Agris on-line Papers in Economics and Informatics

Volume VII Number 3, 2015

Transformation Econometric Model to Multidimensional Databases to Support the Analytical Systems in Agriculture

J. Tyrychtr, A. Vasilenko

Department of Information Technologies, Faculty of Economics and Management, Czech University of Life Sciences Prague, Czech Republic

Anotace

Aplikace ekonometrických modelů v zemědělských podnicích je velmi složitý proces vyžadující znalost nejen ekonomie, ale i statistických a matematických metod samotnými pracovníky v zemědělství. Řešením může být aplikace ekonometrických problémů do analytických systémů pro podporu rozhodování manažerů farem. Pro takové řešení je nutné navrhnout multidimenzionální databázi pro podporu online analytického zpracování dat (OLAP). V článku je představena nová metoda (nazvaná TEM-CM) pro formální transformaci ekonometrického modelu do konceptuálního modelu dat pro tvorbu multidimenzionálních schémat. Tato nová metoda umožňuje formalizovat proces převodu produkční funkce v zemědělství na multidimenzionální model dat a přispívat tak k efektivnějšímu návrhu datových skladů a OLAP databází pro podporu rozhodování v zemědělských analytických systémech.

Klíčová slova

Multidimenzionální databáze, OLAP, ekonometrický model, produkční funkce, konceptuální návrh, zemědělství.

Abstract

Econometric model application in farms is a very complex process requiring knowledge not only the economy but also statistical and mathematical methods in agriculture workers themselves. The solution may be an application of econometric problems in analytical decision support systems for farms managers. For such a solution is necessary to design a multidimensional database for support online analytical data processing (OLAP). This paper proposes a new method (called TEM-CM) for formal transformation of econometric model to the conceptual data model for creating multidimensional schemes. This new method allows to formalize the process of transferring production function in agriculture to multidimensional data model and thus contribute to a more efficient design of data warehouses and OLAP databases for decision support in the agricultural analytics systems.

Key words

Multidimensional database, OLAP, econometric model, production function, conceptual design, agriculture.

Introduction

Analytical systems in the agricultural sector can help farm businesses to enhance their production potential and production efficiency by the ability to effectively support the management, analysis, planning and decision-making activities of managers and specialists. For example, some analytical issues in the context of precision agriculture (Lips et al., 2013, Cabrera Garcia et al., 2013), econometric (Čechura, Taussigová, 2013, Čechura, 2010, Kroupová, 2010) and information technology (Schulze et al., 2007, Rai et al., 2008) can be easily implemented using the concepts of Business Intelligence. Interdisciplinary approach to agriculture requires very high standards for data management. Special attention should be dedicated to the development of operational and analytical data to support the use of OLAP (Online Analytical Processing). OLAP describes an approach to decision support, which aims to gain knowledge from the data warehouse, or more precisely, from data marts (Abelló, Romero, 2009). There are currently several approaches to store of analytical

data. Among the most important are called multidimensional, relational, hybrid or desktop OLAP (more on this subject deals (Burstein, Holsapple, 2008)).

In this paper, a new method TEM-CM stage data for OLAP is introduced. The aim of the method is to allow formal transformation of econometric model to the conceptual data model for creating multidimensional schemes. This new method allows to formalize the process of transferring production function in agriculture to multidimensional data model and thus contribute to a more efficient design of data warehouses and OLAP databases for decision support in the agricultural analytics systems. In the OLAP approach, data are stored in an analytic database using a special scheme instead of the traditional relational schema. This approach is different compared to the modeling of operational OLTP databases (Online Transaction Processing). At the conceptual and logical level can be OLAP defined by three basic activities (Abelló, Romero, 2009):

- dimension analysis and modeling,
- data warehouse modeling,
- and realization of a dimension changes.

Multidimensional modeling (Burstein, Holsapple, 2008, Novotný et al., 2005, Datta, Thomas, 1999, Codd et al., 1993) is a basic part of OLAP solution. At the farm, the OLAP databases are rather exception. However, it is currently possible to find literature focused on an OLAP databases in agriculture. There are papers (Pardillo et al., 2010, Schulze et al., 2007), in which are described the proposals of OLAP database application in the agricultural context.

The authors of the proposal dealing with OLAP databases, however, do not consider the proposal in the context of production (or cost) functions in agriculture. The production function represents the relationship between the size of inputs (factors of production) and the size of the farm production output. Based on the identified production function, it is possible to create a multidimensional database for the collection of relevant and objective data. However, for the creation of such a database is necessary to transform the production function to conceptual data model.

This reason of the production function in the OLAP solution is that the farm can perform analytical processing of summarized and aggregated data to help answer questions such as the:

- How big was the total production [in thousands of CZK] farm in 2011?
- How big was the milk production [kg] on the farm in 2011?
- How big was the milk production [kg] for Holstein cattle in March 2011?
- Knowledge production function within the Business Intelligence (Tyrychtr et al., 2015) can also help agricultural enterprises in matters:
- How much will change production when you change the workforce of unit?
- How much will change production when you change the acreage of cultivated land unit?

All of these questions can be answered effectively by using OLAP technology. For most database is necessary to formulate model step by step, first step should be to define its structure in general. The overall task of designing a database is to map the real world application into formal data model of the database management system (Rai et al., 2008).

Database design is a process that produces a series of database schema for the particular application (Fahrner, Vossen, 1995).

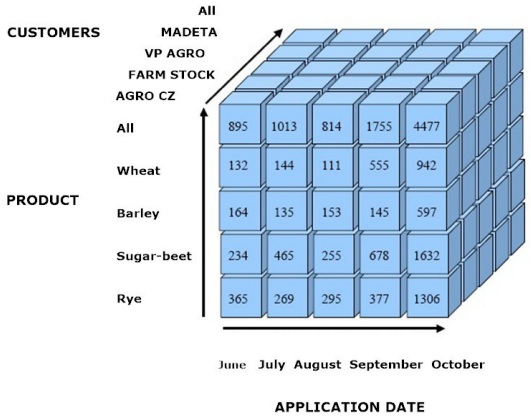
Multidimensional modeling (Mylopoulos, 2009) is the process of data modeling in the universe of discourse with the modeling structure to provide a multidimensional data model. Multidimensional models categorize data, either as facts associated with numerical measure or as dimensions that characterize the facts and are generally in plain text.

Facts are objects that represent the object of required analysis to be analysed to better understand his behaviour. Multidimensional models are currently mostly based on a relational data model, or on the construct of data cubes (Pedersen, 2009a).

Data cube (figure 1) is a data structure for storing and analysing large amounts of multidimensional data (Pedersen, 2009b). It is generally interpreted as a basic logical structure to describe multidimensional databases, as well as relations for the relational databases. Data cube represents the abstract structure but, unlike traditional relational structures in the relational data model is not defined clearly. There are many approaches to the formal definition of the operators' data cubes (comprehensive overview is available on the paper (Vassiliadis, Sellis, 1999). Generally we can tell, that the data cube is constructed of dimensions and measures.

Dimension is a hierarchical set of dimensional values that provide categorical information characterizing a particular aspect of data (Pedersen, 2009c).

Measures (monitored indicators) of data cube are mostly quantitative data that can be analysed.



Source: self-authored

Figure 1: An application context of data cubes.

Materials and methods

In this paper we first formalizes the notation of econometric model and multidimensional data model through mathematical apparatus. Subsequently, we propose a new method of TEM-CM through formal rules. Chapter Conclusion will be focused on a new method of TEM-CM demonstrated by the application of individual rules for the econometric model in analytical systems. Finally, conclusions are formulated article and opportunities for further research proposal OLAP in an agricultural context.

Econometric model

Econometric model (ECM) represent mathematical model, which is a mathematical-statistical formulations of economic hypotheses. It expresses the dependence of economic variables on variables that explain the hypothesis. Most often in the economic literature (Kroupová, 2010) is used Cobb-Dougles production function, which can be characterized by constant elasticity of production factors, invariability in economies of scale among businesses and convexity isoquant function towards the beginning. Cobb-Dougles production function has general form (Felipe, Adams, 2005, Kroupová, 2010):

$$y = \alpha x_l^{\beta_j} x_p^{\beta_p} x_k^{\beta_k}, \text{ where} \quad (1)$$

y… amount of output,

$x_{l,p,k}$ … amount of *l-th, p-th* a *k-th* input,

α, β … parameters of production function.

In the agricultural environment ECM is often composed of more than one equation. There are stochastic equations with random variable and definition equation (identity function).

If we have ECM in symbolic form (Tvrdoň, 2006):

$$y_{1t} = \gamma_{11}x_{1t} + \gamma_{12}x_{2t} + \gamma_{13}x_{3t} + \gamma_{14}x_{4t} + u_{1t}$$
$$y_{2t} = \beta_{21}y_{1t}\gamma_{21}x_{1t} + \gamma_{25}x_{5t} + u_{2t}$$
$$y_{3t} = y_{1t} + y_{2t}, \text{ then} \quad (2)$$

y_s is an endogenous s-type variable and its value in the period *t*- $y_{st,}$ index s = (1, 2, … g), t = (1, …, n). x_r is r-th exogenous variable with value in period *t* - x_{rt}, where number of exogenous variables is equal to k, then r = (1, 2, … k). Delayed endogenous variable express effects of variables of period *t-z*, where z = (1, 2, …, t-z). u_{st} is random variable in s-th equation of explained endogenous variable in period t. β_{is} is structural parameter in i-th equation in s-th model undelayed endogenous variable and γ_{ir} in i-th equation of model of r-th predetermined variable.

Multidimensional model

Data cube is basic structure of multidimensional databases and it is used as basic amount of inputs and output for every operators based on multidimensional databases (Datta, Thomas, 1999).

If we have quaternion <*D,M,A,f*>, where all four components indicate the properties of the data cube. Then, these properties are (Datta, Thomas, 1999):

Set on *n* dimension $D = \{d_1, d_2, ..., d_n\}$, where every d_i in name of dimension obtained from domain $dom_{dim(i)}$.

Set of *k* measures $M = \{m_1, m_2, ..., m_k\}$, where every m_i in name of measure obtained from domain $dom_{measure(i)}$.

Set of names of dimension and measures is disjoint; tj. $D \cap M = 0$.

Set of *t* attributes $A = \{a_1, a_2, ..., at\}$, where every a_i is name of attribute obtained from domain $do_{mattr(i)}$.

Viewing one to many $f: D \rightarrow A$, exists for each dimension and each set of attributes.

The view is such that a set of attributes associated to the dimension in the pair are disjoint i.e. $\forall i,j,i \neq j, f(d\_i) \cap f(d\_j) = 0$.

Constellation schema

Like the design of conceptual schema operational database is useful for the design of multi-dimensional diagram make use of some type of design approach. The most commonly are used style star schema, Snowflake schema and constellation diagram (or galaxy/integrated/ hybrid scheme). This article uses the constellation diagram (Kimball, 1998), which is formed by the dimension tables shared by two or more fact tables. This is an equivalent star schema or flakes, with the difference that this scheme consists of several fact tables. The constellation schema is obtained when the multidimensional model is composed of two or more cubes, which eventually share some dimensions (Boulil et al., 2014).

Results and discussion

To create a new formal method TEM-CM (Transformation of Econometric Model to the Conceptual Model) is the first to define the representation of the econometric model and multidimensional representations of the proposal scheme, which is based on a constellation diagram type.

Formal representation of econometric model

Let us set Y a X, where

$Y = \{y_s\} \cup \{y_{st}\}$ is a finite set of endogenous variables in the model.

$X = \{x_r\} \cup \{x_{rt}\}$ is a finite set of exogenous variables in the model.

$Rel \subseteq (X \times Y) \cup (Y \times Y)$ is a set of structural relations.

Formal representation of multidimensional schema proposal

Schema Constellation is defined by five elements (*Ent, Key, Att, Ass, getKey*), where:

Ent is a finite nonempty set of data model entities,

Key is a finite nonempty set of data model keys,

Att is a finite nonempty set of data model attributes,

$Fact \subseteq Ent$ is a finite set of separate entities out of constellation scheme

$Dim \subseteq Ent$ is a finite set of entity dimensions

Each entity $e \in Ent$ is described by collection of keys and collection of attributes, i.e. accordingly we can assume that:

$\forall e \in Ent: \exists(\{k \in Key\} \cup \{a \in Att\})$

getKey is function, which returns entity key in constellation schema, i.e. accordingly we can assume that:

$\forall e \in Ent: getKey(e): Ent \rightarrow Key_e \subseteq Key$

$Ass \subseteq (Dim \times Fact)$ is finite set of relation between entities.

New methods TEM-CM

For the defined sets and conditions are proposed following transformation rules, which are divided into two phases. The first phase of transformation provides dimension and facts from econometric variables into the constellation diagram. The second phase is building relationships between dimensions and facts.

Phase 1: Constellation scheme creation

Rule 1.1: Creating facts table in empty constellation scheme for every endogenous variable in econometric model.

$\forall y_s \in Y: c_s\ Fact$ and $\forall y_{st} Y: c_{st} \in Fact$

Rule 1.2: Creating dimensions to constellation scheme for every exogenous variable in econometric model.

$\forall x_r \in X: ds \in Dim$ and $\forall x_{rt} \in X: d_{rt} \in Dim$

Rule 1.3: If there exists time variable in econometric model, we have to create time dimension

$\forall x_{rt} \in X: d_{rt} \in Dim_{time}$

Phase 2: Definition of relationship between entities in conceptual model

Rule 2.1: If there is a relationship between exogenous variable x and endogenous variable y and function getKey that returns a set of keys for these variables, then it will be associations between the fact table and the dimension to which they apply:

$$\forall (x,y) \in Rel: (d,c,K) | (d \in Dim) \wedge (c \in Fact) \wedge ((d,c) \in Ass) \wedge (K \subseteq K_d \cup K_c | (K_d = getKey(d)) \wedge (K_c = getKey(c)))$$

Application of new method TEM-CM

Formally defined rules are demonstrated in the application context. The following example represents a situation where the total production

of the farm is dependent on plant production and livestock production and for each of these productions are monitored difference measurement (indicators).

Let us ECM (2) and simplified case study:

y_{1t} … gross plant production in the period t,

y_{2t} … gross livestock production in the period t,

y_{3t} … gross agriculture production in the period t,

x_{1t} … amount of subsidies,

x_{2t} … fundamental production funds in plant production,

x_{3t} … the amount of labour in plant production,

x_{4t} … weather condition,

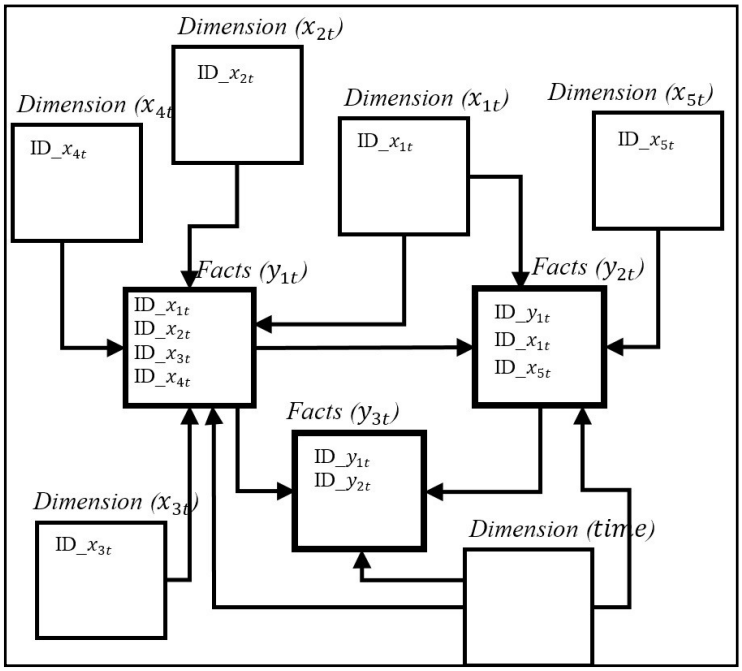
x_{5t} … livestock numbers,

u_{1t}, u_{2t} … random component in the period t,

In first phase is defined fact table into empty constellation scheme for gross agriculture production, gross plant production and gross livestock production (rule 1.1). In next steps are defined dimensions for every exogenous variable in econometric model (subsidies, plant production funds, labour amount in plant production, weather conditions, and number of livestock). If model (2) contain time variable *t,* we have to define time dimension into our model. In second phase (rule 2.1) is defined association between fact table and dimension by generated key. For example equation $y_{2t} = \beta_{21}\, y_{1t}\, \gamma_{21}\, x_{1t} + \gamma_{25}\, x_{5t} + u_{2t}$ express amount of subvention and number of livestock has relation with gross livestock production (fact table y_{2t}). Random parts u_{1t}, u_{2t} (eventually another different variables out of TEM-CM rules) cannot be transformed. For example in application context is such equation express in form: $y_{1t} = 3.45x_{1t} + 1.32x_{2t} + 1.07x_{3t} + 0.43x_{4t} + 284.36$. So random variables u_{1t}, u_{2t} and parameters β, γ are represented by numbers and it is not necessary to contain it to our scheme. Result of TEM-CM method is represented in figure 2.

Conclusion

In paper has been introduce new method TEM-CM for econometric model transformation into conceptual model (constellation scheme) by mathematical apparatus. TEM-CM method represent formalized rules. These rules are to automate the process of designing a conceptual model and streamline the decision-making for agricultural managers. Analytic system can be established with econometric support. Method TEM-CM provide formalized rules for conceptual model for analytic data in agriculture business. This method has impact on analytic expert systems and their application in agriculture. Method TEM-



Source: self-authored

Figure 2: Result of TEM-CM method.

CM represent partial method in whole OLAP design. In future research can be TEM-CM used for process design to make fully automated conceptual model. Questionable also remain current methods for designing logical and physical schema that will be also part of future research. The effort is to streamline and simplify the design process analytical systems with support of econometric for farms, especially considering the current high cost of investment in Business Intelligence.

Acknowledgements

The results and knowledge included herein have been obtained owing to support from the Internal grant agency of the Faculty of Economics and Management, Czech University of Life Sciences in Prague, grant no. 20141040, "New methods for support of managers in agriculture".

Corresponding author:
Ing. Jan Tyrychtr, Ph.D.
Department of Information Technologies, Faculty of Economics and Management
Czech University of Life Sciences in Prague, Kamýcká 129, Prague 6 – Suchdol, Czech Republic
E-mail: tyrychtr@pef.czu.cz

References

[1] Abelló, A., Romero, O. Encyclopedia of Database Systems. Ling L., Özsu, T. M. USA: Springer US, On-Line Analytical Processing. 2008, p. 1949-1954. ISBN 978-0-387-35544-3.

[2] Boulil, K., Le Ber, F., Bimonte, S., Grac, C., Cernesson, F. Multidimensional Modeling and Analysis of Large and Complex Watercourse Data: An OLAP-Based Solution. Ecological Informatics. 2014, vol. 24, p. 90-106. ISSN 1574-9541.

[3] Burstein, F., Holsapple, C. Handbook on decision support systems. 2008, Springer. ISBN 978-3-540-48712-8.

[4] García, S. C., Tamayo, J. E. I., Carbonell-Olivares, J., Cabrera, Y. P. Application of the Game Theory with Perfect Information to an Agricultural Company. Agricultural Economics. 2013, Vol. 59, No. 1. ISSN 0139-570X.

[5] Čechura, L. Estimation of Technical Efficiency in Czech Agriculture with Respect to Firm Heterogeneity. Agricultural Economics. 2010, Vol. 56, p. 183-191. ISSN 0139-570X.

[6] Čechura, L., Taussigová, T. Avian Influenza and Structural Change in the Czech Poultry Industry. Agricultural Economics. 2013, Vol. 59, No. 1. ISSN 0139-570X.

[7] Codd, E. F., Codd, S. B., Salley, C. T. Providing OLAP (on-Line Analytical Processing) to User-Analysts: An IT Mandate. 1993, Codd and Date, Vol. 32.

[8] Datta, A., Thomas, H. The Cube Data Model: A Conceptual Model and Algebra for on-Line Analytical Processing in Data Warehouses. Decision Support Systems. 1999, Vol. 27, No. 3, p. 289-301. ISSN 0975-8887.

[9] Fahrner, Ch. Vossen, G. A Survey of Database Design Transformations Based on the Entity-Relationship Model. Data & Knowledge Engineering. 1995, Vol. 15, No. 3, p. 213-250. ISSN 0169-023X.

[10] Felipe, J., Adams, F. G. "A Theory of Production" the Estimation of the Cobb-Douglas Function: A Retrospective View. Eastern Economic Journal. 2005, Vol. 31, No. 3, p. 427-445. ISSN 0094-5056.

[11] Kimball, R. The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses. 1998, John Wiley & Sons. ISBN 978-0-471-25547-5.

[12] Kroupová, Z. Technická efektivnost ekologického zemědělství České republiky. Ekonomická Revue. 2010, Vol. 2, p. 61-73. ISSN 1212-3951.

[13] Lips, M., Schmid, D., Jan, P. Labour-use Pattern on Swiss Dairy Farms. Agricultural Economics. 2013, Vol. 59, No. 4. ISSN 0139-570X.

[14] Mylopoulos, J. Database design. 2009, Ling L., Özsu, T. M., Springer US, p. 708-710. ISBN 978-0-387-35544-3.

[15] Novotný, O., Pour, J., Slánský, D. Business Intelligence: Jak využít bohatství ve vašich datech. Prague: Grada Publishing, 2005. ISBN 80-247-1094-3.

[16] Pardillo, J., Mazón, J.-N., Trujillo, J. Extending OCL for OLAP Querying on Conceptual Multidimensional Models of Data Warehouses. Information Sciences. 2010, Vol. 180, No. 5, p. 584-601. ISSN 0020-0255.

[17] Pedersen, T. B. Encyclopedia of Database Systems. Ling L., Özsu, T. M. USA: Springer US, 2009a, Multidimensional Modeling, p. 1777-1784. ISBN 978-0-387-35544-3.

[18] Pedersen, T. B. Encyclopedia of Database Systems. Ling L., Özsu, T. M. USA: Springer US, 2009b, Cube, p. 538-539. ISBN 978-0-387-35544-3.

[19] Pedersen, T. B. Encyclopedia of Database Systems. Ling L., Özsu, T. M. USA: Springer US, 2009c, Dimension, p. 836-836. ISBN 978-0-387-35544-3.

[20] Rai, A., Dubey, V., Chaturvedi, K. K., Malhotra, P. K. Design and Development of Data Mart for Animal Resources. Computers and Electronics in Agriculture. 2008, vol. 64, No. 2, p. 111-119. ISSN 0168-1699.

[21] Schulze, Ch., Spilke, J., Lehner, W. Data Modeling for Precision Dairy Farming within the Competitive Field of Operational and Analytical Tasks. Computers and Electronics in Agriculture. 2007, Vol. 59, No. 1, p. 39-55. ISSN 0168-1699.

[22] Tvrdoň, J. Ekonometrie (Econometric). 2006, Prague. Czech University of Life Sciences Prague. ISBN 80-213-0819-2.

[23] Tyrychtr, J., Ulman, M., Vostrovský, V. Evaluation of the State of the Business Intelligence among Small Czech Farms. Agricultural Economics. 2015, Vol. 61, No. 2, p. 63-71. ISSN 0139-570X.

[24] Vassiliadis, P., Sellis, T. A Survey of Logical Models for OLAP Databases. ACM Sigmod Record. 1999, Vol. 28, No. 4, p. 64-69. ISSN 0163-5808.