SUPPORTING OF FORECASTING MODELS IN INFORMATION-ANALYTICAL SYSTEMS OF RAILWAY TRANSPORTATION

Eugene Kopytov[1], Vasilijs Demidovs[2]

[1]Transport and Telecommunication Institute
Lomonosov Str.1, Riga, LV-1019, Latvia
Fax: +371 7100660. E-mail: kopitov@tsi.lv

[2]State Join-Stock Company "Latvian Railway"
Turgenev Str. 21, Riga, LV-1050, Latvia
Fax: (+371)-7234366. E-mail: dem@ldz.lv

The paper presents the approach of building Informational-Analytical Systems at the railway with the help of modern achievements in the sphere of Data Warehouse. Application of Data Warehouse technologies allows building a lot of system models on the basis of a huge accumulated quantity of data. The suggested conceptual scheme of building virtual models gives opportunity to have several analysed data models concurrently and to create quickly new ones. The offered approach can be applied to forecasting of railway transportation depending on the control action at the system with account of effect of the environment.

Keywords: *forecasting, railway transportation, information system, data warehouse, virtual model*

1. INTRODUCTION

At present time European Union aims to make traffic more flexible in Europe, and the EU White Book on Transport, published in 2001, describes the Union transportation policy until 2010 [1]. It emphasizes the importance of the transportation of loads on railways instead of using the roads. There is a note that community funding should be redirected to shift the focus to the use of railway transport [2].

Railway transport in the Latvian Republic is an important economic field of the country, more than 13 thousand people worked on State Joint-Stock Company "Latvian Railway" in 2003, the asset of the Company made up 195,366 million LVL, its own capital making up 102,056 million LVL [3]. High dynamics of changes of the transport services market and strong competition suggested by other transporters demands a complex analytical research with the account of all data accumulated in different spheres of the railway activities as well as developing multi-variant predictions for different periods of time. It requires, in its turn, developing a sole integrated Dataware system, which should combine in itself data from different Information Systems of the Latvian Railway [4].

Generally, the process of decision-making for the further period of the railway functioning is the analysis of the information (previous, present and future) characterizing objects of management and working out on its basis a certain set of control actions, which provide the achievement of the chosen targets, set (as a rule) by the efficiency criteria. Thus, the task of developing the Decision Support System (DSS) on the railway includes:

- establishing an Information System to provide the completeness and trustworthiness of the stored data;
- developing a system of predicting passenger and cargo flows on the basis of the prediction models;
- working out a system of making reasonable decisions for managing the process of transportation.

Note that in the frame of the given paper the second task is of greatest interest but it cannot be solved in isolation without solving the first task and the third one.

2. PROBLEMS OF DEVELOPING AN INTEGRATED INFORMATION-ANALYTICAL SYSTEM FOR THE LATVIAN RAILWAY

At the present time the Latvian Railway employs Information Systems (IS) responsible for different spheres of its business shown in Figure 1. The information generated by these systems

allows evaluating the efficiency of performance and to execute prospective planning. For solving these tasks in each system there have been worked out specialized local program sets for requirements of any level managers. But these systems are mainly oriented towards performing account-controlling and enquiries functions and the final user doesn't actively interact with them in the course of making optimal managing decisions. Most of these programs have been developed on the basis of outdated technologies, they can work only with small volumes of data, and it is difficult to support this software. As an example we'll give a structure of information support of a passenger transportation system, which combines four Information Systems (Express, Ticket-cash machines #1, Ticket-cash machines #2, Cashiers on site) developed at different times on different program and apparatus basis [5]. These systems generate more than 12 million transactions per year [6]. Under a *transaction* we understand a registered fact of selling a ticket or a group of tickets, which is not equivalent of a trip. Tickets may be single or season, one-way or return. The data received from these systems have different format and degree of aggregation. For example, data from hand sale enter the system being already aggregated according to the type of the ticket, stations of sale, departure and destination. To get information about the number of transported passengers the data should be subjected to additional procession.

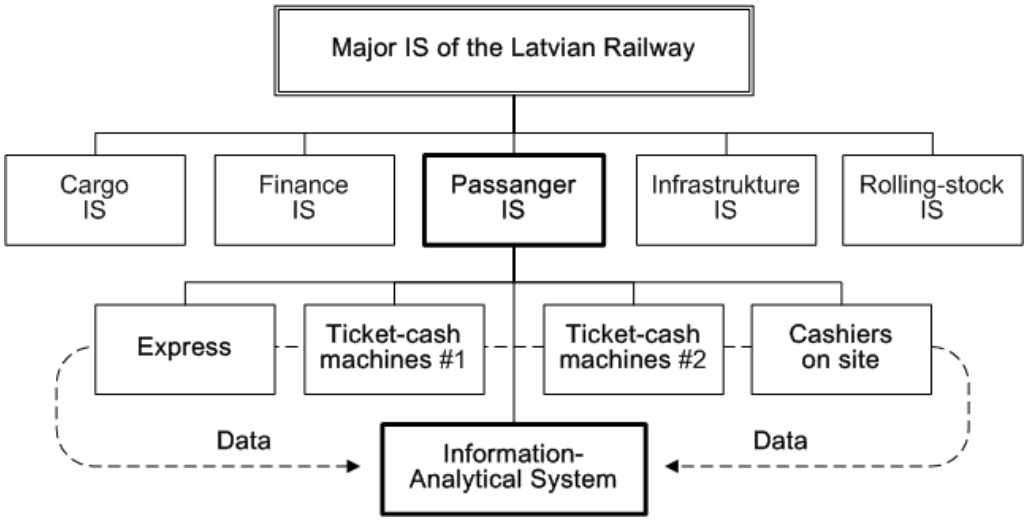

Figure 1. *Information Systems of the Latvian Railway*

The basics of modernizing and developing Information Systems on the Latvian Railway were laid in 1997 when the general conception of developing Latvian Railway IS was worked out and adopted [7]. This conception gave priorities to seven IS and first of all to building a high speed network of data transmission (1GB WAN), a finance system and a single integrated database as a general information resource of the whole enterprise.

Application of new technologies such as Data Warehouse (DW) does not eliminate all existing problems, but allows performing alienation of the existing systems and consolidation of analytical data in DW, thus freeing operative systems from improper functions, and modernizing the systems actually separately [4].

In the sphere of passenger transportation this task was simplified by the fact the four existing systems (see Figure 1) were actually not inter-connected and could function separately. Integration of data from these systems was performed only at the analysis stage [6, 8]. Nevertheless after introducing the analytical system exploitation of the old one continued on for half a year.

Thus in the recent five years there has greatly increased interest to the questions of building Information-Analytical Systems, which allow performing a multilevel analysis, planning and managing the processes of passenger and cargo transportation on railway and other types of transport. These systems enable managers to take prompt, effective decisions in the new, quickly changing economic conditions.

It proves the actuality of carrying out research to improve the dataware system of the decision-making tasks in the Latvian Railway in the sphere of passenger transportation. Recently DSS has stopped being the prerogative of the certain layer of managers and has a form of pyramidal model (see Figure 2), and, therefore, the increased requirements of accessibility and data safety and reliability are laid down to it.

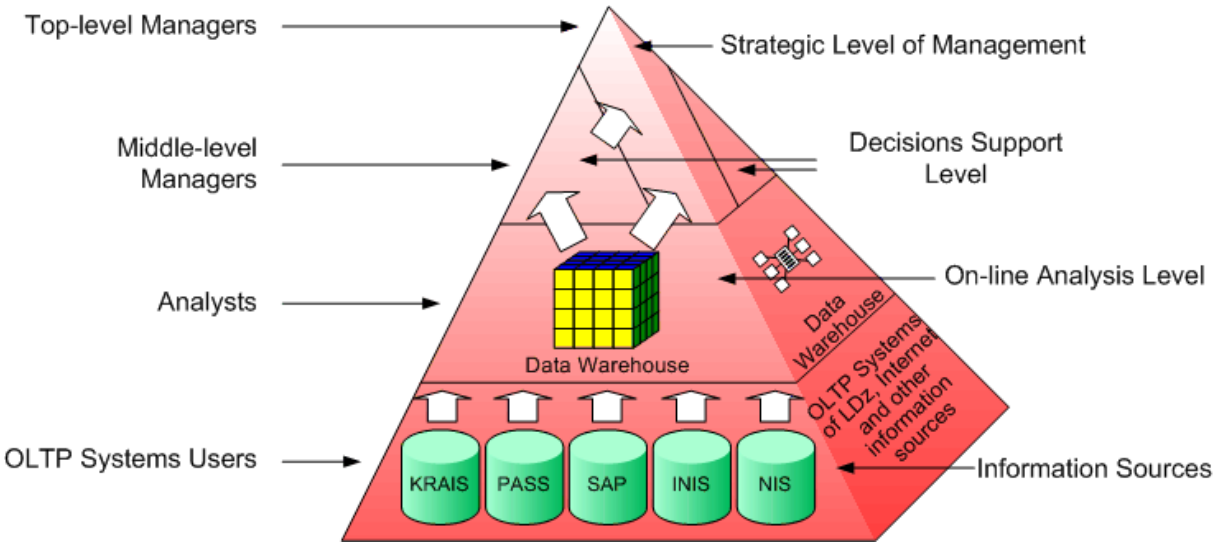

Figure 2. *Pyramidal Model of Decision Support System*

The analysis of the situation in the Latvian Railway has shown that the questions of data trustworthiness are the key questions in building Decision Support Systems. Therefore let's consider them in more detail.

Reliability of prediction depends on the reliability of data stored in Data Warehouses. DW architecture is based on the schemes type "star" and "snowflake". It is peculiar for these schemes to have tables of factors where all transactions and all aggregates of transactions are described, as well as tables of measurements for each entity. Here, the concept of transaction can differ from the similar concept in the initial data obtained from On-Line Analytical Processing (OLAP) systems [7]. One of the obligatory measurements of the given scheme is a measurement of time. Thus, a time indicator is present in the data prepared for analysis. This time indicator is formed periodically once per week, decade, month, quarter or year during the data processing. Forming of data for the certain period is being performed on the basis of Referenced Data System (RDS). Data situated in RDS are the subjects to continual changes, for example, old trains are cancelled and new ones are appointed, schedule and routes of trains are changed. The preparation of data for the long period of time demands registration of all changes in RDS for the processed period. When there is a necessity of the creation of the new physical data model, which is impossible to receive on the basis of the existing ones, we shall need all initial data. Therefore, for the flexibility of the system in Data Warehouse it is necessary to store data received from the initial systems and transformed into one form. Therefore, time factor reflected in temporal model RDS should be taken into account [4, 5].

3. VIRTUAL MODELS IN THE FORECASTING TASKS OF RAILWAY TRANSPORTATION

Forecasting plays an important role in the tasks of managing and planning transportation flows on transport and is the main part of DSS. Prediction in the railway transportation, which is a rather complicated system, is constructed on the experience gained by the system for the prolonged period; at the same time very large volumes of the accumulated historical data are used. In many cases we

assume that behaviour of the data in the past will be a good guide to its behaviour in future. When solving the problem of prediction usually a standard admission is entered: if in the past the system reacted in a certain way on the similar events, then, with higher level of probability it is possible to suppose that it will react in the same way in future as well.

Comprehensive planning of transport company activity demands presence of a certain set of models that adequately describe functioning of the railway and application of various mathematical methods for forecasting of transportation process. During the forecasting we need to develop various decision strategies to deal with future uncertainties.

In general case the considered forecasting task is to anticipate behaviour of railway system in the future taking in account possible control impacts on the system and environment behaviour.

Let vector $Y(t) = \{y_1(t),\ y_2(t),\ ..., y_k(t)\}$ characterizes the status of railway system at the moment of time *t*, where $y_j(t),\ j = \overline{1,k}$ are the system's indicators observed at the moment of time *t*. Let the prehistory of a system is known, i.e. vectors $Y(t_1), Y(t_2), ..., Y(t_n)$ are given, where $t_1, t_2,\ ..., t_m$ are moments of time in the past and present.

As the authors suggest [4-5], we can predict the state of railway system $Y^*(t_e)$ for future moment of time t_e using the prediction model in the following form:

$$Y^*(t_e) = F_1\left[Y(t_i),\ i = \overline{1,m};\ C(t_e); Z^*(t_e)\right], \tag{1}$$

where F_1 is a certain functional;

$C(t_e) = \{c_1(t_e),\ c_2(t_e),\ ..., c_q(t_e)\}$ is the vector of control impacts on the system in the period of time from t_m to t_e;

$Z^*(t_e)$ – the vector of predicted values of random factors, characterizing the impacts of environment on the investigated system in the moment of future time t_e.

So, using model (1) we have to consider a set of possible management strategies $C_1(t_e),\ C_2(t_e),\ ...,\ C_n(t_e)$ for several possible variants of environment behaviour $Z_1^*(t_e),\ Z_2^*(t_e),\ ...,\ Z_n^*(t_e)$ in the period of time from moment t_m to moment t_e. Thus, for *n* versions of environment behaviour we obtain accordingly forecasts $Y_1^*(t_e), Y_2^*(t_e), ..., Y_n^*(t_e)$ that give opportunity to use the situation control in the considered system.

The given approach suggests that some physical models having states $Y_1^*(t_e), Y_2^*(t_e), ..., Y_n^*(t_e)$ are parallel created without changing and locking the initial Real model stored in Data Warehouse (see Figure 3).

Let us note that considered approach has some disadvantages:

- for storage of created Prediction Models it is necessary to have huge amount of free disk space, and amounts of the stored data are increased proportionally to the quantity of models;
- it is necessary to generate a new Prediction Model at any new version of control actions $C_j(t_e)$.

For solution of the given problems the method of the virtual models is offered [5, 8].

Main principles of virtual models are the following:

- in a Data Warehouse are located only metadata of Models Repository, describing a set of virtual models constructed on the basis of real model and vectors of control impacts;
- we can describe all effects at the system in metadata as the functional dependences.

Virtual Models are actually templates of the mathematical prediction models with the calculated values of evaluating unknown parameters and prediction variables are taken from the Data Marts Repository at the time of carrying out the analysis, and it allows exercising prediction for other time intervals. Developing of virtual data models makes it possible to reduce the Data Warehouse size. There is no need of blocking the Real Model of data since each analyst works with its own set of Virtual Models.

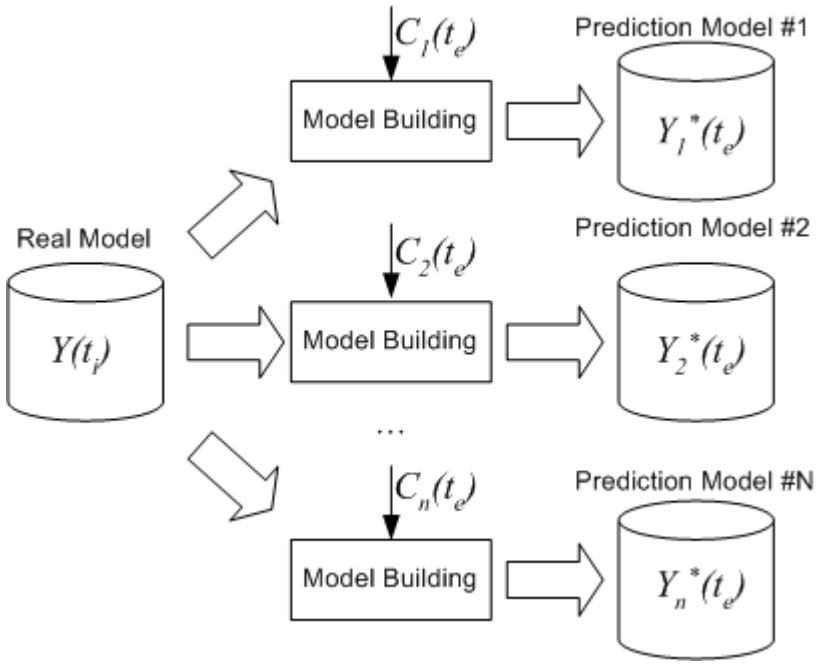

***Figure 3**. Conceptual scheme of Prediction Models building in anticipatory system*

As *an example* that shows an opportunity of the suggested approach let us consider the forecasting task of railway passenger transportation in the regions of Latvia. The search of optimal solutions on the future stages of railway transportation demands the construction of a set of prediction models, on the basis of which various strategies of behaviour of the railway system have to be considered.

To analyse and predict the number of passengers transported by railway in *i*-th region of Latvia during the year it is suggested to employ the multiple linear regression models of the form:

$$y_i^* = \beta_0^* + \beta_1^* \mathrm{x}_{i,1} + \beta_2^* \mathrm{x}_{i,2} + ... + \beta_\mathrm{n}^* \mathrm{x}_{i,n}, \qquad (2)$$

where y_i^* is the predicted value (estimate) of the dependent variable Y (a number of passengers transported by railway in the regions) for the *i*-th observation (region of Latvia);

$\beta_0^*, \beta_1^*, ..., \beta_n^*$ – estimated regression coefficients (evaluations of the unknown parameters);

$x_{i,1}, x_{i,2}, ..., x_{i,n}$ – values of independent (predictor) variables X_1, X_2, ..., X_n (accompanying factors) for the *i*-th observation.

Using formula (2) several Prediction Models, which allow evaluating the influence of the main social-economic factors on the passenger transportation by the railway in the regions of Latvia, have been created [9].

On the basis of the suggested virtual models' approach there appears a flexible possibility of suggested Prediction Models implementation in Data Warehouse, shown in Figure 4. Prediction Models are built using stored in Data Warehouse statistical data about passenger transportation in the regions as well as by the indicators of the regions social-economic development in the observed period. Data Marts Repository is created taking in account the suggested business logics and external factors and consists of two domains:

- Real Data Marts, which data physically exist on hard discs;
- Virtual Data Marts, the data are formed at the access moment.

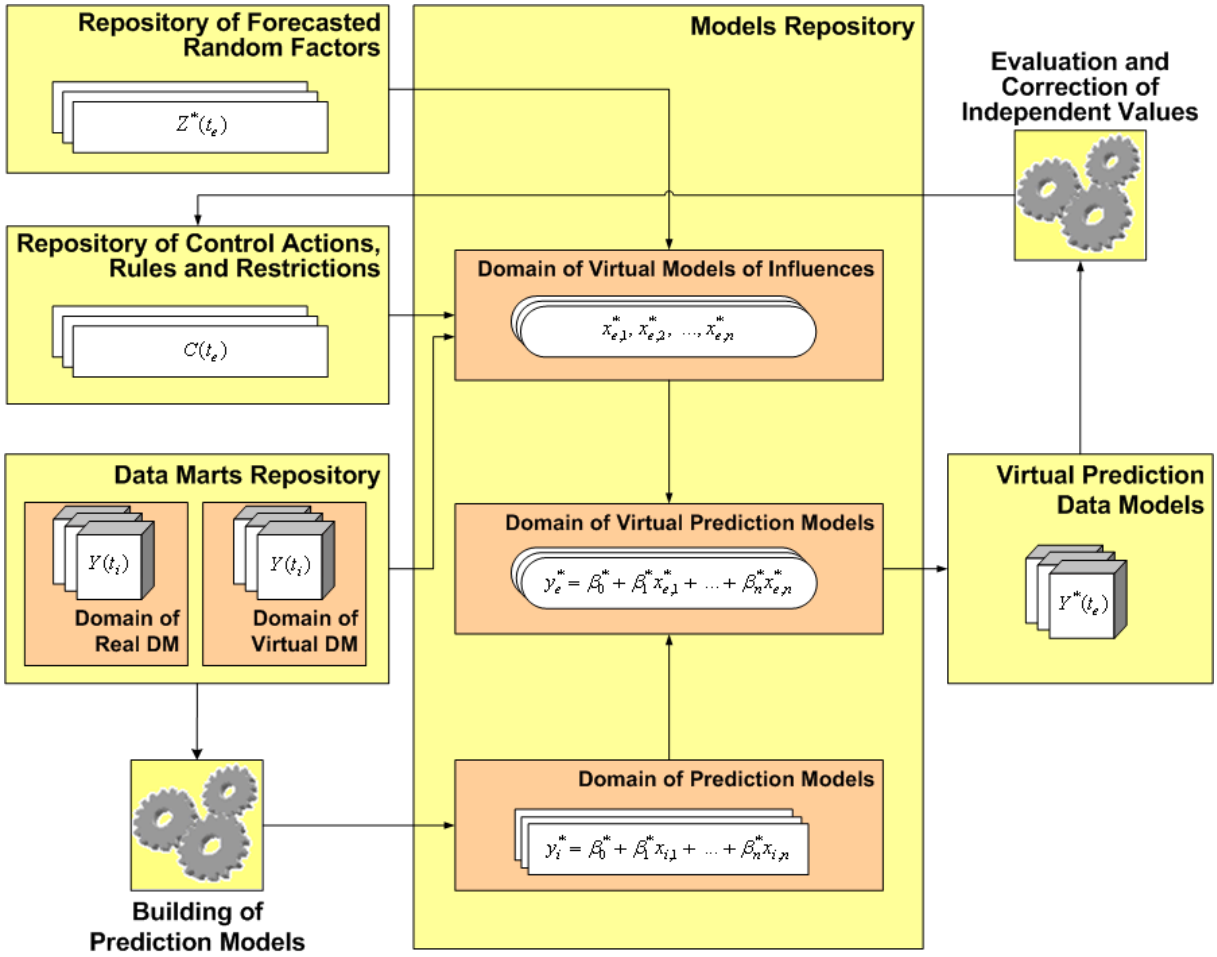

Figure 4. *Conceptual schema of Virtual Prediction Data Models building approach*

Calculated estimates of unknown parameters $\beta_0^*, \beta_1^*, ..., \beta_n^*$ together with corresponding values of predictor variables $x_{i,1}, x_{i,2}, ..., x_{i,n}$ are loaded in the related basic table stored in Domain of Prediction Models. Domain of Virtual Prediction Models consists of a set of views (virtual relations). For each kind of model the corresponding view is created. Views, that realize multiple linear regression models (2), are using evaluations of unknown parameters $\beta_0^*, \beta_1^*, ..., \beta_n^*$ stored in Domain of Prediction Models. Estimates of predictor variables $x_{i,1}^*, x_{i,2}^*, ..., x_{i,n}^*$ are on-line calculated using corresponding data, stored in Data Marts Repository, Repository of Control Actions, Rules and Restrictions and Repository of Forecasted Random Factors, by formula

$$x_{e,n}^* = F_2\left[Y(t_i);\ C(t_e);\ Z^*(t_e)\right], \tag{3}$$

where F_2 is a certain functional;

$Y(t_i)$ is a vector of variables that characterized the passenger transportation stored in Data Marts Repository; for example, train timetable in the prehistory;

$C(t_e)$ is a vector of control impacts on the system in the future period stored in Repository of Control Actions, Rules and Restrictions; for example, the number of railway stations in the region, train timetable, tariff policy, and others;

$Z^*(t_e)$ is a vector of predicted values of factors characterizing the impacts of environment on the investigated system stored in Repository of Forecasted Random Factors; for example, population density in the region, density of the unemployed population in the region, the number of schools per a unit of territory in the region, bus timetable, USD exchange rate, and others.

The considered schema of Virtual Prediction Data Models building allows evaluating the current state of the transportation system, making prediction for future $Y^*(t_e)$, defining necessary correcting impacts $C(t_e)$ and testing them. The correction procedure of management is the following. On the basis of developed recommendation the decision is made, which is evaluated on the developed Virtual Prediction Data Model, and again the prediction is performed taking into account the made decision. Thus, the system allows recognizing operational degradation recursively and in proper time reacts to them [10].

In the process of the receipt of new data a virtual model changes its status reacting to the deviations from the predicted behaviour. Thus, when using self-correction mechanism the correction of the predicted future situation happens as well. A continual testing of the reliability of a prediction is taken place and in case of negative deviations new correction impacts are defined.

Up to the present moment we assumed that the prediction is being performed for the future period, i.e. $t_e > t_n$. But model (2) has practical application for case $t_e < t_n$ as well, i.e. the prediction is being performed for the past period. This allows assessing the efficiency of the undertaken management decisions in comparison with other alternative decisions. The suggested method helps to choose an appropriate strategy to avoid common mistakes in the future. The authors consider in [4] as an illustration of the suggested approach the task of Latvian Railway losses analysis from the sizes of currency corridor.

CONCLUSIONS

The approaches to building Information Analytical Systems suggested by the authors have realized in a number of Decision Support Systems on the Latvian Railway. Employing Virtual Prediction Models makes it possible to change the values of the predictor variables at the time of carrying out the analysis. It allows taking account of any suppositions in predictions or predictions of change of the independent analysed variables, thus realizing the prediction of the type "what will be if …" or "what would be if …".

The experience of realizing virtual models has shown that they can easily set the rules and restrictions managing the process of calculations and make the process of data transformation in Decision Support System transparent and easily controlled. Complex analysis of a number of possible alternatives of the events development in the system of transportation for different variants of managing affects and different behaviour of the external medium and working out the suggestions for optimal system management in different situations – without reiterative increase of the disc space in relation to the basic variant and blocking the data warehouse in the time of carrying out the analysis.

The offered approach can be applied to forecasting of railway transportation depending on the change of charges policy as control action at the system with account of effect of the environment. The external effects which are taken into account during the predicted period of time, for example in a case with freight traffic, can be the registration of activity of other carriers inside the country and in the adjacent states, change of a dollar exchange rate and the prices for transported cargoes etc.

References

[1] Kazatsay Z. The Hungarian Transport policy – Accession to the European Union. http://www.imprint-eu.org/public/Papers/IMPRINT5_Kazatsay.pdf (2005, June 22).

[2] Pohjamo S. Logistics Event. http://www.logisforum.fi/eng/news/pohjamoeng.doc (2005, June 22).

[3] *Annual Report of State Joint-Stock Company "Latvijas Dzelzceļš" for 2003*. Riga: SJSC "Latvijas Dzelzceļš", 2004. 37 p. (In Latvian)

[4] Kopytov E., Demidovs V., Petoukhova N. Principles of Creating Data Warehouses in Decision Support Systems of Railway Transport. In: *Computing Anticipatory Systems. CASYS 2003 - Sixth International Conference: Conference Proceedings 718.* /Edited by D. M. Dubois. Published by The American Institute of Physics, pp. 497-507.

[5] Kopitovs J., Demidovs V., Petoukhova N. Virtual Models in Forecasting Systems of Railway Transportation. In: *Proceedings of the International Conference "Modelling and Simulation of Business System".* /Edited by H. Pranevicius, E. Zavadskas and B. Rapp. May 13-14, 2003, Vilnius, Lithuania. Kaunas: Technologija, 2003, pp. 265-268.

[6] Kopytov E., Petoukhova N., Demidov V. Methodology of Huge Data Volume Processing System Development for Analysis of the Latvian Railway Passengers Transportation. In: *Proceedings of VI International Conference "TransBaltica 2001"*. Riga: RMS Forum, 2001, pp. 201-208. (In Russian)

[7] Korovkin S., Levenets I., Romanov I., and etc. Resolving of Problem of Complex On-Line Analysis of Information in Data Warehouse. – http://www.citforum.ru/database/articles/art_11.shtml (2005, June 22).

[8] Demidovs V. Virtual Models in Decision Support Systems of the Latvian Railway. In: *Transport and Telecommunication,* volume 5(2). Riga: Transport and Telecommunication Institute, 2004, pp. 25-35. (In Russian)

[9] Demidovs V., Bogdanovs J. Regression Models for Forecasting of Traffic Flow in Decision Support System on the Latvian Railway. In: *Abstracts of the International Conference "Reliability and Statistics in Transportation and Communication (RelStat '04)"*. Riga: Transport and Telecommunication Institute, 2004, pp. 37-38.

[10] Allgood G. Mapping Function and Structure for an Anticipatory System: what Impact will it have and is it Computationally Feasible, Today? – http://www.ornl.gov/~webworks/cpr/pres/107928_.pdf (2005, June 22).