Multifactor Modelling System with Cloud Based Pre-processing

Ventsislav Nikolov, Danko Naydenov

Abstract: *In this paper a practical multifactor modelling system is described implementing heuristic searching of polynomial formula that describe an unpredictable time series by a set of predictable series. The data needed for the calculations are pre-processed by a cloud computing system to fill the missing values. The features and problems of both pre-processing and formula searching systems are presented and analyzed.*

Key words: *Multifactor model, Heuristic Searching, Evolutionary, Cloud Computing, Missing Values*

INTRODUCTION

The processes in economics, physics and other fields of human activity could be separated in two categories.

- Predictable processes. For this kind of processes theoretical models exist for estimation with satisfactory accuracy of the future values. If these models are given then by feeding a model with its arguments for an arbitrary future moment an estimation could be obtained. Alternatively, the theoretical model could be obtained empirically but most often its accuracy is worse compared to the existing theoretical representation.
- Unpredictable processes. Such processes could not be modelled theoretically because there is not known method to calculate any arbitrary values. Here only empirical modelling can be performed as an only possibility. This kind of processes is often very important for the enterprise systems.

In this paper a software system is presented that models an unpredictable process by a set of predictable processes. The model is built as a best fitting polynomial formula [8] producing estimations as close as possible to the observed historical values. The processes are represented as time series which, because of their different origin and nature, often are incomplete and require pre-processing stage. This task could be effectively solved using more complicated methods than a simple linear interpolation. The method used here [6] is implemented using the cloud computing approach.

MULTIFACTOR MODELLING

The observations of a process are represented as an ordered sequence of values with a given constant time horizon. All such discrete time series are called factors. Some of them are used as modelling series called explanatory factors. The modelled series is called target factor and it should represent an unpredictable process which is of main interest. The explanatory factors for a given target factor should represent predictable processes. The first problem in the multifactor modelling is the selection of explanatory factors for a given target factor. The second problem is the building of the polynomial expression, especially including non-linearity to the model. The third problem is that the number of observations should be the same for both the target series and all explanatory factors. The series however often are with missing and incomplete data values. For example, the time series could represent historical interest rate values for different institutions in different countries in which there are different holiday non-working days or the database could contain some corrupted values, etc.

In Fig.1 the communication between the subsystems of the multifactor modelling and the missing values filling are shown. When the target and explanatory factors are known they are checked and if there are missing values the series are sent to the cloud based sub-system to generate the needed values and then the data is returned back to the multifactor system where the formula building stage is performed.

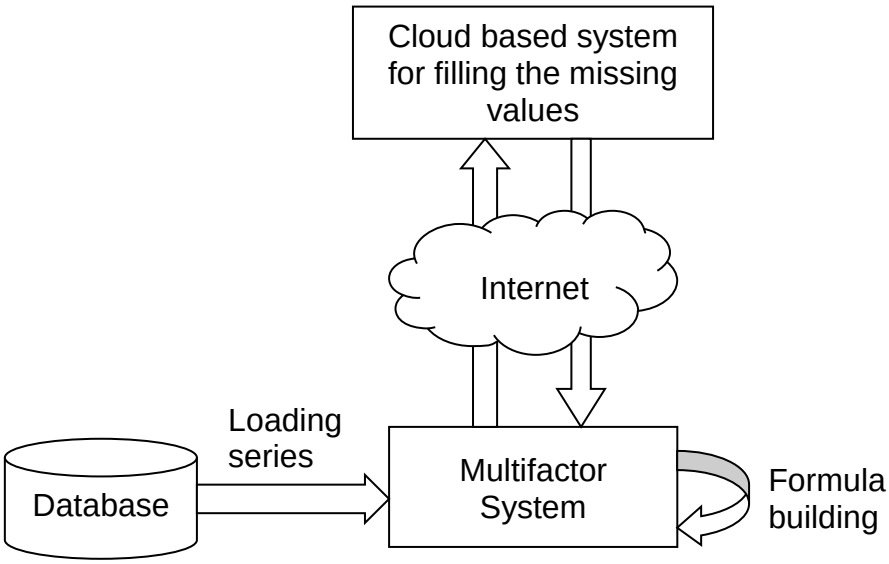


Fig.1 The multifactor system and cloud system filling missing values

CLOUD COMPUTING

Here some basic pieces of information are presented as reasons to use cloud computing for the implementation of the sub-system filling the missing values. Generally, the restoring of the missing values in the factors is a fundamental and common task. For this reason, solving the problem requires choosing an approach which allows wide and general solution. The current trends are the Internet accessibility to become easier and at the same time the communication speed through the global network to increase. This fact naturally leads to the choice of approaches that will yield the greatest accessibility of the algorithm for filling the missing values, namely to create a web application. In addition to the general availability another major advantage is the possibility of changing or updating the algorithm without the need to distribute a new version or recompilation of the applications that use it. However, the standard web approaches may also lead to the following problems.

- Reliability of service. If an application using a local library does not work then either the code file is missing or it is unavailable for use. It is easy to detect that in the very early stages of the application work. However, when a web application is used the unavailability issue is not so trivial. A communication should be initialized between the running application and the web service. The service may not be available either due to a communication environment failure or because of a breakdown of the service server.
- Response time. This indicator is determined by two components. The first one is the time for data transfer over the Internet and the second one is the time for applying the algorithm. It is assumed that the transfer time is relatively constant. In contrast, the time for the algorithm running is very significant for the server loading. Because a standard web application can be accessed by many users simultaneously, it can significantly

slow down the response time.

The approach that overcomes these problems and at the same time is a web based is the cloud computing. This is a modern technology that ensures reliability and at the same time leads to good response time due to the fact that each cloud application may be smoothly scaled to meet a peak load. In addition to solving these two problems, another major advantage of the cloud computing is its price. The basic cloud vendors provide free limited in time or a free quota of used resources.

The cloud system for filling the missing values is based on Platform as a Service (PaaS) cloud service which is the most appropriate for software development because it is similar to the development of a standard web application and also there is no need to install and maintain any system software, which reduces the maintenance costs of the application.

FORMULA SEARCHING IMPLEMENTATION

Initial formula building

After the filling of the missing values there are available observations of the target factor, which is selected beforehand but is unpredictable, that should be modelled using only a subset of all available factors, that are predictable, but it is not known which of them could best model the target factor. The model is generally built by the trial and error approach performed in two sub-steps.

- Selecting the explanatory factors for the target to participate in the formula. In the implemented system three approaches are used for automatically suggestion of explanatory factors:
  - The most correlated factors to the target factor;
  - The most uncorrelated factors each other;
  - By clustering. All available factors are clustered and then the target factor is classified in one of the clusters [11]. The explanatory factors are selected to be all factors that belong to that cluster.

  These three approaches automatically suggest a set of factors, but they could be manually changed by the user of the system adding new or removing some of the suggested factors. That flexibility allows additional checks or attempts to be performed.
- Building a polynomial formula using the selected explanatory factors. In addition to the explanatory factors the other components of the formula are the basis functions and regression coefficients. The set of the basis functions is preliminary selected and consists of well known mathematical functions like: sine, cosine, tangent, cotangent, square root, logarithm, power function, etc. The formula building consists of the following sub-stages:
  - Finding the basis functions combination to the explanatory factors – Fig.2. If the number of the explanatory factors is m and the basis functions are k then all possible combinations of basis functions over the explanatory factors are:

$$k * k * \ldots * k = k^m \quad (1)$$

(1) (2) ... (m)

Note that the explanatory factors are unique, and their order is preserved but the basis functions could repeat over the explanatory factors. Then there are k possible basis functions for the first factor. All of them can be combined with k possible basis functions for the next factor and so on to the last m-th factor. Thus, the combinations of the basis functions over the explanatory factors are permutations of k indistinguishable values in m positions [9]. Given a basis functions combination, e.g. $f_3, f_1, \ldots, f_2$, the target factor estimation is calculated as follows:

$$\hat{y} = \beta_1 f_3(x_1) + \beta_2 f_1(x_2) + \ldots + \beta_n f_2(x_n) + \beta_{n+1} \quad (2)$$

and the error ε, shown in Fig.2, is the distance to the observed target. Depending on the explanatory factors nature it is possible some basis functions to be unsuitable for the factor values. For example, a big factor value applied to exponential basis function will produce a value near to infinity and would cause a calculation error. Similarly, a square root applied to a negative explanatory factor value causes undefined operation.

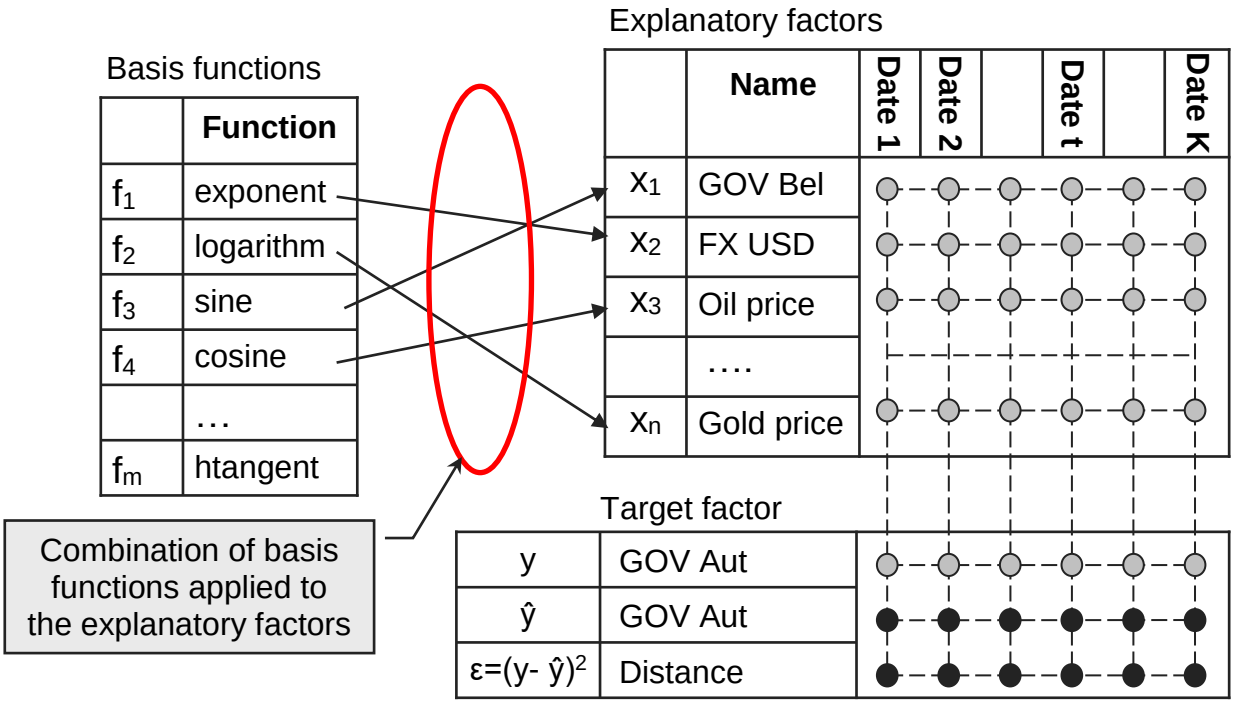
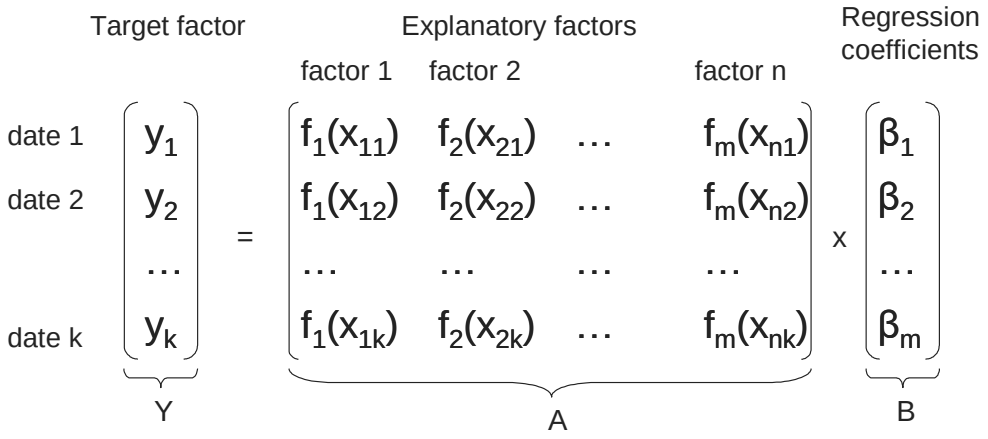


Fig.2 Basis functions combination over the explanatory factors and error (distance) calculation

- Finding the regression coefficients. When a permutation of basis functions is found the regression, coefficients are calculated using the ordinary least square method [1, 2, 3] solving a simple matrix equation (3).



$$\hat{B} = (A^T A)^{-1} A^T Y \quad (3)$$

and the target estimation in matrix notation is

$$\hat{Y} = A \times \hat{B} \tag{4}$$

After the regression coefficients calculation, only terms with significant coefficients can be preserved [10]. Thus, factors reducing may be performed as logically second sub-stage of the explanatory factors selection.

The calculation of the regression coefficients is an easy task that consists of only simple matrices operations like multiplication and inversion. The matrix inversion can be fast implemented, for example by LU decomposition. However, the basic functions permutations, that should be searched, are too many taking into account that the number of explanatory factors and basis functions is at least several dozens. The main object is to find a satisfactory basis function permutation from amongst all of them that are also called solutions. These permutations are searched using the following heuristic optimization methods.

- Evolutionary heuristic.
  This approach requires the following settings to be specified beforehand (generally obtained from a settings file for the system):
  - Number of iterations “I”;
  - Number of temporary solutions ”S”;
  - Number of the best temporary sub-solutions ”H”;
  - Number of attempts ”C” used to avoid solutions repetition;

  The algorithm consists of the following steps.
  1) Initial generation of ”S” temporary solutions;
  2) Calculation distances:
     a. Regression coefficients calculation;
     b. Calculation of the target factor estimation, using (4), and the Euclidean distance (ε in fig.2) is calculated to the target factor. A decay factor can be used to take more importance into account (y and ŷ are the target factor and its estimation correspondingly):

        Euclidean distance without decay:

$$d = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{5}$$

        Euclidean distance with decay λ:

$$d = \sqrt{\frac{1}{\sum_{i=1}^{N}\lambda^{i-1}}\sum_{i=1}^{N}\lambda^{N-i}(y_i - \hat{y}_i)^2} \tag{6}$$

     c. Saving the information for the best solution - basis functions permutation, regression coefficients, target factor estimation, distance value.
  3) Selection of the best H solutions according to the calculated distance. Two alternatives are used [5]:
     a. Roulette wheel;

b. Truncation selection;

One of these two alternatives is selected and it is not changed until the algorithm finishes.

4) Recombination - producing S solutions from the selected H solutions
   a. Splitting at a randomly chosen point, thus producing two subsets of partly solutions: left parts and right parts;
   b. Two random elements are selected - one from the left and one from the right parts – fig.3. The selected parts are merged producing a new solution. If it is already generated a new pair of elements are selected. This is done until the number C is reached and after that the solution is accepted even though it may exist, otherwise an infinite loop may arise. Such solution generation is performed until S solutions are obtained.

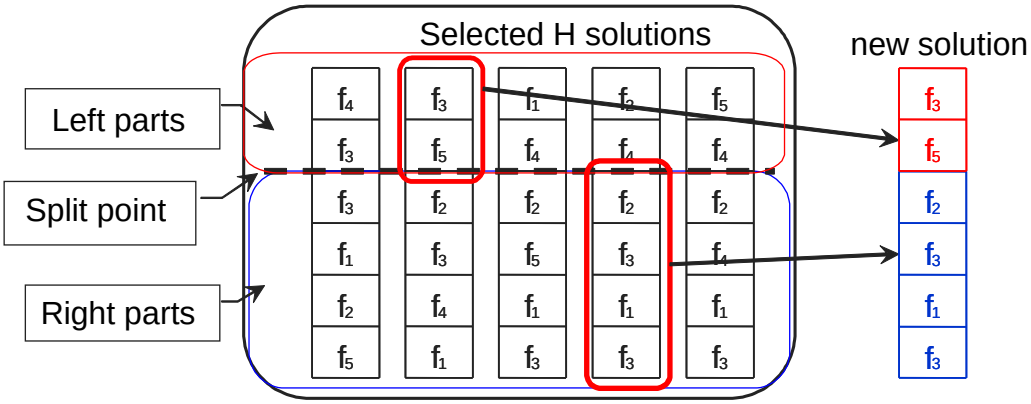


Fig.3 Recombination

5) If the current iteration number is I then finish, else go to the step 2.

- Random searching

  Here only the number of iterations should be specified before starting the algorithm. The following steps are performed.

  1) A new random solution is generated.
  2) The distance is calculated identically to the step 2 of the evolutionary heuristic method.
  3) If the current iteration number is I then finish, else go to the step 1.

Formula calibration

The formula found using some of the above two methods is not guaranteed to be the best solution. But it is near to the optimal solution and most often it is good enough for practical usage. Taking into account that periodically new observations are appended to the target factor, the formula must also be periodically recalculated in order to keep the accuracy good enough. The recalculation can be performed either with the already selected explanatory factors or with new factors. Most often the formula is recalculated with the old factors and if the results are not satisfactory then the explanatory factors selection stage is re-performed. The process is shown in its time development in fig. 4.

RESULTS

The multifactor modelling system described here is used for market series in financial portfolio modelling. Both the filling missing values and the formula searching are written in Java and implemented as cloud based service and software library respectively. The formula building is performed locally, and its major criterion is the accuracy of modelling. The target series is estimated by the modelling formula and several statistics are used to asses the error between the estimated target and real target series. The results are shown

in table 1 where the average values of the distance (with decay 0.96), correlation, R^2 and adjusted R^2 are shown as fitting statistics between 500 target factors and their estimations. The basis functions are: natural logarithm, sine, cosine, tangent, cotangent, square root and power functions with degree 1, 2 and 3.

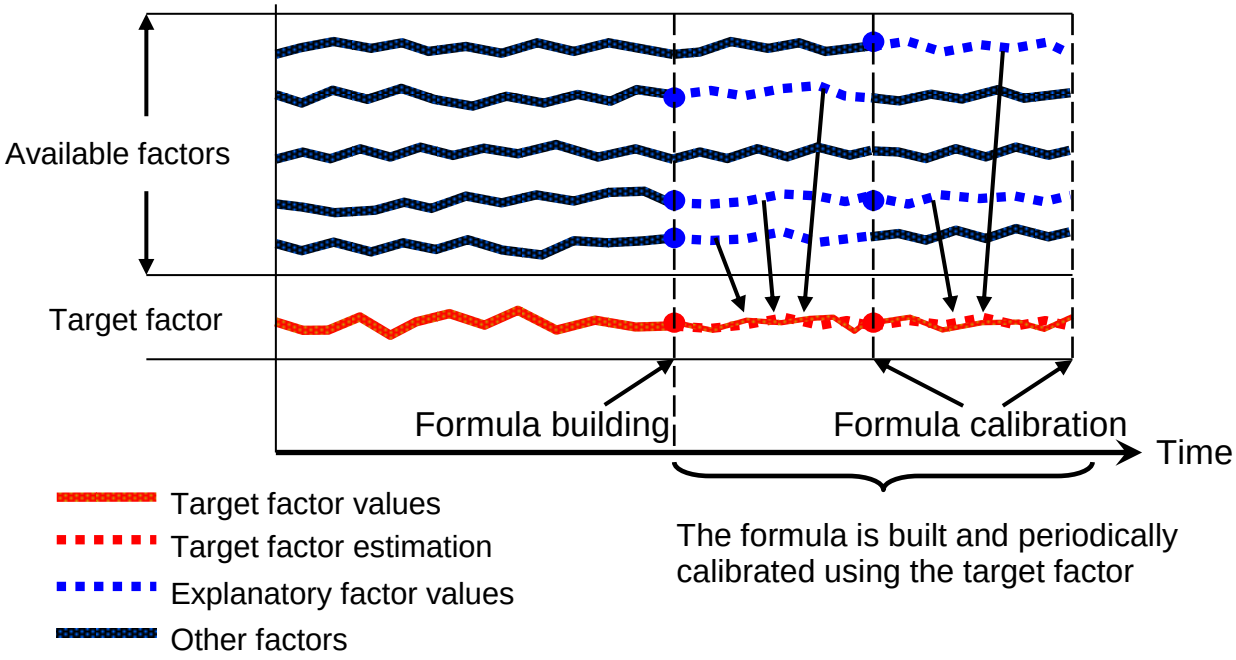


Fig.4 Formula building and calibration in time

Table 1. Average results for 500 target series each with 109 explanatory factors

Statistic	Without basis functions	With basis functions	
		Evolutionary method	Random method
Distance	1.57051	1.27608	1.35635
Correlation	0.99491	0.99545	0.99578
R^2	0.98984	0.99092	0.99158
Adjusted R^2	0.98204	0.98395	0.98512

Taking into account that the number of the explanatory factors correspond to the degrees of freedom, better solutions emerge when this number is near to the number of historical observations that is the time series size. It is possible the matrix of the explanatory factors in (3) to be non-positive definite in which case the solution does not exist. This problem is overcome by:

- Using of various enough basis functions in order to obtain a positive definite matrix of explanatory factors transformations;
- Matrix perturbation methods [4, 7]. These methods modify the matrix with minimal possible values in order to transform it to be positive definite. The practically used software solutions need such methods in order to continue the work even with some error that should be acceptable.

CONCLUSIONS AND FUTURE WORK

The implemented software solution comprises the formula building sub-system and the cloud based sub-system realizing the modified Roweis algorithm. In the former the searching of basis functions combinations is realized as a possibility to improve the accuracy and to overcome problems like dealing with non-positive definite matrices. The latter is realized as a cloud service improving the availability and reliability as well as giving

certain advantages as making the often-used algorithms to be reusable, realizing separation of concerns, etc. Currently in the formula building system the evolutionary search performs more machine operations and it is more time-consuming compared to the random search. However, as it can be seen from the results, it produces better solutions. The weakest point of the current software implementation of the formula building is the explanatory factors selection. Further work could be done to improve that stage. Even though the results show that the practical implementation is fully usable, and it is applied solving real world problems. The multifactor system could also be entirely based on cloud environment. In this case some problems emerge mainly because of the data accessibility because the series in the database are often too many and also there are difficulties in the format of the data to be transmitted.

REFERENCES

[1] Cameron, C. Trivedi, P. Regression Analysis of Count Data. Cambridge university press, 1998.

[2] Draper, N. Smith, H. Applied Regression Analysis. Wiley Series in Probability and Statistics, 1998.

[3] Hamilton, J. Time Series Analysis. Princeton University Press, 1994.

[4] Kim, J., A. Malz, J. Mina. Long Run Technical Document. RiskMetrics Group, 1999.

[5] Koza, J. Genetic Programming. MIT Press, 1992.

[6] Nikolova, N., D. Toneva-Zheynova, D. Naydenov, K. Tenekedjiev. Imputing missing values of environmental multi-dimensional vectors using a modified Roweis algorithm, IFAC WorkshopDynamics and Control in Agriculture and Food Processing (DYCAF), June 13-16.

[7] Rebonato, R., P. Jaeckel. The most general methodology for creating a valid correlation matrix for risk management and option pricing purposes. Journal of Risk, Vol.2, No.2. Winter 1999.

[8] Recktenwald, Gerald. Numerical Methods with Matlab: Implementations and Applications. Prentice Hall, 2007.

[9] Rosen, K. Discrete Mathematics and Its Applications, Forth Edition. AT&T, 1998.

[10] Steel, R. Torrie, J. Principles and Procedures of Statistics. McGraw-Hill, 1960.

[11] Xu, R. Wunsch, C. Clustering. Wiley, 2009.

ABOUT THE AUTHORS

Dr. Ventsislav Nikolov
Senior Software Developer
Eurorisk Systems Ltd.
31, General Kiselov Str., 9002 Varna, Bulgaria
E-mail: vnikolov at eurorisksystems dot com

Danko Naydenov
Senior Software Developer
Eurorisk Systems Ltd.
31, General Kiselov Str., 9002 Varna, Bulgaria
E-mail: sky at eurorisksystems dot com