Virtual Data Models in Anticipatory System of Railway Transportation

Eugene Kopytov
Transport and Telecommunication Institute
Lomonosova Str.1, Riga, LV-1019, Latvia
Fax: +371 7100660 – e-mail: kopitov@tsi.lv
http://www.lza.lv/scientists/kopytov.htm, http://www.tsi.lv

Vasilijs Demidovs
State Join-Stock Company "Latvian Railway"
Turgeneva Str.21, Riga, LV-1547, Latvia
Fax: (+371)-7234366 – e-mail: dem@ldz.lv

Abstract
The application of Data Warehouse technologies allows constructing the set of prediction models of the railway transportation on the basis of the huge amount of the accumulated data. The big size of the disk space is necessary for these models allocation, therefore the method of the virtual data models is offered. The period of life of the created models is not limited in time, they do not influence the real data, and they do not occupy the place on the storage devices. The examined virtual data models are developed within the frames of the creation of the Decision Support System of Latvian Railway. However the received results have a universal character.
Keywords: anticipatory system, data model, forecasting, railway transportation

1 Introduction

Forecasting has always played an important role in the tasks of managing and planning transportation flows in the Decision Support Systems (DSS) of the railway. This problem became especially acute in European countries at the beginning of the 90s due to social and political changes in the Baltic countries, Russia and Eastern Europe. The change of economic conditions brought about the change of the existing number of factors and of their impact degree and the accumulated statistics became useless since it couldn't explain new regularities and tendencies of the future events development.

The search of the effective solutions of the future railway functioning should be based on the forecast of a great number of characteristics of transportation processes. The external effects which are taken into account during the prediction period of time can be the following: the registration of activity of other carriers inside the country and in the adjacent states, changes of the dollar exchange rate and the prices of transported cargoes and passengers, new road and highways construction etc. As a result, different versions of the anticipatory models have been created, in which various combinations of the effects controlling and influencing the transportation system have been considered.

For this purpose it is offered to create up-to-date Data Warehouse, which can store large amounts of information to be used for prediction [1]. However, the necessity to create a lot of models generates, in its turn, some problems connected with a sharp increase of the data storage volume and of the time needed for the solution of the anticipatory task, therefore the method of the virtual anticipatory models is offered [2]. The given models don't require substantial growth of the disk space because we have to store only metadata of the model in the data repository instead of the traditional approach of the full model storing. It is possible to construct a big multitude of virtual models, which can operate in parallel. The period of life of the created model is not limited in time, they do not influence the real data, and they do not occupy the place on the storage devices. With them it is possible to work at any time and they will always take into account the freshest data and depending on their deviations from the design values they will give the prediction with the correcting impact. The principle of the models construction gives possibility to change control impacts on the transportation system during the anticipatory process. The examined virtual data models are developed within the frames of the creation of the DSS of Latvian Railway.

The methodology of building and improving the Decision Support Systems in the railway suggested by the authors implies synergy of modern technologies and approaches in the sphere of databases and their integration in the existing technologies [3], [4], and [5]. The present paper considers the questions of building informational support of Decision Support Systems employing the technologies of Data Warehouses, Data Marts, intellectual data processing, anticipatory systems and virtual models.

2 Building of Virtual Prediction Models of Railway System

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 $y_i(t)$ is the system's indicator with number i observed at the moment of time t. Let the prehistory of a system is known, i.e. vectors $Y(t_1), Y(t_2), ..., Y(t_m)$ are given, where $t_1, t_2, ..., t_m$ are moments of time in the past and present. Let assume that t_1 is an initial moment of system observation, t_m is a current moment of time, and $t_1 < t_2 < ... < t_m$.

The considered task is to anticipate behaviour of railway system in the future taking in account possible control impacts on the system and environment behaviour. As it was suggested by authors [2], in the general case we can anticipate 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; $Z^*(t_e)$ is the vector of predicted values of random factors, characterizing the impacts of environment on the investigated system in the moment of time t_e.

For obtaining a more reliable forecast of the system behaviour with various combinations of administrative effects and for developing suggestions for the optimum control choice, it is necessary to create some models of evolution of the system in the future. 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 versions 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.

By choosing for each vector $Z_j^*(t_e), j=\overline{1,n}$ the corresponding control strategies $C_j(t_e)$ using model (1) for $Z^*(t_e)=Z_j^*(t_e); C(t_e)=C_j(t_e)$ we can anticipate the system status for the moment of time t_e defined by vector $Y_j^*(t_e)$. 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 model stored in Data Warehouse (see Figure 1). Thus, for n versions of environment behaviour we obtain accordingly forecasts $Y_1^*(t_e), Y_2^*(t_e), ..., Y_n^*(t_e)$ those give opportunity to use the situation control in the considered system.

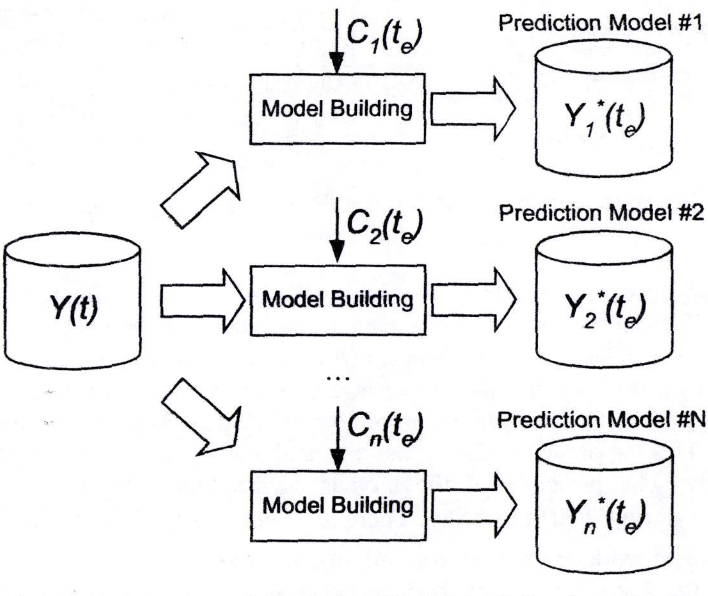

Figure 1: Conceptual scheme of Prediction Models building in anticipatory system.

Naturally, the given method makes stronger demands to used Information Systems, because there is multiversion of stored objects. This approach has some disadvantages. For storage of 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. Here it is necessary to generate new versions of models at variation of control actions $C_j(t_e)$. For solution of the given problems the creation of the virtual models of system which are not demanding a disk space as in a Data Warehouse in addition 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, is offered (see Figure 2) [1].

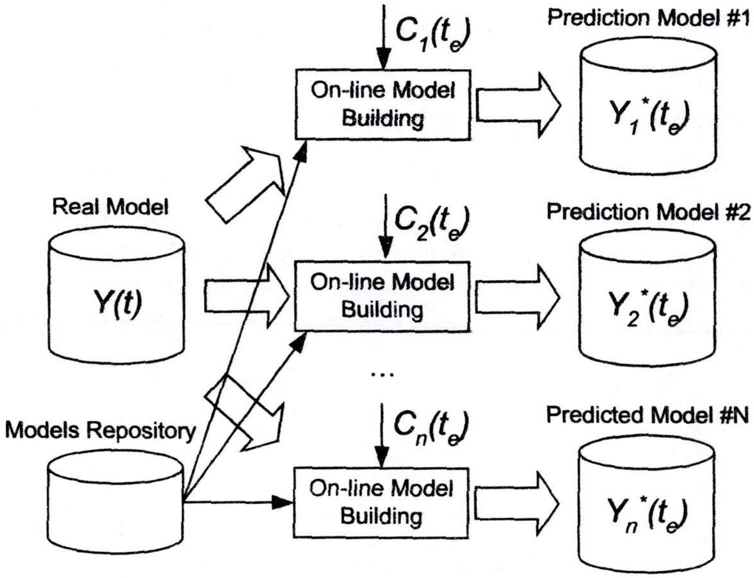

Figure 2: Conceptual scheme of Prediction Models building in anticipatory system using Virtual Data Models.

The second approach is more flexible, taking into account the fact that we can describe all effects at the system in metadata as the functional dependences. Controlling input data in these dependences, we can control effects at the system as a whole on-line. It is achieved by the fact that the model is created at the moment of addressing to it on the basis of its specification incorporated in Models Repository stored in Data Warehouse. This approach allows having and using a set of virtual models simultaneously. The principle of the models construction allows changing control actions $C_j(t_e)$ during their researches. There is no need of blocking the real data model since each analyst works with its own set of virtual models.

Let's note that the suggested method can be used to anticipate railway transportation depending on the change of the tariff policy as the managing impact on the system with the account of the influence of environment. The external influences considered in the forecasting process can appear, in the case of cargo transportation, the account of other carriers' activities within the country and in the neighbouring countries, the change of the dollar rate and of the transported cargoes process, etc.

Examples of practical realization of the approach of virtual data models in anticipatory system on railway are considered in next sections.

3 Application of Virtual Models in Data Transformation

Reliability of prediction in considered anticipatory system depends on the reliability of data stored in Data Warehouses and on quality of data transformation processes. Realization of complicated algorithms of data transformation in DSS and the Executive Information System is connected with the problems of providing the trustworthiness of the received results. In a number of works [5], [6], [7] it is suggested to employ the mechanism of views for data transformation. But there is no any development of this idea in the above works. A *view* is a virtual relation, which is made at the moment of the query entry on the basis of the basic relations [8]. Views give possibility to perform complicated data transformation within the database, they are convenient for the operations of clearing data in operational databases when building Operational Data Store, for transforming data in Data Warehouse, aggregating for Data Mart and On-Line Analytical Processing [7].

But in some cases, to get the required result, very complicated views are needed. Development, testing, optimisation and transformation of such views connected with the future change of the calculations business logics are not a trivial task. Under these conditions provision of the received results trustworthiness also causes big problems. Realization of sophisticated algorithms of data transformation in DSS is connected with the problems of providing the received results reliability. To resolve these tasks at the stage of data transformation, it is suggested to use the method of hierarchically related views making it possible to get a universal mechanism of managing the data transformation processes and giving additional possibilities in working out and testing the processes [9]. The idea of the given method is to split complicated query logics into separate steps – views with a possibility to test each stage. Thus, transforming one major view into several simple ones we get a graph of interrelated views shown in Figure 3.

Dynamic transformation and aggregation of data requires large calculating resources at the moment of applying to the *final view*. It causes a big load on the server and interferes with the work of other applications. Sometimes the logic of data transformation is so complicated that the server resources are not enough even in the monopoly mode. Therefore, in such cases it is reasonable to insert intermediate *materialized views* in the chain of hierarchical views [9], which actuality are supported by special scripts (see Figure 4). This allows to reduce the run-time of some particular queries by tens times.

A materialized view is a relation really existing in the database, which is periodically pre-planned data replenished by the virtual prototype. Replenishment of the basic relation is performed by special utilities:

EXPORT – exporting data into the subsequent file of standard format;

IMPORT or LOAD – loading data from the file to the basic relation.

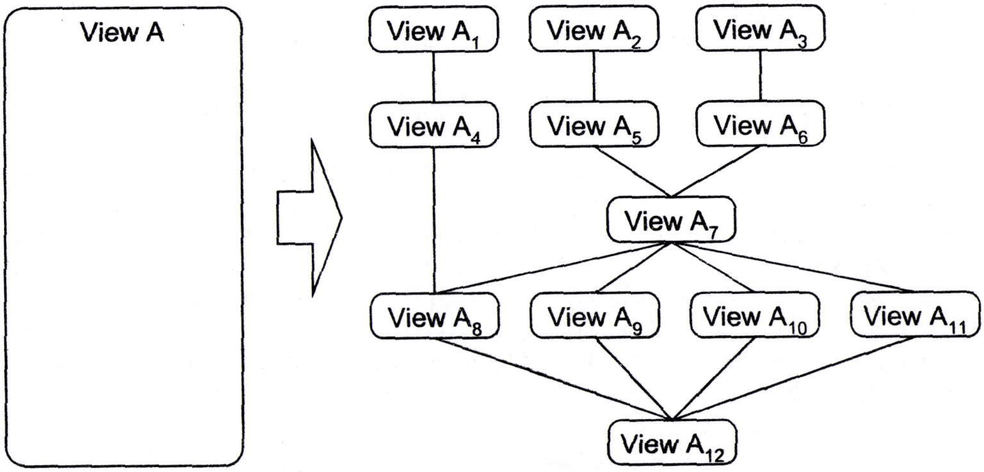

Figure 3: Transformation schema of complicated view.

These utilities are less exigent to the server resources, their work is easy to control and the process of data charging goes much quicker than when employing the query language SQL expression:

INSERT INTO <*table name*> SELECT * FROM <*view name*>.

Using utilities IMPORT and LOAD allows engaging the mechanism of relational integrity of the relational data model. At the moment of performing the IMPORT operation all restrictions are checked and the data, which do not meet the given criteria, are expelled. With the LOAD operation error data are entered into special EXCEPTIONS-relations. After that they can be analysed, corrected and loaded into the corresponding relations.

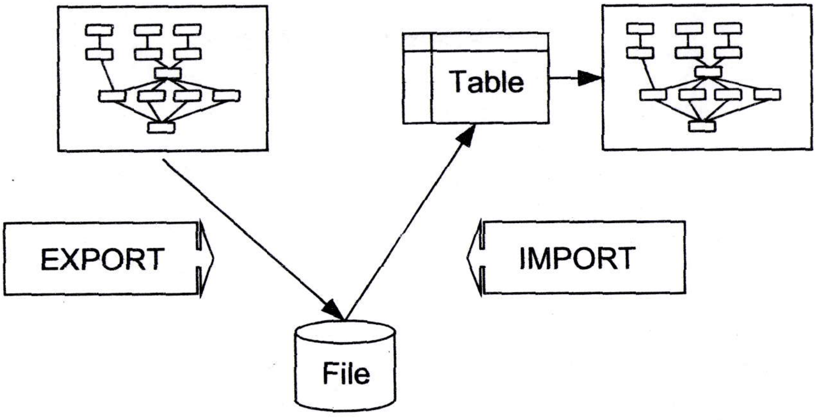

Figure 4: Materialization of intermediate views.

Example of the scheme of data transformation using virtual views is presented in Figure 5. The received relation is used to form a report about the transported passengers in the different sectors and lines of Latvian Railway, there being accounted passengers not only arriving and leaving any sector or line but also those transiting through them non-stop [9].

The given approach allows splitting the process of data transformation into separate stages and steps. At the preliminary stage of the transformation process there is formed the general part of data abstractions of different aggregation level in the form of basic relations. And on these basis there are formed all necessary reports, actual and virtual Data Marts, which are the concern of final users. Analysts can get access to more detailed data of the preliminary stage by performing the drill down operation.

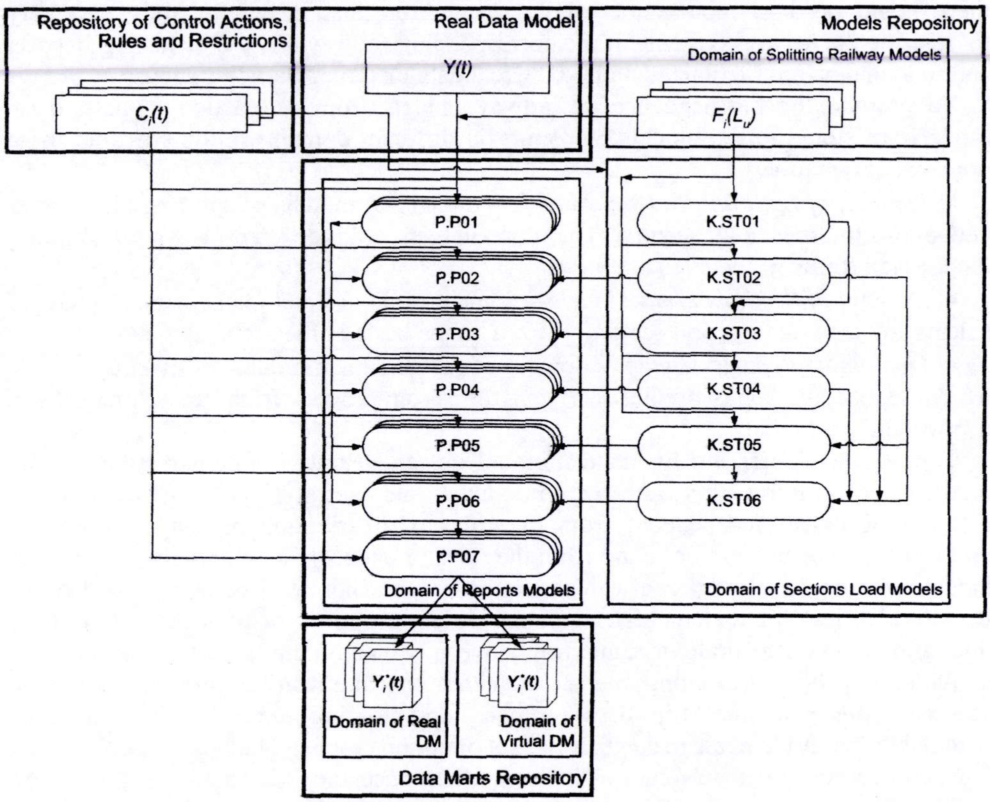

Figure 5: Conceptual model of the system of the analysis of passenger flows.

The developed conceptual system model consists of four fragments:

- *Real Data Model* – cleared and consolidated data of On-Line Transaction Processing (OLTP) systems;
- *Models Repository*, which consists of several model domains:
  - Models of lines and sectors of the railway;

- Models of calculating load on lines and sectors;
- Models of making reports;

• *Repository of Control Action, Rules and Restrictions* – rules and restrictions managing the process of calculations;

• *Data Marts Repository,* which consists of two domains of data marts (DM):
  - *Real Data Marts*, which data physically exist on hard discs;
  - *Virtual Data Marts*, which data are formed at the access moment.

The suggested conceptual scheme of building virtual models using sole and independent from the railway network mathematical model gives opportunity to have several analysed data models concurrently and to create quickly new ones. This approach widens the borders of the system application on other railways as well. And there is no need to rebuild the whole system for this; it will be only necessary to describe the network scheme of the corresponding railway and the managing impacts and to load new transactional data into the real data model in Data Warehouse.

To provide the independence of railway models from calculation models, these models are stored in the models repository in different domains. Let's spell out these domains designation:

• *Domain of Splitting Railway Models* – domain of models of splitting the railway network into lines and sectors. The domain can hold different ways of splitting the Latvian Railway and even other railways.

• *Domain of Sections Load Models* – domain of an abstract mathematical model of calculating load on lines and sectors realized on the basis of hierarchical views.

• *Domain of Reports Models* – domain of mathematical models of making reports, which realize the logic of the data transformation process from Real Data Model to Virtual Data Models.

Thus, in developing a mathematical model we can abstract from a particular railway model and splitting this railway into lines and sectors, as well as develop a universal model independent from the rules of distribution passenger-kilometres on disputable sectors and lines and from the rules of choosing an alternative route. This model specifies the algorithm of calculating passenger-kilometres between two stations. It is divided into logic elements calculating different options of trips through abstract lines and sectors: trip along one line; trip between two adjoining lines; trip between two non-adjoining lines; trip along one sector within one line; trip between two adjoining sectors within one line; trip between two non-adjoining sectors within one line. Altogether theses elements make up a full set of options for calculating any kind of trip.

Such division of models can, in its turn, when necessary (for example, for getting different types of reports) have a single model of splitting the railway network for different mathematical models.

Application of virtual models together with strict system formalization makes it possible to reduce the time of developing new reports and to guarantee their reliability.

4 Application of Virtual Models in Forecasting Tasks of Passenger Transportation

Comprehensive planning of transport company activity demands presence of set of the models that adequately describe functioning of the railway and application of various mathematical methods for forecasting of the passenger flows. As an example, that shows an opportunity of the suggested methods, research of influence of social and economic parameters of Latvian regions development on demand of passenger transport service is carried out.

To analyse and predict the demand for passenger transportation in the regions of Latvia it is suggested to employ the multiple linear regression model of the form:

$$y_i^* = \beta_0^* + \beta_1^* x_{i,1} + \beta_2^* x_{i,2} + \ldots + \beta_n^* x_{i,n}, \tag{2}$$

where y_i^* is the number of railway passengers for the i-th observation – predicted value (estimate) of the dependent variable y; $\beta_0^*, \beta_1^*, \ldots, \beta_n^*$ are estimated regression coefficients – evaluations of the unknown parameters; $x_{i,1}, x_{i,2}, \ldots, x_{i,n}$ – values of independent predictor variables (accompanying factors) $x_1, x_2, \ldots, x_n$ for the i-th observation.

In the course of the suggested research several models of a multiple regression, which allow evaluating the influence of the main social-economic factors on the volumes of passenger transportation by the railway transport in the regions of Latvia, have been received. One of the models built according to the statistic data for the year 2003 for cities and regions of Latvia (except Riga and Jurmala cities and regions of Riga and Ogre) is given by formula:

$$\begin{aligned} y_i^* = {} & 81316 + 16579x_{i,1} - 608246x_{i,2} + 39142x_{i,3} - 73812x_{i,4} - 25477000x_{i,5} + 565904x_{i,6} - \\ & - 88572x_{i,7} + 24809x_{i,8}, \end{aligned} \tag{3}$$

where

y_i^* – predicted value of the dependent variable determining the number of passengers transported in i-th region during the year;

$x_{i,1}$ – the population density in i-th region;

$x_{i,2}$ – the number of enterprises per a unit of territory in i-th region;

$x_{i,3}$ – the number of enterprises per 1000 residents in i-th region;

$x_{i,4}$ – the density of the unemployed population in i-th region;

$x_{i,5}$ – the number of schools per a unit of i-th region's territory;

$x_{i,6}$ – the number of buses per a unit of i-th region's territory;

$x_{i,7}$ – the number of buses per 1000 residents in i-th region;

$x_{i,8}$ – the number of railway stations in i-th region.

On the basis of the suggested virtual models' approach there appears a flexible possibility of realizing the conceptual schema of prediction for multiple linear regression models considered above, shown in Figure 6. The suggested approach gives a prompt possibility of a simultaneous employment of the developed Prediction Models, which are kept in the Domain of Prediction Models of Models Repository.

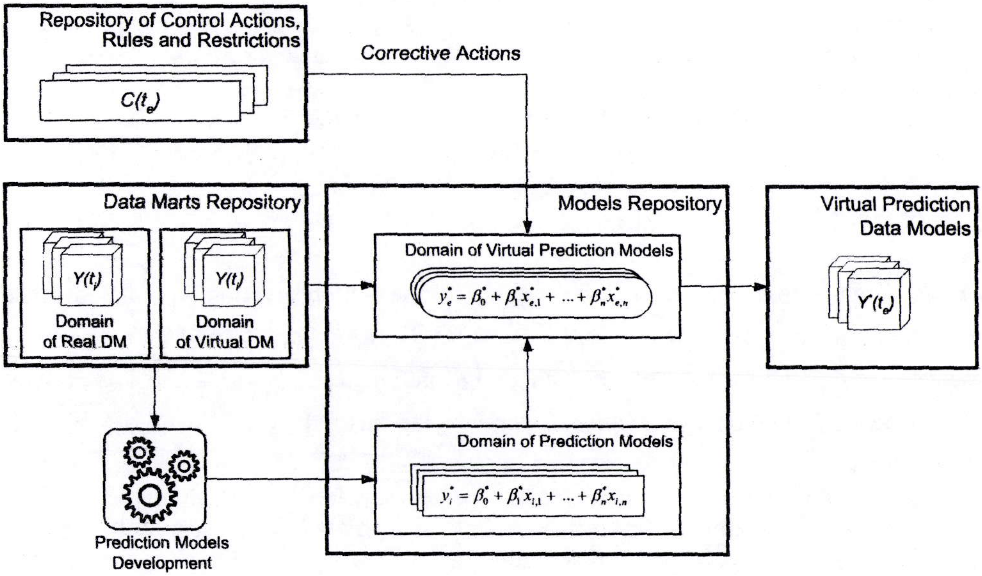

Figure 6: Conceptual schema of Virtual Prediction Models.

On the basis of these models in the Domain of Prediction Models there are built Virtual Prediction Models and placed in the Domain of Virtual Prediction Models. Virtual models are actually templates of the mathematical prediction models with the calculated values of evaluations of unknown parameters, the values of predictor variables are taken from the Data Marts Repository at the time of carrying out the analysis, and it allows executing prediction for different periods of time. 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 to take account of any suppositions in predictions or predictions of change of the analysed predictor variables, thus realizing the anticipation of the type "what will be if..." or "what would be if...".

5 Conclusions

The suggested method of building virtual models has been used in solving different tasks connected with data processing and forecasting in Information-Analytical Systems of the Latvian Railway. There have been considered several practical questions on data transformation and business logics realisation in the task of analysing passenger flows along lines and sectors in the system of databases. There has been shown the possibility of using virtual models in building Data Marts for the tasks of predicting passenger transportation. 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.

References

[1] Kopytov E., Demidovs V., Petoukhova N. (2004). Principles of Creating Data Warehouses in Decision Support Systems of Railway Transport. *Computing Anticipatory Systems: CASYS'03 – Sixth International Conference.* Ed. by Daniel M. Dubois, Published by The American Institute of Physics, AIP Conference Proceedings 718, pp. 497-507.

[2] Kopitovs J., Demidovs V., Petoukhova N. (2003). Virtual Models in Forecasting Systems of Railway Transportation. *Proceedings of the International Conference "Modelling and Simulation of Business System".* Ed. by H. Pranevicius, E. Zavadskas and B. Rapp. –Kaunas: Technologija, Lithuania, pp. 265-268.

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

[4] Simon A.R. (1995). *Strategic Database Technology: Management for Year 2000.* Morgan Kaufmann Publishers, Inc., San Francisco, California.

[5] Chaudhuri S., Dayal U., Ganti V. (2001). Database Technology for Decision Support Systems. *IEEE Computer,* Nr.1, pp. 48-55.

[6] Schaffer S. A Relational Approach to Data Transformations. (2000). *DB2 magazine*, Volume 5(3), – San Francisco: CMP Media, pp. 46-54.

[7] Smith J., Li C., Castelli V., and Jhingran A. (1998). Dynamic Assembly of Views in Data Cubes. *Proceeding of the Symposium on Principles of Database Systems (PODS).* – Seattle, Washington, pp. 274-283.

[8] Connolly T., Begg C. (2001). *Database Systems: A Practical Approach to Design, Implementation and Management.* Third Edition. – Harlow: Addison Wesley Longman.

[9] Demidovs V. (2003). Use of Methods for the Increase of Data Reliability in Decision Support Systems of the Latvian Railway. *Transport and Telecommunication,* Volume 4(1). – Riga: Transport and Telecommunication Institute, pp. 16-22. (in Russian)