit parade

The equation of the straight line in coordinates (*m;S*) will have the form

$$S_n(m_n) = km_n + b$$

Angular coefficient *k*

$$k = \frac{\Delta S_n}{\Delta m_n} = \frac{m-1}{1-m} = -1$$

Define **b** from the boundary conditions:

$$S_n = 1 \text{ at } m_n = m,\ 1 = -m + b \rightarrow b = m + 1$$

The dependence of the rating on the place will look like

$$S_n(m_n) = m + 1 - m_n$$

A song's rating is the sum of the ratings for the places the song has visited, i.e.

$$S = \sum S_n = m + 1 - m_1 + m + 1 - m_2 + \cdots + m + 1 - m_n = (m+1) \cdot n - \sum m_n$$

Then the main formula of the hit parade is

$$S = (m+1) \cdot n - \sum m_n$$

$\sum m_n$ – sum of places the song has visited.

The restriction imposed on the main formula of the hit parade: the position number m_n - is a positive integer, the number of weeks *n* - is also a positive integer. Accordingly, the song's *S* rating is also a positive integer.

Rating units.

1 point is the main and minimum unit of measurement. It is the rating of the song that held the last place in the hit parade for one week. In other words, 1 point = 1 week × 1 place. The song rating, calculated using the basic formula, is measured in points and is numerically equal to the area under the graph of the song's rankings, built in the form of a bar chart in coordinates (*n;m*).

1 week of leadership (WL) - the rating of a song that stayed in the charts for one week in first place. 1 WL= *m* points.

The procedure for transferring the rating from points to leadership weeks.

1. The rating of the song in points is divided by the number of places in the hit parade *m*.
2. Round off the resulting value to the nearest lower whole value, the resulting value will be the number of weeks of leadership without taking into account the "remainder" in points.
3. Multiply the result obtained by the number of places in the hit parade *m*.
4. Subtract the resulting value from the result of the song in points - you will get the "remainder" of the result in points.
5. Write down the answer in the form of WL + points.

The rating of a song, expressed in weeks of leadership, allows you to determine the minimum number of weeks in which it can be recruited. If the rating of a song is equal to any number of WL without a "remainder" in points, this number is the minimum number of weeks required to gain this rating. If the rating is equal to any number of WL+ any number ("remainder") of points, the minimum number of weeks is equal to the number of WL+1.

1 set (H) - the rating of the song that stayed in the hit parade for *m weeks* and visited each place once. It is the largest rating unit.

The rating of such a song is numerically equal to the sum of integers from 1 to m, and can be calculated using the arithmetic progression formula.

$$\mathrm{H} = \frac{1+m}{2} \cdot m, \text{points}$$

Relationship between rating units.

$$1 \text{ point} < 1 \text{ WL} < 1H; \;\; 1 \text{ WL} = m, \text{points}; \; 1\mathrm{H} = \frac{1+m}{2} \cdot m, \text{points} = \bar{m}m, \text{points}; \;\; 1H = \bar{m} \text{ WL}$$

The complete result of the song.

$n\ p(n)\ S -$ weeks in the hit parade, peak (number of weeks at peak), song rating. Example: song [5 4 2 1 1 2 2 6 8 10], $m = 13$. The outer result of the song is 10 1 (2). The complete result of the song is 10 1 (2) 99. The rating of the song, according to the main formula of the hit parade

$$S = (m+1)n - \sum m_n = (13+1)\cdot 10 - (5+4+2+1+1+2+2+6+8+10) =$$
$$= 140 - 41 = 99 \text{ points.}$$

An example of calculating the rating of a song in different values: $S = 358$ points; $m = 13$; 1 WL = 13 points; $H = \frac{1+m}{2}\cdot m = \frac{1+13}{2}\cdot 13 = 91$ points; in leadership weeks $S = \frac{358}{13} = 27{,}53 \rightarrow 27$ WL; $27 \cdot 13 = 351$ points $\rightarrow 358 - 351 = 7$ points; $S = 358$ points $= 27$WL $+ 7$points; in sets $S = \frac{358}{91} = 3{,}93 \rightarrow 3$H; $3 \cdot 91 = 273 \rightarrow$ $358 - 273 = 85$ points $= \frac{85}{13}$ WL $= 6$ WL $+ 7$ points; $S = 358$ points $= 3$H $+ 6$ WL $+ 7$ points.

The indirect parameters of the song, discussed below, can be obtained from the full result: Average score $(\bar{S})$ – rating divided by the number of weeks spent by the song in the hit parade

$$\bar{S} = \frac{S}{n} = \frac{(m+1)n - \sum m_n}{n} = m + 1 - \frac{\sum m_n}{n} = m + 1 - \bar{m}.$$

Average place $(\bar{m})$ - is the sum of all places on which the song held, referred to the number of weeks spent by the song in the hit parade:

$$\bar{m} = \frac{\sum m_n}{n} = m + 1 - \bar{S}.$$

The average score and the average place of a song can take, in contrast to the rating and place, fractional values, i.e. when calculating the specified parameters, the restriction imposed on the main formula of the hit parade does not apply. At the same time, the average place and the average score, as well as the rating of the song, are positive values.

Compression ratio (*k*) - the ratio of the number of weeks spent by a song in the hit parade to the minimum number of weeks for which a rating of a given song can be gained:

$$k = \frac{n}{n_{min}}.$$

How the compression ratio is calculated.

1. Divide the song's rating by the number of hits in the m.
2. Round the resulting value up to the nearest whole value. The result is the minimum number of weeks for which you can get this rating. In other words, to determine the minimum number of weeks, it is necessary to calculate the song's rating in leadership weeks (see above).
3. Divide the number of weeks the song spent in the hit parade by the value obtained in ex. 2. The resulting value is the compression ratio.

Compression Ratio Properties: The compression ratio is dimensionless by definition. In any hit parade, the minimum value of the compression ratio is equal to one, and the maximum value is equal to the number of places in the hit parade. That is, $1 \le k \le m$. In this case, the compression ratio can take fractional values.

Proof. *The first critical case* - the song stayed in the charts for n weeks in first place. The rating of such a song is $S = (m + 1 - 1) \cdot n = mn$ points. The minimum number of weeks for which you can get the *mn* points rating is: $n_{min} = \frac{S}{m} = \frac{mn}{m} = n;$

$k = \frac{n}{n_{min}} = \frac{n}{n} = 1.$

The second critical case - the song lasted n weeks at the last place of the charts. The rating of such a song $S = (m + 1 - m) \cdot n = n$ points. The song stayed on the charts for $n = n$ weeks. The minimum number of weeks for which you can score n points is $n_{min} = \frac{S}{m} = \frac{n}{m}$. Compression ratio $k = \frac{n}{n_{min}} = \frac{n \cdot m}{n} = m$. The graph (fig. 2) shows the change in the compression ratio for a song that lasted n weeks at the last place of the charts for the case *m*=13. As can be seen from the graph, the compression ratio after a certain period *T*, measured in weeks, reaches its maximum value *m*. The number of the week at which the value of the compression ratio reaches its maximum can be determined by the formula $n_m = m + TZ$, where $Z \ge 0 - \text{any integer}$.

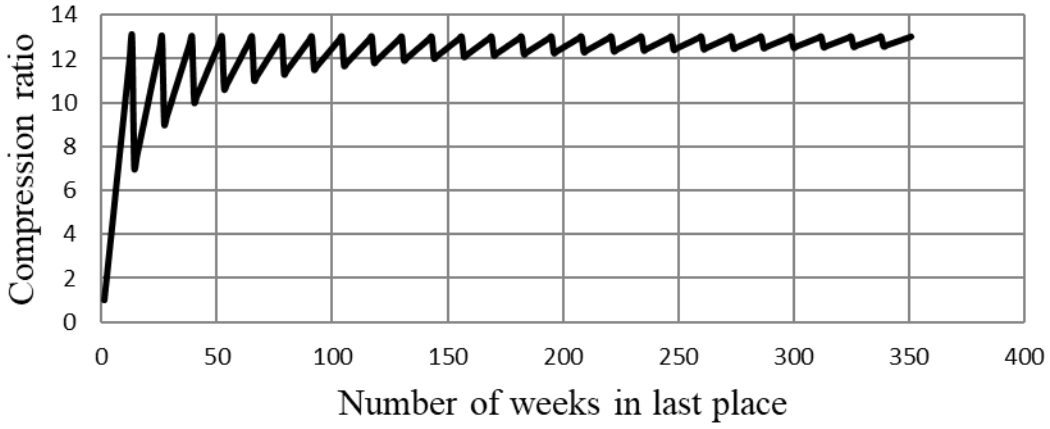

Fig. 2. Changing the compression ratio for a song that lasted n weeks at the last place of the charts (*m*=13)

The "lower" value of the compression ratio changes with the "conditional" period $T = m$. The number of the week in which the compression ratio falls to the "lower" value is determined by the formula $n_n = 1 + mZ = 1 + TZ$. The "lower" value of the compression ratio corresponding to the week defined above is determined by the formula:

$$k_n = \frac{1 + mZ}{Z + 1} = \frac{1 + TZ}{Z + 1} = \frac{n_n}{Z + 1}.$$

At the initial moment, when $Z = 0$ and $n_n = 1$, the "lower" value of the compression ratio is minimal and equal to one

$$k_n = \frac{n_n}{Z + 1} = \frac{1}{0 + 1} = 1.$$

With an increase in the number of periods *Z*, the "lower" value of the compression ratio will also increase, each time, with each new period, approaching the maximum value of *m*:

$$\lim_{Z \to \infty} \frac{1 + mZ}{Z + 1} = \lim_{Z \to \infty} \frac{\frac{1}{Z} + m}{1 + \frac{1}{Z}} = \frac{0 + m}{1 + 0} = m,$$

which once again confirms that the maximum possible compression ratio is numerically equal to the number of places in the hit parade. If a song has been on the hit parade for *n* = *mZ* weeks, which corresponds to the compression ratio period, and its average score is expressed as an integer, the compression ratio can be determined using the formula below. In this case, the compression ratio will be the maximum possible for a song that lasted any number of weeks *n* in one place m_n of the hit parade:

$$k = k_{max} = \frac{m}{\overline{S}} = \frac{m^2}{S}$$

The compression ratio of the set is in the range of $1 \leq k \leq 2$. In this case, a set means a song that has visited all the places of the hit parade once. The external result of such a song is *m* 1(1). The rating of the set is numerically equal to the sum of integers from 1 to *m*. According to the formula of arithmetic progression

$$H = \frac{1 + m}{2} \cdot m$$

The minimum number of weeks for which you can get this rating is

$$n_{min} = \frac{H}{m} = \frac{1 + m}{2 \cdot m} \cdot m = \frac{1 + m}{2}$$

Dial compression ratio

$$k = \frac{n}{n_{min}} = m : \frac{1 + m}{2} = \frac{2m}{1 + m}$$

Since the number of hits in the hit parade lies within $1 \leq m < \infty$, it is necessary to consider two critical cases.

First critical case $m \to 1$ $\lim\limits_{m\to 1} \frac{2m}{1+m} = \frac{2\cdot 1}{1+1} = 1$

Second critical case $m \to \infty$ $\lim\limits_{m\to \infty} \frac{2m}{1+m} = \lim\limits_{m\to \infty} \frac{2}{\frac{1}{m}+1} = \frac{2}{0+1} = 2$

Thus, for any values of the number of places in the hit parade *m*, the compression ratio of the set lies within: $1 \leq k \leq 2$

Below (fig. 3) is a graph of the dependence of the compression ratio of a set on the number of places in the hit parade.

Fig. 3. The compression ratio of the set from the number of places in the hit parade.

Example. The song "On my Moon" by the "Dead Dolphins" group stayed in the chart of the "Chartova Dozen" $n = 13\,weeks$ and scored $S = 117\,points$. Average song score $\bar{S} = \frac{S}{n} = \frac{117}{13} == 9$ points. The average song score is expressed as an integer, i.e. the rating of the song is equivalent to 13 weeks of being only in the fifth place of the charts, because $\bar{m} = m + 1 - \bar{S} = 13 + 1 - 9 = 5$. The number of weeks spent by a song in the hit parade corresponds to the compression ratio period $n = T = mZ = 13 \cdot 1 = 13$ weeks. Consequently, the compression ratio of this song is the maximum possible for a song that lasted n weeks at one place in the hit parade, in this case - the fifth. $k = k_{max} = \frac{m}{\bar{S}} = \frac{13}{9} = 1{,}444$.

Group parameters.

Group - a collection of songs. Peak of the group (P_A) – the highest place to which the songs of the group rose in the hit parade. The ceiling - the maximum score scored by the band's song. The ceiling is determined by the song and is measured in points. Annual ceiling (C_a)– the maximum score scored by a song of a group within a year, and selected from all the years in which the group was

presented in the hit parade. Factual ceiling (C_F) - the maximum score scored by the group's song in the entire history of the group's stay in the hit parade. Correlation between ceilings: $C_f \geq C_a$. Ceiling difference - the difference between the actual and annual ceiling of the group: $\Delta C = C_f - C_a$; $\Delta C \geq 0$. Ceiling groups can be classified as follows: groups with zero difference in ceilings; groups with non-zero difference in ceilings. In turn, groups with a non-zero difference in ceilings can be divided into: single-ceiling groups - groups in which the annual and actual ceilings are determined by one song $(\text{the number of ceilings } N_C = 1)$, two-ceiling groups - groups for which the annual and actual ceilings are determined by different songs $(\text{number of ceilings } N_C = 2)$. Groups with zero difference in ceilings are single-ceiling.

"Peak-to-ceiling" ratio.

This ratio is represented in the form of the base *R* in the degree of difference between the peak of the song, which determines the group's ceiling in the hit parade, and the peak of the group. That is R^{Δ}, where

$$\Delta = P_C - P_A.$$

If the peak of the group is equal to the peak of the song that defines the ceiling of the group, the difference between the peaks is zero, and this ratio is called "one-to-one", since any number to the zero degree is one. If a $\Delta \neq 0$, this ratio is "ambiguous". In this case, the value of the quantity $\Delta - \text{degree of ambiguity}$. The "peak-to-ceiling" ratio can also be - annual - in relation to the group's annual ceiling, actual - in relation to the actual group ceiling, within the considered year or other period.

Example. Group "Naive". The band's annual and actual ceilings are defined by the song "Memories of Past Love". The song held the charts in 2007 and 2008 and scored 358 points: 245 points in 2007 and 113 points in 2008. The band's annual ceiling $C_A = 245$ points. Actual group ceiling $C_F = 358$ points. Difference of ceilings $\Delta C = C_f - C_A = = 113$ points. The group is single-ceiling ($N_C = 1$), as the annual and actual ceilings are determined by one song.

Example. Spleen group. The group's annual ceiling in the "Chartova Dozen" is determined by the song "My Heart", which scored 263 points and completely stayed in the hit parade within 2001. The actual ceiling of the group is determined by the song "Mayak", which was in the charts in 2007-2008 and scored 270 points, of which 215 points in 2007 and 55 points in 2008. $C_a = 263$ points; $C_f = 270$ points; the difference between the ceilings is $\Delta C = C_f - -C_A = 270 - 263 = 7$ points. The group is two-ceiling $(N_C = 2)$, since the annual and actual ceilings are determined by different songs.

Annual rating parameters.

As a rule, the rating is calculated for the year.

1. The number of songs hitting the hit parade per year.
2. Average score of the year - the sum of the ratings of all the songs that hit the hit parade for the year, referred to the number of songs that hit the hit parade for the year.
3. Overall score - a place in the hit parade, which corresponds to the average score.
4. Number of groups.
5. If the hit parade contains groups from different countries - the number of countries. In this case, the percentage of groups from different countries can be calculated.

Items 2-5 can also be defined for the TOP (for example, TOP-13, TOP-20, TOP-10) of the hit parade.

The condition for the correct filling of the rating

If the table is filled in correctly, the sum of ratings (in points) of all songs that have been in the hit parade for a certain period is equal to the product of the set *H* for this hit parade by the number of *Z* programs released during the given period. When calculating the number of programs released during the period under review, the final program, which summarizes the results of this period, should not be taken into account.

$$\sum_{i=1}^{n} S_i = H \cdot Z$$

If the sum of the ratings is not equal to the given product, the rating is filled in incorrectly. The error can be found by the formula

$$\theta = H \cdot Z - \sum_{i=1}^{n} S_i$$

From this error, you can find the place of the hit parade in which one is admitted, provided that the rating is checked after each new program entered into the archive and the errors are eliminated if they are found. Also, the place containing the error can be found using the formula below if the error does not exceed the number of places in the hit parade:

$$m_0 = m + 1 - \theta$$

Example. For the "Chartova Dozen" hit parade (*m*=13) the set is *H*=91 points (it was determined earlier). In the event that 51 hits of the hit parade were released during the year, with the correct filling of the table, the sum of the ratings of the songs that have visited during the year should be equal to

$$\sum_{i=1}^{n} S_i = H \cdot Z = 91 \cdot 51 = 4641 \text{ points.}$$

Comparison of results for different years.

Comparison can be made for any parameters - the number of songs, programs, average score, average TOP score. When comparing TOPs for different years, you can determine the record scores for different places at the end of the year. Record is the highest score for any place at the end of the year. Records are of two types: a fundamental record means that at the end of any other year the song would have taken a higher place than at the end of the year in which this song set a record, a non-fundamental record means that at the end of any other year the song would also have taken place as at the end of the year in which she set a record. The actual top of the hit parade is the rating of the song that took first place in the consolidated rating in the entire history. The annual top of the hit parade is the maximum rating of a song within a year, or a record score for first place at the end of the year.

Example. The annual ceiling of the "Chartova Dozen" is determined by the song "Nobody" by the "Kukryniksy" group and amounts to 328 points. The actual ceiling is determined by the song "Dance of the Evil Genius" by the "King and the Fool" group and is 405 points. $\Delta C = C_F - C_A = 405 - 328 = 77$ points. The hit parade has two ceilings, because the annual and actual ceilings are defined by two different songs.

Conclusions

On the basis of the classical theory of mathematical statistics, using the example of the statistics of Russian hit parades collected by the author, a method for comparing and predicting the position of an object in the rating is shown.

References

1. Armstrong J.S. Forecasting for Marketing // Quantitative Methods in Marketing. London: International Thompson Business Press, 1999. P. 92 – 119.
2. Jingfei Yang M. Sc. Power System Short-term Load Forecasting: Thesis for Ph.d degree. Germany, Darmstadt, Elektrotechnik und Informationstechnik der Technischen Universitat, 2006. 139 p.
3. Chervyakov N.I., Tikhonov E.E., Tikhonov E.E. Application of neural networks for forecasting problems and problems of identification of forecasting models on neural networks /Neurocomputers: development, application. 2003. № 10-11. P. 25-31.
4. Lebedyantsev V.V., Ozerova M.I. The influence of the architecture of the neural network and the initial data on the operation of the neural network for classification problems // In the collection: Information technologies in science and production. Materials of the VII All-Russian Youth Scientific and Technical Conference. Editorial Board: A.G. Yanishevskaya (ex.ed.) [et al.].2020. P. 145-152.
5. P.P. Bozhenko, R.U. Stativko Brief description of neural networks. Implementation of an expandable neural network//Bulletin of Youth Science of Russia. 2019. №5. P. 1.