Towards Econometric OLAP Design in the Intelligence System

JAN TYRYCHTR[1], IVAN VRANA[2]
[1]Department of Information Technologies
[2]Department of Information Engineering
Czech University of Life Sciences in Prague
Kamýcká 129, 165 00 Praha 6 - Suchdol
CZECH REPUBLIC
[1]tyrychtr@pef.czu.cz, [2]vrana@pef.czu.cz

Abstract: The econometrics is currently a perspective area of enterprises growth and economy which is based on application of econometric models requires a good knowledge of mathematics and statistics. To better support decision making in the areas of economic analysis as much as possible, it is necessary to apply econometric problems with intelligent decision support systems for managers and experts. In this paper we focused on the improvement of the method of transformation to better support the design of the analytical (multidimensional) databases of econometric based intelligence systems. We create formal rules of the new TEM method for transformation econometric model through the using of mathematical notation. Based on the proposed approaches to the transformation of an econometric model to the multidimensional schema paying attention to measurement of quality and we compare it with original TEM-CM method. Application of innovative TEM method is illustrated in agriculture context. We also create a prototype of multidimensional OLAP databases and applications to ensure econometric analyses.

Key-Words: OLAP, multidimensional design, database design, econometrics, econometric system, analytical system

1 Introduction

The area of econometrics has developed rapidly in the last three decades and its use can be found in a number of areas such as determining the level of interest rates, estimates of the price elasticity of oil demand, production analysis of business, etc.

Econometrics has become an interesting tool that provides the ability to extract useful information about important business matters of the company and the economy. However, application of econometric models is a nontrivial process that requires a good knowledge of mathematics and statistics. To support decision making in the areas of economic analysis as much as possible, it is necessary to apply econometric problems with intelligent decision support systems for managers and professional and expert staff at various levels of management.

To cope with this problem, several attempts have been made in the last two decades. One of the first studies [10] examines the feasibility of developing an artificially intelligent econometrician as an active decision support system. Brown et al. [7] described the econometric based system to estimate daily cotton market prices. Brandl et al. [5] used genetic algorithm for automated econometric decision support system on foreign exchange market. Various approaches were extensively used, mainly to solve automation econometric methods (e.g., [3], [29], and [21]). Previous studies no focused on the actual design of the databases of these econometric systems and even on the use of new approaches to Business Intelligence concepts.

First such effort was the development of the TEM-CM method [26] used for formal transformation of econometric models to the conceptual model of a multidimensional database as the basis for an econometric system based on online analytical processing. There is a number of shortcomings. Our research addresses the improvement of the method of transformation to better support the design of the multidimensional databases of econometric based systems.

Application of this method is illustrated in a case study on agriculture. Authors dealing with the design of the analytical system in agriculture ([15], [20], [24], [14], [17], [2], [4], [27], and [12]) have yet to consider the proposal of econometric functions. Yet econometric analysis (e.g. [6], [9], and [18]) for agricultural businesses presents a great potential for the improvement of their production and technical efficiency.

A common problem is how to create online analytical solutions for such econometric analysis of data in intelligent systems. Therefore our aim is to create an innovative method for transformation of econometric models to analytical database (so-called TEM method).

2 Theoretical Background

In this section we provide the relevant theoretical background of the online analytical processing and econometric models.

2.1 The OLAP

Online Analytical Processing (OLAP) represents an approach for the decision support which aims to gain information from the data warehouse or data marts [1]. OLAP allows the aggregation of data and inspection of the indicators from different points of view. OLAP gains aggregated data by grouping various analytical data from a multidimensional database.

Multidimensional data analysis is based on the fact that decision-makers need aggregate data related to a particular topic which will be also assessed on the basis of certain factors. Aggregated data are typically modelled as a data cube which is the default model for OLAP. Data cube is a data structure for storing and analysing large amounts of multidimensional data [19]. Data cube allows utilizing benefits of multidimensional view on data and through OLAP operators such as roll-up, drill-down, slice-dice, pivoting etc., perform OLAP questions.

There are many approaches to the formal definition of operators' data cubes (a comprehensive overview can be found in [28]). Generally, the data cube consists of dimensions and measures. Dimension represents the factors based on which analysis of summarized data is carried out. Analysts often need to group the data and therefore to assess each dimension at different levels of detail. For this reason it is important to organize the data into multi-dimension hierarchies. Dimension hierarchy specifies aggregation levels and granularity. For example the Time dimension can be defined as a multi-level hierarchy of levels day → month → quarter → year. Measures (monitored indicators) of the cubes are mainly quantitative data that can be analyzed. Common examples include sales, profits, revenues and costs.

2.2 Econometric Models

Econometric model (EM) represents a mathematical model which is a mathematical-statistical formulation of economic hypotheses. It expresses the dependence of economic variables on variables that explain the hypothesis.

Most often used in the economic literature is 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 (e.g., [11]):

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

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 enterprise environment EM is often composed of more than one equation. There are stochastic equations with random variables and identity equations in the model.

In a standard linear model, mathematically [25]:

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

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). Time-delayed endogenous variable expresses effects of variables of period $t - z$, where z = (1, 2, …, t−z). u_{st} is a random variable in s-th equation of explained endogenous variables 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.

3 Methodology

In this section we describe the methodology to design OLAP solution for econometric analysis.

First we create formal rules of the new TEM method for transformation econometric model through the use of mathematical notation. We measure of quality of proposed approach and we compare it with original TEM-CM method. For the quality assessment of the schemas we use a measurement of quality in data warehouse by [8], [13], and [23]. The measurement for data warehouse are following:

- NFT(Sc). Defined as a number of fact tables of the schema.
- NDT(Sc). Number of dimension tables of the schema.
- NSDT(Sc). Number of shared dimension tables. Number of dimension tables shared for more than one star of the schema.
- NAFT(Sc). Number of attributes of fact tables of the schema.
- NADT(Sc). Number of attributes of dimension tables of the schema.
- NASDT(Sc). Number of attributes of shared dimension tables of the schema.
- NFK(Sc). Number of foreign keys in all the fact tables of the schema.

Application of innovative TEM method is illustrated in agriculture context. Finally we create a prototype of multidimensional OLAP databases and applications to ensure econometric analyses. We use Microsoft Excel 2013 and MS PowerPivot.

4 The TEM method design

In this section we create a formalized new method TEM for transformation econometric models to star schema of multidimensional database.

4.1 Formal representation

For a formal definition of the rules of the TEM method, let us have a set *Y* and set *X*, where:

$Y = \{y_s\} \cup \{y_{st}\}$ is a finite set of endogenous variables,

$X = \{x_r\} \cup \{x_{rt}\}$ is a finite set of exogenous variables and

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

Schema *Star* is any set with 5 elements (*Ent*, *Key*, *Att*, *Ass*, *getKey*), where:

Ent is a non-empty finite set of entities in the schema,

Key is a finite non-empty set of keys in the schema,

Att is a finite non-empty set of attributes in the schema,

$Fact \subseteq Ent$ is a finite set of facts in the schema,

$Dim \subseteq Ent$ is a finite set of dimensions in the schema,

$Measure \subseteq Fact$ is a finite set of measurements in the schema.

Each entity $e \in Ent$ is described by the collection of keys and attributes

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

$getKey$ is a function that returns the key entities in the star schema

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

$Ass \subseteq (Dim \times Fact)$ is a finite set of relationships of the entities.

4.2 The design of rules for TEM method

Phase 1: Creation of the basic star schema.

Rule 1.1: Creation of measurements into an empty schema for each endogenous variable of the econometric model is defined by:

$\forall y_s \in Y : m_s \in Measure$ and $\forall y_{st} \in Y : m_{st} \in Measure$

Rule 1.2: The creation of the dimension into the schema for each exogenous variable in the econometric model is defined by:

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

Rule 1.3: If there is a time variable in the econometric model, then we create time dimension:

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

Phase 2: Creation of relations between entities in the star schema.

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 to these variables, then we create associations between a fact and a dimension:

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

4.3 Application of rules of the TEM method

To verify the rules, we consider the econometric model (2) and thereto simplified semantic context of the example in agriculture:

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

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

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

x_{1t} … the amount of grants (subsidies),

x_{2t} … basic production funds in the crop production,

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

x_{4t} … climatic conditions,

x_{5t} … number of livestock,

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

The example is a situation where the total production of the agricultural company is dependent on crop production and livestock production. For each of these three endogenous variables we should observe the different measures.

In the first phase we create measurements in the fact table in an empty star schema for gross agricultural production, gross crop and livestock production (rule 1.1).

Subsequently, in accordance with rule 1.2 we create a dimension into the star schema for each exogenous variable in our econometric model (the amount of subsidies, basic production funds in crop production, the amount of labour in crop production, climatic conditions, and number of livestock). Since the model (1) includes a time variable t, the time dimension is created.

In the last phase (rule 2.1) we form an association through the generated keys between the fact table and dimensions. Thus, for example, the equation $y_{2t} = \beta_{21} y_{1t} \gamma_{21} x_{1t} + \gamma_{25} x_{5t} + u_{2t}$ indicates that the level of subsidies and livestock have a relation with gross production of livestock production (i.e. with a measures of y_{2t} in the fact table).

In the application context the equation may be expressed in this form: $y_{1t} = 3{,}45 x_{1t} + 1{,}32 x_{2t} + 1{,}07 x_{3t} + 0{,}43 x_{4t} + 284{,}36$. So random components u_{1t}, u_{2t} and parameters β, γ are already expressed numerically. Therefore, random components u_{1t}, u_{2t} (or other variables that are not listed in the rules of the TEM method) are depicted in the schema.

5 Comparison of multidimensional schemas

The original TEM-CM method and above mentioned variant of transformation of econometric model are possible for the design of multidimensional schemas. Given the generally increasing complexity of analytical databases we should pay attention to the evaluation of their quality during their development.

In this part of our work we verify whether the results of measuring the quality and understandability of the multidimensional schema for the two mentioned variants of transformation presented are significant or not. This is important for determining which of these approaches we should select for transformation of the econometric model.

5.1 Quantitative comparison

The evaluation was performed for the first and second variant of transformation. We collectively illustrate the resulting schemas in Figure 1.

Variant 1 (TEM-CM) **Variant 2 (new TEM)**

Fig. 1 - TEM logical schema for EM with 3 equations, variant 1 and 2 (self-authored).

The evaluation results are shown in Table 1. The results show that the formed schema according to the first variant is structurally complicated, the total value of measurements is greater than of the second type of design.

Table 1: The result of quality assessment schema (self-authored).

Variant 1 (TEM-CM)		**Variant 2 (TEM)**	
Measures	Value	Measures	Value
NFT(Sc)	3	*NFT(Sc)*	1
NSDT(Sc)	2	*NSDT(Sc)*	0
Sum	**5**	**Sum**	**1**

Another possible indicator of the quality of the resulting schemas is the measurement of understandability. Evaluation is performed again for the first and second variant of transformation. The evaluation results are shown in Table 2.

Table 2: The result of understandability assessment schema (self-authored).

Variant 1 (TEM-CM)		**Variant 2 (TEM)**	
Measures	Value	Measures	Value
NFT(Sc)	3	*NFT(Sc)*	1
NDT(Sc)	6	*NDT(Sc)*	0
NFK(Sc)	12	*NFK(Sc)*	6
NMFT(Sc)	3	*NMFT(Sc)*	3
Sum	**24**	**Sum**	**10**

From the results of measurement (Table 2) we can see that the first variant in terms of understandability is significantly worse. Again, the second option is more suitable for the transformation of EM.

6 The creation of the prototype

To obtain an accurate preview of the future econometric OLAP solution we create a prototype of a multidimensional database through the TEM method. Creation of the prototype will allow us to verification of our method. For the creation of the prototype we follow the design of data warehouses by [22]. For the creation of the prototype we use conventional farming production function (according to [21]).

6.1 Conceptual design of the prototype

To create a conceptual schema we apply rules of the TEM method. Application of the TEM method for the production function leads to the identification of measurements and dimensions. The fact table for the whole schema is only one. The measurements identified by rule 1.1 is therefore a subset of the fact table.

Table 3: Description of the results of TEM (self-authored).

Rule 1.1	Measure	$y_{kt} = 234{,}25{L_{kt}}^{0{,}244}{WU_{kt}}^{0{,}616}{K_{kt}}^{0{,}100}$
Rule 1.2	Dimension	Land (L) Work (WU) Capital (K)
Rule 1.3	Dimension	Time

Result of the applied rules (Table 3) allows the creation of conceptual schema, which we further complete by a logical design.

6.2 Logical design of the prototype

In the logical design we apply the rule 2.1, which allocates to each identified dimension an appropriate association with the fact table. Since of the TEM method does not affect the creation of attributes relative to the semantics of variables in the model, it is advisable to add other possible attribute of dimensions to the schema after the use of the TEM method.

For the Land dimension we will include the attribute Acreage which will contain data about the hectare acreage of agricultural land. For the dimension of Work, we create an attribute Number, which will include the number of workers. In the Capital dimension we add the Size attribute expressed as a sum of tangible and intangible fixed assets in thousands of crowns.

To the time dimension we add the attributes that will allow econometric analysis in the long term. From an economic nature it is irrelevant to perform the analysis in the short term, since most of the factors in the equation are unchanged in the short term. It is also necessary to choose the type of data granularity.

The snapshot granularity is suitable for the econometric analysis. Data is entered into a database at the same time intervals (e.g. every quarter). Thus, the time dimension takes into account both the year and the quarter. In these intervals, it is possible to identify changes in various dimensions. The resulting logical schema is illustrated in Figure 1.

Fig. 1 - Logical schema of the prototype (self-authored).

6.3 Physical design of the prototype

After the creation of the logic schema our next step is to design the physical schema. The essence is in supplementing the logical model by physical characteristics that are typical for the OLAP technology and specific database system.

However, at this stage of acceptance of our solution the optimum specific setting of the proposed database solution is not important but the opportunity to examine the proposed logical model is. Therefore, we used MS PowerPivot for the physical design of the prototype which is sufficient to build our prototype.

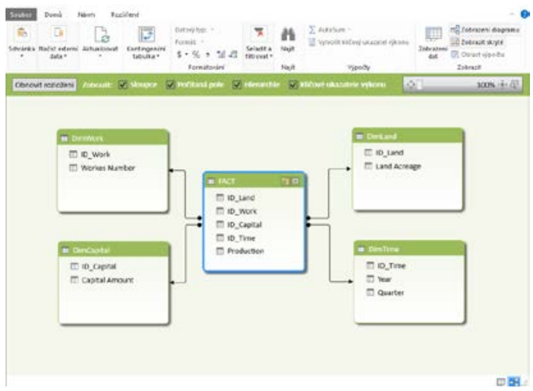

Fig. 2 - Diagram in PowerPivot for the prototype (self-authored).

First in PowerPivot we integrate data into the fact table and each dimension and create a relationships proposed by logical schema (Figure 2). Integrated data do not represent specific data of the agriculture company. Data are only theoretical and represent a medium-sized farm in the period of 2010 - 2012 with the average number of workers in the interval <3; 6>, capital (millions CZK) in the range of <2.5; 4> and land area (ha) in the interval <90;

120>. Integrated tables contain attributes designed in the logical schema. Other attributes, especially those that could add dimensions to individual hierarchy, are not included in the prototype.

The physical approach of the design of the prototype allows us to verify that the logical model proposed by the TEM method is feasible and has practical importance in the design of econometric based intelligence systems. Depending on the needs of econometric analyses it is particularly necessary to take into account the form of integrated data.

7 Conclusion

The fact that many companies will reach full production power and technical efficiency was the motivation for our research. In this paper we seek such approaches that can improve the development of econometric intelligent systems. These systems can easily help interpret to the company management the econometric analysis from which it is possible to obtain relevant knowledge about their economic performance.

In this paper we presented the TEM method. The subject of this method is the transformation of econometric models into multidimensional schemas. Input element is just econometric equations that are transformed into a conceptual and logical schema of analytic databases.

The proposed method provides system engineers methodological framework for designing structures of multidimensional databases. This approach based on client applications of the OLAP type enables us to gain analytical data and present it through dashboards in the form of PivotTables, graphs and other special outputs. Track information on total production costs or consumption developed in time for the whole company or its parts. The key benefit is the specific econometric ability to analyze factor - factor that can be implemented in a manner that allows the user to select different combinations of factors (for example, number of employees, price, quantity of input factors, the population in the region, etc.).

Due to the fact that similar research on econometric analysis through OLAP was not carried out, the results and benefits bring new insights into the development of Econometric Intelligent Systems.

In future research we will look for approaches for process design to make fully automated transformation econometric models to OLAP database.

Acknowledgements:
This work was conducted within the project DEPIES – Decision Processes in Intelligent Environments funded through the Czech Science Foundation, Czech Republic, grant no. 15-11724S.

References:

[1] Abelló, A., & Romero, O. (2009). On-line analytical processing. In Encyclopedia of Database Systems (pp. 1949-1954). Springer US.

[2] Abdullah, A. (2009). Analysis of mealybug incidence on the cotton crop using ADSS-OLAP (Online Analytical Processing) tool. Computers and Electronics in Agriculture, 69(1), 59-72.

[3] Assaf, T., & Dugan, J. B. (2007, January). Decision automation for predictive analysis models. In 2007 Annual Reliability and Maintainability Symposium (pp. 335-340). IEEE.

[4] Bimonte, S., Pradel, M., Boffety, D., Tailleur, A., André, G., Bzikha, R., & Chanet, J. P. (2013). A New sensor-based spatial OLAP architecture centered on an agricultural farm energy-use diagnosis tool. International Journal of Decision Support System Technology (IJDSST), 5(4), 1-20.

[5] Brandl, B., Keber, C., & Schuster, M. G. (2006). An automated econometric decision support system: forecasts for foreign exchange trades. Central European Journal of Operations Research, 14(4), 401-415.

[6] Bravo-Ureta, B. E., Solís, D., López, V. H. M., Maripani, J. F., Thiam, A., & Rivas, T. (2007). Technical efficiency in farming: a meta-regression analysis. Journal of productivity Analysis, 27(1), 57-72.

[7] Brown, J. E., Ethridge, D. E., Hudson, D., & Engels, C. (1995). An automated econometric approach for estimating and reporting daily cotton market prices. Journal of Agricultural and Applied Economics, 27(02), 409-422.

[8] Calero, C., Piattini, M., Pascual, C., Serrano, M. A., Piattini, M., Genero, M., ... & Ruiz, F. Towards DW Quality metrics. In Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW 2001) Interlaken, Switzerland (June 4, 2001).

[9] Cechura, L. (2014). Analysis of the technical and scale efficiency of farms operating in LFA. AGRIS on-line Papers in Economics and Informatics, 6(4), 33.

[10] Dolk, D. R., & Kridel, D. J. (1991). An active modeling system for econometric

analysis. Decision Support Systems, 7(4), 315-328.

[11] Felipe, J., & Adams, F. G. (2005). " a theory of production" the estimation of the cobb-douglas function: A retrospective view. Eastern Economic Journal, 31(3), 427-445.

[12] Fountas, S., Carli, G., Sørensen, C. G., Tsiropoulos, Z., Cavalaris, C., Vatsanidou, A., Liakos, B., Canavari, M., Wiebensohn, J. & Tisserye, B. (2015). Farm management information systems: Current situation and future perspectives. Computers and Electronics in Agriculture, 115, 40-50.

[13] Gupta, R., & Gosain, A. (2010, September). Analysis of Data warehouse quality metrics using LR. In International Conference on Advances in Information and Communication Technologies (pp. 384-388). Springer Berlin Heidelberg.

[14] Han, M., & Ju, C. (2008, August). Research and Application on OLAP-based Farm Products Examination Model. In Electronic Commerce and Security, 2008 International Symposium on (pp. 858-861). IEEE.

[15] Karmakar, S., Laguë, C., Agnew, J., & Landry, H. (2007). Integrated decision support system (DSS) for manure management: A review and perspective. Computers and Electronics in Agriculture, 57(2), 190-201.

[16] Kroupova, Z. (2010). Technická efektivnost ekologického zemědělství České republiky. Ekonomická revue, 2, 61-73.

[17] Nilakanta, S., Scheibe, K., & Rai, A. (2008). Dimensional issues in agricultural data warehouse designs. Computers and electronics in agriculture, 60(2), 263-278.

[18] Nowak, A., Kijek, T., & Domanska, K. (2015). Technical efficiency and its determinants in the European Union agriculture. Agric. Econ.(Zemedelská Ekonomika), 6, 275-283.

[19] Pedersen, T. B. (2009a) Cube. On-line analytical processing. In Encyclopedia of Database Systems (pp. 538-539). Springer US.

[20] Rai, A., Dubey, V., Chaturvedi, K. K., & Malhotra, P. K. (2008). Design and development of data mart for animal resources. Computers and Electronics in Agriculture, 64(2), 111-119.

[21] Recio, B., García-Mouton, E., Castellanos, M.T., Morató, M.C. & Ibáñez, J. (2010). An econometric system to assess the economic impact of water restriction policies in Spain. Spanish journal of agricultural research, (3), 526-537.

[22] Rizzi, S., Abelló, A., Lechtenbörger, J., & Trujillo, J. (2006, November). Research in data warehouse modeling and design: dead or alive?. In Proceedings of the 9th ACM international workshop on Data warehousing and OLAP (pp. 3-10). ACM.

[23] Serrano, M. A., Calero, C., Sahraoui, H. A., & Piattini, M. (2008). Empirical studies to assess the understandability of data warehouse schemas using structural metrics. Software Quality Journal, 16(1), 79-106.

[24] Schulze, C., Spilke, J., & Lehner, W. (2007). Data modeling for Precision Dairy Farming within the competitive field of operational and analytical tasks. Computers and Electronics in Agriculture, 59(1), 39-55.

[25] Tvrdoň, Jiří. 2006. Ekonometrie. Praha : Česká zemědělská univerzita v Praze, 2006. 80-213-0819-2.

[26] Tyrychtr, J., & Vasilenko, A. (2015). Transformation Econometric Model to Multidimensional Databases to Support the Analytical Systems in Agriculture. AGRIS on-line Papers in Economics and Informatics, 7(3), 71.

[27] Uyan, M., Cay, T., & Akcakaya, O. (2013). A Spatial Decision Support System design for land reallocation: A case study in Turkey. Computers and electronics in agriculture, 98, 8-16.

[28] Vassiliadis, P., & Sellis, T. (1999). A survey of logical models for OLAP databases. ACM Sigmod Record, 28(4), 64-69.

[29] Yu, L., Wang, S., & Lai, K. K. (2008). Forecasting China's foreign trade volume with a kernel-based hybrid econometric-AI ensemble learning approach. Journal of Systems Science and Complexity, 21(1), 1-19.