ISSN 0005-1055, Automatic Documentation and Mathematical Linguistics, 2008, Vol. 42, No. 4, pp. 221–225. © Allerton Press, Inc., 2008.
Original Russian Text © V.M. Grinyak, S.M. Semenov, 2008, published in Nauchno-Tekhnicheskaya Informatsiya, Seriya 2, 2008, No. 7, pp. 20–24.

Automating the Solution of the Task of Classifying Clients According to the Stages of Relations in Modern Corporative Information Systems.

V. M. Grinyak and S. M. Semenov

Received April 9, 2008

Abstract—This paper considers a model for the task of controlling relations with clients by estimating the stability of sales. A procedure is suggested for building a system for supporting decision-making about the main parameters of the task—a trend model, an analyzed period, and the number of such periods. The efficiency of the suggested procedure is demonstrated on real data from the point of view of the prospects of its creation in modern corporative information systems.

DOI: 10.3103/S0005105508040031

INTRODUCTION

Corporate Information System (CISs) are customarily understood at present as the assembly of the means for collecting, transmitting, and preprocessing the data that accompany the activity of enterprises and organizations, structuring and analyzing these data and working out the set of possible management actions. The main task of modern corporative information system is to form the information basis necessary for plant administrations to make management decisions. This being the case, the content and form of information presentation must provide the smallest possible number of erroneous decisions.

Corporative systems always exist within the framework of some business-model (business-process) of an organization [1]. The functionality of modern CISs includes the solution of such tasks as strategic and operative planning, as well as commodity and business accounting in various areas. The enumerated tasks are supported technologically by program modules with the names “Finance Planning and Control,” “Personnel Management,” “Module of Business and Tax Accounting,” “Logistics,” “Production Planning and Control,” “Motor Transport Control,” etc [2, 3].

The management of relationships with clients (Client Resource Management—CRM) is one of the main constituents of modern management accounting. The task of CRM is to accumulate information about clients that can be used to carry out marketing research and to form the policy of an enterprise with respect to various clients. The functionality of CRM methodologically implies introducing catalogues for database objects—clients (dealers, partners, providers, buyers, or competitors), forming a certain set of user classifiers and attributes, assigning them to clients, and reflecting the catalogues of clients with their grouping by a chosen set of classification signs. Subsystems of a CRM with various contents are available in practically all modern corporate information systems that have gained a wide distribution on the domestic market (for example, in such corporate information systems as the Russian systems “Flagman,” “1S,” “Galaxy,” and the foreign systems “Axapta” and “SAP”) [2–5].

One of the tasks that are often realized within the framework of CRMs is to classify clients (as a rule, buyers) by their stage of relationship. The stages can be the following: a potential buyer, one-time buyer, constant buyer, or a lost buyer. From the standpoint of management activity, the results of solving this task can be used to analyze the significance of clients, their reliability and stability and to analyze the efficiency of the work of managers, or the a priory reliability of planning. Classifying the stages of relations with consumers is based on a retrospective analysis of the results of their relationships for a certain period; the number of transactions, sales volume, sales proceeds, sales income, etc. can be the analyzed indices. This being the case, all modern CISs give a user the opportunity to choose the parameters of a task solution (an analyzed index, a period of an analysis, or a number of periods for an analysis) exclusively by intuition or “manually.”

This work considers a model for the task of classifying clients by stages in their relationships and suggests a method for automatically choosing its parameters based on the results of a statistical analysis.

THE MAIN MODEL PRESENTATIONS

A day correlated with a concrete date is usually the period for estimating some index in problems of an economic analysis. Other periods are also often used: “weekly,” “monthly,” “quarterly” periods, etc. The

value of a chosen index X_k in a period with a number k can be expressed by the formula

$$X_k = \sum_{i=(k-1)*n}^{k*n} x_i, \quad (1)$$

where n is the number of days in the period, x_i is the value of the chosen index with the number i, $i = \overline{1, N}$ (N is the total number of days, for which the data were taken), $k = \overline{1, J}$ (J is the number of periods into which N days are divided in such a way that $N = Jn$).

The model of change in the index X in time can be expressed by the formula

$$X_k = G(k) + \eta(k), \quad (2)$$

where $G(k)$ is a function expressing the determinate law for the evolution of the value X (trend), $\eta(k)$ is a random value characterizing the deviation of the actual value of an index from its trend ($\eta(k)$ is considered here and below as an uncorrelated random value with a zero mathematical expectation).

Analyzing the properties of $\eta(k)$ can be taken as a basis for classifying clients by the stages of their relations. The idea of the classification is to group analyzed objects (clients) by their degree of homogeneity for analyzed parameters (by the mean-square deviation of $\eta(k)$ correlated with an average value of $G(k)$, which is expressed in percentage) [3].

Allow us to introduce the value

$$\nu = \frac{\sigma_\eta}{\overline{G}} \times 100\%, \quad (3)$$

where σ_η is the mean-square deviation of the value $\eta(k)$, $\overline{G} = \sum_{k=1}^{J} G(k)/J$ is the average value of the trend $G(k)$. If $G(k) = \text{const}$, ν is the variation coefficient of the value X_k. Having determined the value of ν for each client, one can place a client in various classes of their relationship stage. Thus, when the value of ν is $\nu \in (0.20)$, a client can be placed in the class of "very stable clients"; when the value of ν is $\nu \in (20.50)$, a client can be placed in the class of "average stable clients"; when $\nu \in (50.100)$, he can be considered to be an "unstable client," and when $\nu > 100$, he can be placed among "one-time clients"; the limits of classes are chosen proceeding from the specific activity of a company. Properly speaking, the value ν is not as important for the goals of the functionality of client resource management as the direction of its change is (i.e., the "improvement" or "worsening" of relations with a client).

The main problem in determining the value ν for each client is the choice of a function $G(k)$ and the choice of values n and J in such a way that the task model will be maximally representative.

PROBLEM SETTING AND A METHOD OF SOLUTION

Allow as to consider formula (2). Let us choose the polynomial model

$$G(k) = \sum_{j=0}^{m} a_j k^j. \quad (4)$$

as the model for the trend $G(k)$.

Formula (2) can be written in the generalized form:

$$X = Ka + \eta, \quad (5)$$

where X is the vector of values, X_i, K is the matrix of dimension $J \times m$, a is the vector of the coefficient of a polynomial a_i, and η is the vector of random values $\eta(k)$.

The solution of equation (5) by the least squares method has the following form with respect to the vector a [6]:

$$\hat{a} = (K^T R^{-1} K)^{-1} K^T R^{-1} X, \quad (6)$$

where $\hat{a}$ is the estimate of the vector a and $R = M[\eta\eta^T]$ is the covariation matrix (M is the operator of mathematical expectation). In the event that the $\eta(k)$ values are independent and similarly distributed, the matrix R has a diagonal form in such a way that $R_{kk} = \sigma_\eta^2$, where σ_η is the mean-square deviation of the value $\eta(k)$. With such a statistical uniform precision of $\eta(k)$ formula (6) will have the following form:

$$\hat{a} = (K^T K)^{-1} K^T X, \quad (7)$$

and, along with the estimate of the vector a, the estimate of σ_η can be obtained in the form of the formula:

$$\hat{\sigma}_\eta^2 = \frac{1}{J-m}(X - K\hat{a})^T (X - K\hat{a}). \quad (8)$$

If one designates the error of the solution of equation (5) through $\Delta a = \|a - \hat{a}\|$ by method (7), the corresponding dispersion matrix will be given by:

$$D_a = M[\Delta a \Delta a^T] = \left(\frac{1}{\sigma_\eta^2} K^T K\right)^{-1}, \quad (9)$$

and the estimate of this matrix will have the form:

$$\hat{D}_a = \left(\frac{1}{\hat{\sigma}_\eta^2} K^T K\right)^{-1}. \quad (10)$$

Determining the function $G(k)$. The problem of choosing the model for the trend $G(k)$ is reduced in this case to determining a polynomial order m. The present work suggests a method for choosing a value m based on the probabilistic estimate of the significance of polynomial coefficients in formula (4). Let $\hat{a}_j$ be the estimate of a corresponding polynomial coefficient (4), and $\hat{D}_a(j, j)$ be the coefficient of the matrix (10) lying on the

intersection of the *j*-th line and the j-th column (i.e., on the diagonal). Since the probability distribution of values $\hat{a}_j$ found according to formula (7) is close to the normal distribution independently of the distribution of values $\eta(k)$ [6], the probability that the zero values of a_j are outside the region of their probable values can be expressed by the formula:

$$P(a_j \neq 0) = \int_0^{2|\hat{a}_j|} f_{\hat{a}_j}(\tau)d\tau, \quad (11)$$

where $f_{\hat{a}_j}(\tau)$ is the normal distribution density function with average $\hat{a}_j$ and dispersion $\hat{D}_a(j, j)$. In the event that the value $P(a_j \neq 0)$ exceeds some threshold U_a, the decision is made that a_j is nonzero. Thus, the polynomial order (4) is determined by the maximal *j* for which $P(a_j \neq 0)$ is higher than an assigned probabilistic threshold U_a.

Determining the length of a period *n*. The choice of the number of days *n* in a period for which data will be summed in the course of an analysis should be made based on estimating the value $\hat{\sigma}_\eta$ from formula (8). Allow us to consider the value x_i, i.e., the value of a chosen index in a day with the number *i*. The essence of the problem consists in the fact that zero values make up a considerable part of a sample of x_i in the problems of an economical analysis (for example, sales to some client can occur only 2–3 times a month). This is why the initial data x_i are customarily transformed to some "enlarged" data X_k according to formula (1) and the question arises about choosing the length of this "enlarged" period in such a way that zero values will not, if possible, occur among X_k. (It is considered that $x_i \geq 0$).

Let $G(k)$ be the model for the trend of the value $X(k)$ and $Z = {}_k\min(G(k))$ be the minimal value of this trend within the interval $k = \overline{1, J}$. Let $\hat{\sigma}_\eta$ be the estimate of the mean-square deviation of a random model constituent, obtained by formula (8). Then the probability of that the *Z* found is above zero can be estimated by the formula

$$P(Z > 0) = \int_0^{\infty} f_Z(\tau)d\tau, \quad (12)$$

where $f_Z(\tau)$ is a normal distribution density function with an average of *Z* and a dispersion of $\hat{\sigma}_\eta^2$. Thus, the choice of *n* is reduced to the search for its probable values in increasing order, solution of problem (12) for them, and the choice of such an *n*, starting with which the value $P(Z > 0)$ will exceed some threshold U_Z. This being the case, the problem about the determination of *n* must be solved together with the problem about the determination of a polynomial order $G(k)$.

Determining the number of periods *J*. The number of periods, according to which problem (3) will be solved, must be chosen from the following considerations. On the one hand, *J* must be sufficient to "smooth" singular "random surges" of the value X_k. On the other hand, *J* must not be too high so that a considerable change in the value ν is excessively "smoothed" by retrospective data $X(k)$ and can be discovered at the very beginning of this process.

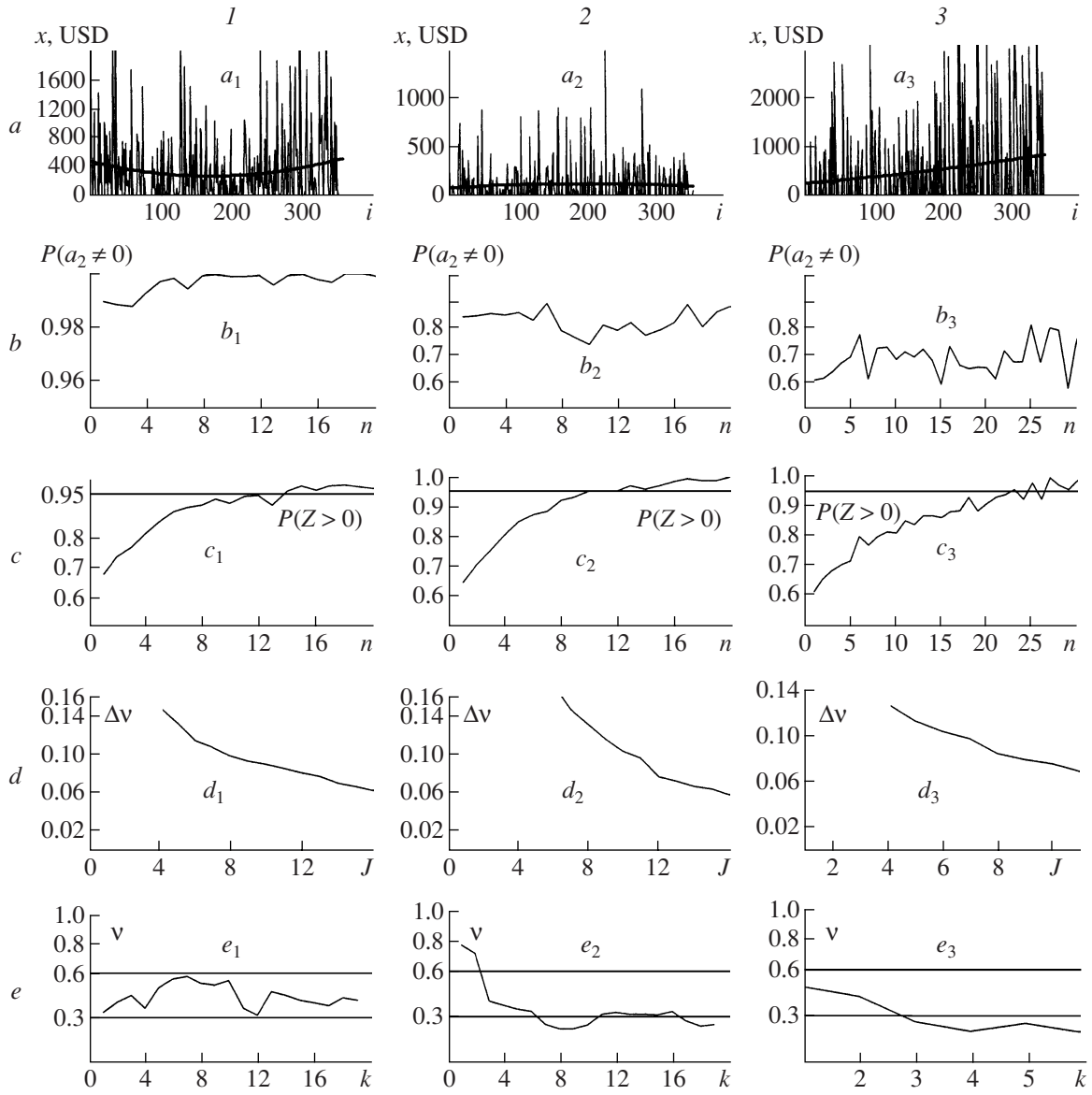
Let ΔX be the value of a random surge of X_k so that $\Delta X = G(k) - X_k$. In the event that the value ΔX has the normal distribution with the mean-square deviation σ_η, one can assert that $|\Delta X| < 2\sigma_\eta$ with a probability of approximately 0.95. For practical use, it seems expedient to choose such a *J*, starting with which the value ν would change not more than by U_ν in the event of a singular random surge $|\Delta X| = 2\sigma_\eta$.

A concrete value of U_ν is chosen by reasoning from the limits of ν for various classes of relationship stages (for example, if the limits of classes are chosen as 30%, 60%, and 90%, than U_ν can be made equal to half of the interval or 15%). The dependence of the value $\Delta\nu = |\nu - \nu_{\Delta X}|$ on *J* at $|\Delta X| = 2\sigma_\eta$ (where $\nu_{\Delta X}$ is the value of ν in the event that there is a random surge ΔX in data) is calculated by modeling problem (3) on the available set of data X_k.

THE RESULTS OF COMPUTATIONAL MODELING

The computational modeling of the problem of classifying clients by their stage of relationship was carried out based on real information about the annual dynamics of sales in a large company that trades in spare parts. The figure reflects the results of problem solution for three different clients (three columns of graphs). The graphs of curve *a* show the value of daily proceeds from sales of goods during a year and the value of the function for the trend $G(k)$, which was determined as a second-order polynomial. The graphs of curve *b* show the result of determining polynomial degree (4). One can see that the value of the coefficient a_2 is nonzero for the first client (b_1) with a high probability, and, consequently, the function for the trend $G(k)$ of the estimate of this client's sales is expedient to present as a second-order polynomial; the probability that the coefficient a_2 is nonzero is not high for the second and third clients, and, consequently, it suffices to model the function for the trend $G(k)$ of these clients by a first-order polynomial (from the practical experience of modeling, the value of the probabilistic threshold U_a is recommended to be taken as equal to not less than 0.95).

The graphs of the curve *c* show the result of solving the task of choosing the number of days in the period for which data must be summed in the course of the



The results of modeling the task of classifying clients by the stage of their relationship.

analysis. One can see that it suffices to sum 12–13 values for the first and second clients (c_1, c_2), and not less than 25–27 values for the third client (c_3). The graphs of the curve *d* show the results of modeling problem (3) to calculate the dependence of the value $\Delta\nu = |\nu - \nu_{\Delta X}|$ on J at $|\Delta X| = 2\sigma_\eta$. It is evident that if $U_\nu = 15\%$, the solution of the task of classifying clients by the stage of their relationship can be performed correctly for the first and third clients (d_1, d_3) on data of 4–5 periods, and that the number of such periods must be not less than 7 for the second client (d_2).

Lastly, the graphs of curve e show the result of solving the task of classifying clients by the stage of the relationship with parameters determined according to the suggested procedure. The following limits of classes are assigned:

0–30%—a constant buyer of the first type—a more stable buyer;

30–60%—a constant buyer of the second type—a less stable buyer;

more than 60%—an unstable buyer.

The graphs show that the first client (e_1) is within the limits of the class "constant buyer of the second type," the second client (e_2) gradually passes from the stage "unstable buyer" to the boundary state between the classes "constant buyer," the third client (e_3) has improved his stage within a year—he has passed from the class "constant buyer of the second type" to the class "constant buyer of the first type."

The procedure for solving the problem of classifying clients by the stage of their relationship that is suggested in this work has been adopted to the data of the corporative information system "1S," namely "1S: Trade Management 8" and "1S: Manufacturing Enterprise Management 8" and realized in the form of the adaptation of the platform "1S: 8," the main purpose of which is to give a manager recommendations for choosing the values of problem parameters (the size of an analyzed period and the number of periods and limits of classes) that are most appropriate for use in the program models of indicated typical system configurations. The presented results demonstrate the usefulness and consistency of the procedure.

REFERENCES

1. Zhdanov, B.A., New Logic and Factors of Development of CIS, *Korporativnye sistemy*, 2006, no. 3.
2. Bocharov, E.P. and Koldina, A.I., *Integrirovannye korporativnye informatsionnye sistemy* (Integrated Corporative Information Systems), Moscow: Finansy i Statistika, 2005.
3. Turban, E., McLean, E., Wetherbe, J. Information Technology for Management: Transforming Business in the Digital Economy, John Wiley and Sons, 2002.
4. Bogacheva, T.G., *1S: Predpriyatiye 8.0. Upravleniye torgovlei v voprosakh i otvetakh* (1S: Enterprises 8.0. Trade Management in Questions and Answers: Practical Guide, Moscow: 1S-Publishing, 2006.
5. Shuremov, E.L., *Informatsionnye tekhnologii upravleniya vzaimootnosheniyami s klientami* (Information Technologies for Client Resource Management), Moscow: 1S-Publishing, 2006.
6. Kramer, G. *Matematicheskiye metody statistiki* (Mathematical Methods of Statistics), Moscow: Mir, 1975.