Data warehouse technology for agricultural policy data: a Greek case study

Article *in* International Journal of Sustainable Agricultural Management and Informatics · January 2016

DOI: 10.1504/IJSAMI.2016.082002

2 authors:

Michael Maliappis
Agricultural University of Athens
22 PUBLICATIONS **181** CITATIONS
SEE PROFILE

Dimitrios Kremmydas
European Commission
32 PUBLICATIONS **141** CITATIONS
SEE PROFILE

Data warehouse technology for agricultural policy data: a Greek case study

Michael T. Maliappis*

Laboratory of Informatics,
Department of Agricultural Economics and Rural Development,
Agricultural University of Athens,
75 Iera Odos, 11855 Athens, Greece
Email: michael@aua.gr
*Corresponding author

Dimitris Kremmydas

Laboratory of Agribusiness Management,
Department of Agricultural Economics and Rural Development,
Agricultural University of Athens,
75 Iera Odos, 11855 Athens, Greece
Email: Kremmydas@aua.gr

Abstract: Statistical data for agricultural policy analysis has certain unique features: a multitude of sources of very different nature, a variety of dimensional granularity and different end user requirements. The utilisation of data warehouse (DW) technology is valuable for tackling the above issues and successfully offering data to policy stakeholders and modellers. In this paper, we briefly introduce the DW technology, discuss the DW design issues in the context of policy related data and investigate the several difficulties identified on building and using a DW for monitoring crop responses to climate change for two Greek regions.

Keywords: agricultural data; data warehouse; online analytical processing; OLAP; agricultural policy.

Reference to this paper should be made as follows: Maliappis, M.T. and Kremmydas, D. (2016) 'Data warehouse technology for agricultural policy data: a Greek case study', *Int. J. Sustainable Agricultural Management and Informatics*, Vol. 2, Nos. 2/3/4, pp.243–262.

Biographical notes: Michael T. Maliappis is a researcher in the Scientific Area of Computer Science and Informatics at the Agricultural University of Athens, Greece. He holds a degree in Mathematics, Master of Science in Operational Research and Informatics and PhD in Knowledge Systems. His research interests include knowledge representation and management, knowledge base systems, system analysis and design of information systems, design and implementation of expert systems and ontology design and usage.

Dimitris Kremmydas has received his BSc from the Department of Agricultural Economics and Rural Development (Agricultural University of Athens) and from the Department of Informatics (Hellenic Open University). He also has received his MSc on Integrated Rural Development. He is currently a Teaching Assistant in the Agribusiness Laboratory and is doing his PhD on Agent-Based

Modelling for Evaluating Agricultural Policies. He has published a number of journal articles and several conference papers. His research interests include agricultural policy modelling, mathematical programming, decision support systems, parallel computing, agent-based modelling and the application of operations research methods in agribusiness management.

This paper is a revised and expanded version of a paper entitled 'An online analytical processing (OLAP) database for agricultural policy data: a Greek case study' presented at HAICTA 2015, 7th International Conference on Information and Communication Technologies in Agriculture, Food and Environment, Kavala, Greece, 17–20 September 2015.

1 Introduction

Worldwide, the agricultural sector is receiving a significant amount of state funding through various agricultural policy tools. In a recent report of the Organisation for Economic Co-operation and Development, the included countries (50 countries, accounting for the majority of global agricultural value added) provided an annual average of EUR 469 billion of support to their agricultural producers directly in the years 2013-15 (OECD, 2016). Thus, the efficient allocation of funding in order to accomplish the strategic goals of the policy makers is essential. Agricultural policy analysis is concerned with evaluating the instruments of providing subsidies to the agricultural sector, ex ante or ex post (Alston and James, 2002).

Although this evaluation is based on theoretical models, most often evidence is sought for empirical validation. In fact, as Runge (2006) notes, the agricultural economics subject itself arose in the late 19th century partly due to the fact that the US Department of Agriculture (USDA) had compiled rich datasets some decades earlier. Consequently, data is of prime importance for agricultural policy analysis, since it is utilised by policy makers to make qualitative judgments and by researchers to build quantitative models.

This agricultural policy related data bears certain special features:

1 *There exist many independent sources of information*, e.g., international or national statistical offices, diversified administration databases, field surveys or past data from universities, etc., none of which should be disregarded because agricultural data is actually a scarce resource. Those sources possibly store their data in different formats or/and different database schema definitions.

2 *The related data expands horizontally on various dimensions* since frequently agriculture policy makers pursue multiple goals. Those dimensions may be classified to:

   - Biophysical (weather or soil related, animal population, etc.). A usage example may be to investigate the effect of a policy to a region's biodiversity or soil erosion.
   - Technical/technological (input-output relationships, management practices, etc.). Technical relationships (i.e., what inputs are used for a certain crop in a specific area) are very important factors for farmers' production decisions and thus are directly relevant to agricultural policy.

- Technical/technological (input-output relationships, management practices, etc.). Technical relationships (i.e., what inputs are used for a certain crop in a specific area) are very important factors for farmers' production decisions and thus are directly relevant to agricultural policy.
- Economic (prices of inputs and outputs, production, income, etc.) containing data that is directly related to policies.
- Social (population per community, age pyramid, etc.) since often agricultural policies target at altering (e.g., shrinking/maintaining population, developing skills, etc.) rural societies. See Elizabeth and Mattison (2005) for a typical policy case study where all of the above dimensions are relevant.

3 *The temporal and spatial dimensions are relevant to their finest available detail*. The first is important to note because temporal dimension might not be always recorded, especially in the case of operational databases. Also, normally policy makers are interested on policy effect estimation to the finest administrative unit and the constraint for doing so is data availability.

4 *Dimensions are mostly of hierarchical kind*. For example, the spatial/administrative dimension includes the community at the lowest level and the country at the top; production type can be very specific (e.g., production of milk from goats) concluding to aggregated level (e.g., production from animals); time from daily to yearly; etc. This logical hierarchy is relevant, since it can facilitate the compilation of databases that hold information for different level of detail and can also be useful for presentation purposes to different stakeholders of the policy making process (e.g., a municipality officer may be interested on a more focused view of the data, in contrast with a ministry officer that is interested on an aggregated picture).

5 *Data is utilised by different kind of users, each with diverse needs*. For example, for a high level policy maker or an administration officer, it is sufficient to browse the data through a web interface or browse the results of a data mining procedure while for a modeller the data will ideally be directly imported to his/her model (e.g., by means of a web service).

The above specific features of agricultural policy related data designates data warehouse (DW) technology to be ideal for usage in agricultural policy monitoring and evaluation. DW can effectively facilitate collection of data from different sources explicitly maintaining temporal information; integrate and present multidimensional data; deal efficiently with hierarchical dimensions; and output data in different ways (Boulil et al., 2014; Rai et al., 2008). The application of DW in agricultural policy evaluation is not a straightforward process since DW technology is a set of processes rather than a ready-to-deliver product and there are specific design requirements that are discussed in the rest of the paper.

In the broad domain of agriculture, there are several cases where a DW was introduced to manage statistical data. One of the earliest appearances was that of the USDA's National Agricultural Statistics Service (Yost, 2000). Another attempt was that of the development of a central DW at Indian Agricultural Statistics Research Institute (IASRI) at New Delhi (Chaturvedi et al., 2008; Rai et al., 2007). Abdullah and Hussain (2006) describe an agriculture extension DW that monitors cotton pests in Pakistan. Van Broekhoven (2007) describes a DW system developed for presenting the Belgian

Farm Accountancy Data Network Data. Additionally, DW solutions are also applied to agricultural business problems. For instance, Schulze et al. (2007) use a DW in a dairy precision farming context, to collect data from different dairy enterprises and efficiently derive timeline measures for examining disease treatments.

In the rest of the paper, we briefly introduce the DW technology (Section 2.1), discuss design issues in the context of policy related data (Sections 2.2 and 2.3) and investigate the several difficulties identified on building and using a DW for monitoring crop responses to climate change for two Greek regions (Section 3). Finally, a summary of conclusions is drawn (Section 4).

2 DW technology for agricultural policy

2.1 A brief introduction to DW technology

As Ballard et al. (1998) note, a DW is not a product but rather a solution for transforming plain information to knowledge. More specifically, they define data warehousing as "the design and implementation of processes, tools, and facilities to manage and deliver complete, timely, accurate, and understandable information for decision making. It includes all the activities that make it possible for an organisation to create, manage and maintain a data warehouse or data mart".

The current DW process lifecycle includes a wide set of operations (Casters et al., 2010; Kimball and Ross, 2013) as depicted in Figure 1. The first step towards DW development is *the identification of data sources*. Usually a disparate (i.e., in respect to mean of storage, access protocols, logical organisation, data quality, etc.) set of sources is used and thus an intermediate procedure called 'extract, transform, load' (ETL) is required in order to prepare plain data and load it in the DW engine. Finally, the DW data is not directly accessible by end-users but accessed by means of *reports*, *data mining interfaces* and *online analytical processing* (*OLAP*) cubes. We provide some DW technology term definitions in order for not so familiar readers to be able to follow the rest of the paper.

Figure 1 DW lifecycle process

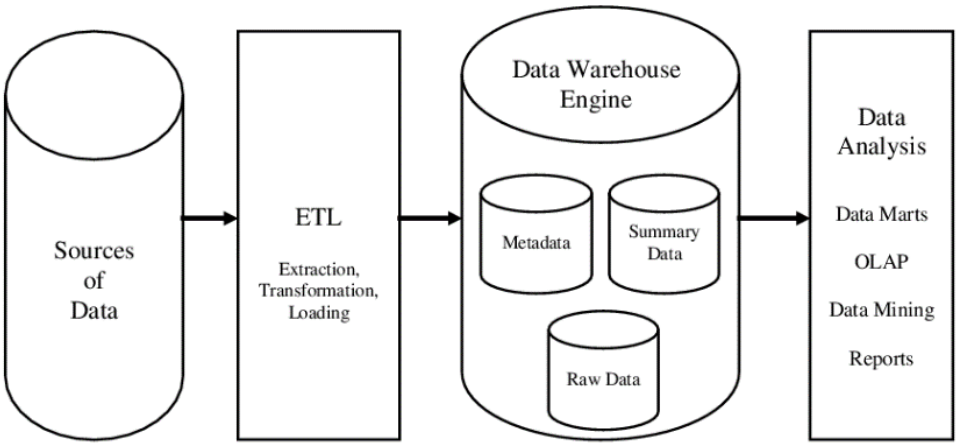
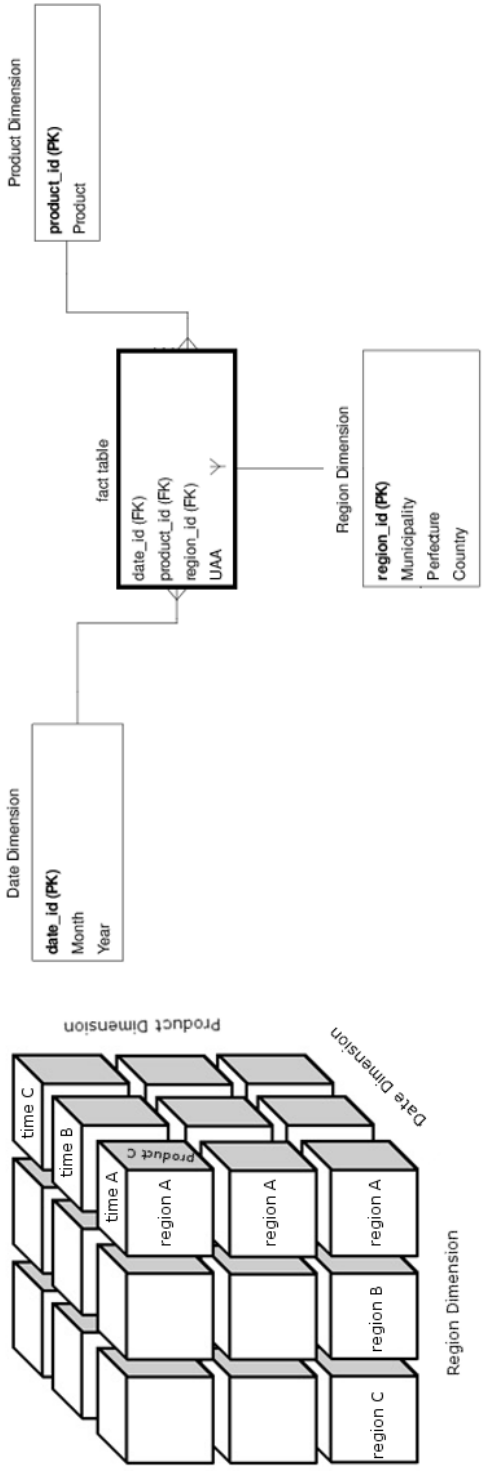


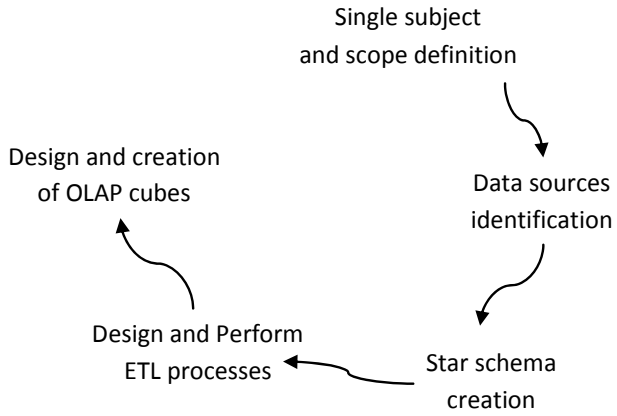
Figure 2 DW main notions visualised



The notions of *fact*, *dimension* and *measure* are central for designing a DW (Malinowski and Zimányi, 2008). Facts are collections of related data items, e.g., the utilised agricultural area (UAA) taken from a census. *Dimensions* are the structures that categorise facts, e.g., the product, the region or the time that the reported utilised area refers to. *Measures* are facts that are aggregated to dimension tuple, e.g., the UAA for a specific region and a specific year. This different dimension aggregation can be depicted as a cube (in the case of three dimensions) or a *hypercube* (more than three dimensions), where each cell of the cube is a measure. Star schema is a relational database schema where one or more central fact tables are linked to one or more dimension tables. DW development process can be abstracted to the mapping of the selected data source data schemas to the DW star schema. A *data mart* is a collection of facts, dimensions and measures that are subject specifics. A DW is actually a collection of data marts.

In Figure 2, we provide a visualisation of a *data mart* and the connection between *facts*, *dimensions*, *measures*, *cubes* and *star schemas*. The fact of the UAA is connected to three dimensions: Date (in month resolution) that UAA was measured, the product and the region that the measurement is referring to. There is a direct correspondence between the star schema and the data mart cube. Each cell of the cube is a measure (what is the UAA of product X in region Y at date Z).

Figure 3 DW design workflow



Source: Adapted from Ballard et al. (1998)

OLAP is the multidimensional view of data stored in a DW. *OLAP cubes* are view structures that correspond to DW cubes like in Figure 2. OLAP functionality includes basic navigation and browsing, i.e., view a *slice* of the data cube (one or more dimensions constant), *dice* (create a sub-cube), *drill down/up* (from aggregated to more detailed dimensions levels or reverse) and *roll-up* (summarises data along a dimension). Also, statistical analyses, time series creation and more complex modelling can be utilised.

Nevertheless, there is a significant literature on DW technology. For a swift introduction, we suggest the guides from Ballard et al. (1998) and Lane et al. (2005) and for going deeper into the subject the books of Ponniah (2011), Adamson (2010) and Kimball and Ross (2013).

Table 1 Sources of statistical information for Greek agriculture

Provider	*Data series name*	*Type*	*Starting year*	*Frequency*	*Geographical coverage*	*Finest geographical resolution*	*Data included*	*Data availability*
EL.STAT.[1]	Census of Agricultural and Livestock Holdings	Census	1961	Every 10 years	Whole of Greece	Municipal districts	Number of plant and animal agricultural holdings and their properties regarding their legal status, agricultural and tenure status, structural properties (type of crop/animal/activity), production methods.	1961, 1971, 1981, 1991 in printed form 2000, 2009 in electronic form
EL.STAT.	Annual Agricultural Statistical Survey	Survey	1961	Annual	Whole of Greece	Municipalities (as defined in the 'Kapodistrias' law)	Agricultural utilised land per type of crop, volume of agricultural (plant and animal) production, utilisation of agricultural machineries.	Online from 1961–2006
EL.STAT.	Farm Structure Survey	Survey	1966	1966, 1977, since 1983 every 2 years (but not 1991 and 2000), since 2010 every 3 years	Whole of Greece	Municipal districts	Number of plant and animal agricultural holdings and their properties regarding their legal status, agricultural and tenure status, structural properties (type of crops/animal/activity), production methods.	Online since 2003
EL.STAT.	Survey on crop production (including permanent cultivations and grape yards)	Survey		Grape yards: yearly survey, grains and other crops/basic survey every 10 years for grape yards/research every 5 years for permanent cultivations	Whole of Greece	Prefecture (NUTS-2)	Cultivating area per crop	Online since 2000
EL.STAT.	Agriculture input and output price index	Index	1967	Monthly	Whole of Greece	760 (output) and 783 (input) price-collection-points, from all Greece	Index of output prices (subsidies and transport costs are excluded) for plant and animal products (as classified in European Economic Accounts) index of input (products and services) prices.	Online since 2001

Note: [1]Hellenic Statistical Authority (EL.STAT).

Table 1 Sources of statistical information for Greek agriculture (continued)

Provider	*Data series name*	*Type*	*Starting year*	*Frequency*	*Geographical coverage*	*Finest geographical resolution*	*Data included*	*Data availability*
EL.STAT.	Agriculture production factors' index (cost index)	Index	1975	Yearly	Whole of Greece	Whole of Greece/155 points of price collection points	Index of production factor wage. It is comprised of three sub-indexes: labour (payment for one day), land (rent), and capital (loan interests and agricultural machinery rent).	Online since 2005
Greek Ministry of Agriculture	FADN/RICA	Survey	1985	Annual	Whole of Greece	Municipal District/~4,000 farms	Accountancy data	Fine detailed data is not freely distributed. Aggregated data is publicly available.
Greek Payment Authority of Common Agricultural Policy Aid Schemes (OPEKEPE)	Registry of farm subsidies	Registry	2013	Yearly	Whole of Greece	Plots	Crop type per plot basis	Not publicly available
National Observatory of Athens (Greece)	Public Database of Meteorological Measurements	Real Data	2006	Daily	Network of 347 scientific meteorological stations	Spatial point	Mean temperature, min-max temperatures, rain, mean wind speed, dominant wind direction.	Online as fixed width text files
European Environment Agency	Corine land cover	Spatial	1990	1990, 2000, 2006	Whole of Greece	25 hectares	44 classes of land use	Online
European Soil Data Centre (ESDAC)	European Soil Database	Spatial	2001	-	Whole of EU-27	1:1,000,000	Soil-related data	Online
EUROSTAT	Land Use and Coverage Area Frame Survey (LUCAS)	Spatial	2006	3-years	Whole of EU-27	270,000 points in EU-27	Land cover, land use and environmental parameters associated with the individual points surveyed.	Online
EUROSTAT	TRADE Database (COMEXT)	Detailed data	1976	1976–1987 is annual, since 1988 is monthly	Whole of EU-27	Intra is from direct collection of information from trade operators/extra is from custom declarations	Value and quantity of goods traded between EU member states (intra-EU trade) and between member states and non-EU countries (extra-EU trade).	Since 2004 are free of charge http://ec.europa.eu/eurostat/web/international-trade/data/database

Note: [1]Hellenic Statistical Authority (EL.STAT).

2.2 Agricultural policy DW design issues

As Jukic (2006) notes, in DW design there are two main schools of thought: *a bottom-up approach* (Kimball and Ross, 2013), where subject-specific data marts are independently created. These are eventually integrated in a common 'dimension bus' forming the DW; a top-down approach (Inmon, 2002), in which the DW (i.e., the collection of individual data marts) is built after the normalised enterprise data model has been set. To further clarify, we will provide the agricultural policy relevant example, where we want to monitor/evaluate the agricultural policy of an EU country.

In the bottom-up approach, we would immediately proceed to creating a data mart for a specific policy measure. Let us assume that due to data availability we choose to create a data mart for direct payments to farmers, using data from the national payment authority. The star schema contains a fact table about the amount of payments, while the included dimensions are: farm production orientation; farm's region (in prefecture resolution), time of payment (in month resolution). After some time, a request to monitor agro-environmental measures is coming, so we create a new data mart with the same procedure, catering for the alignment of the common dimensions, if possible.

For the case of the top down approach, the agricultural policy is fully analysed and all desired reports and OLAP functionality are determined from the beginning. The relevant sources shall be established (e.g., see Table 1 for a list of Greek agricultural sources) and all relevant dimensions be included sketched in a *three-normal-form* dimension star schema. Then, ETL procedures are applied, data marts are populated and OLAP functionality is provided to end users.

What is the most appropriate approach for building a DW for agricultural policy monitoring and evaluation? As discussed in the introduction, agricultural policy related data is connected with many distinguished dimensions and is derived from many independent sources of information. Thus, in the top-down approach, all costs are incurred at the beginning of the project. Furthermore, since many independent administrative departments and stakeholders are involved, project requirements may change, unpredicted problems may arise and thus the final outcome is uncertain. Instead, a bottom up approach will produce usable results in short time, demonstrating the virtues of using DW technology to the stakeholders increasing the motivation for adoption from other departments too. In resume, the complexity of the stakeholder's structure that is most often found in agricultural policy context can be effectively managed with the bottom-up approach.

2.3 Design and implementation of policy related data marts

Regarding the design and implementation of a single data mart, using the bottom up dimensional model, we adapt the DW design lifecycle proposed by Ballard et al. (1998), as shown in Figure 2, and discuss the various steps in the agricultural policy context.

1 *Subject and scope definition*: In the beginning, a single subject shall be defined. The question that is to be answered here is 'what and why do I want to analyse'. Example subjects are: 'monitoring crop yields in response to climate change' (see following section); 'the effect of a specific policy measure on the biodiversity of arable crop fields'.

2 *Data sources identification*: After there is a clear idea on the what-why questions, it is easier to identify the relevant data sources. Regarding agricultural policy related data, apart from the traditional sources (e.g., statistical offices, etc.), there are some emerged data sources worth considering: geo-referenced agricultural and environmental data sources, like meteorological remote sensing systems, satellite images; computer applications where farmers can record their practices (Pinet and Schneider, 2010); data from agricultural institutes and projects (Janssen et al., 2012). In general, there is a variety of data sources that can be proved useful, for example see Table 2 for the case of Greek agriculture. Also, the use of web services can provide easy access to that data.

3 *Star schema creation*: Regarding agricultural policy context, certain dimensions are expected to appear often: Type of activity, temporal and administrative unit classification. Their organisation in a hierarchical way will facilitate the data analysis phase and the related operations (drill up/down, slice, etc.). Other dimensions are also expected to be present (farm size, various categorisation of farming types, etc.) and should be catered accordingly. Also, the logical conformation of the dimensions takes place in this stage and a relevant discussion on dealing with such issues is made in Nilakanta et al. (2008).

   A good systematic method for the conformation of a dimension that is common in different data sources is depicted in Figure 6. A directed graph is constructed where each node represents a resolution level of the dimension. There is also a directed hierarchical positioning of the various dimension levels (e.g., municipalities are connected to regions, regions to countries, countries to continents, etc.). If there are dimensions levels of different data sources that are exactly the same, then they are both written within the same node in a way that the information of the source is maintained (e.g., {LAU-1} and [Municipalities], where {} notes that the level name is found in Eurostat data source and [] in Greek statistical office). Any acyclical subgraph that starts from the lowest level node and ends to the highest can be a hierarchical conformed dimension for two or more data sources, as long as they are present in at least one node.

   To clarify, in Figure 4, there are three such subgraphs: 1, 2, 3, 5, 6, 7, 9, 10 which connects Eurostat and Greek Statistical Office hierarchy with [Prefectures] to be a missing level: 1, 2, 3, 4, 6, 7, 9, 10 which does the same with {NUTS-3} be missing; and 1, 2, 3, 5, 6, 8, 10 which connects the above data sources and FADN source. A researcher will select the dimension merge that is more suitable for his needs (i.e., missing dimension is not important) and there is also the option of attaching more than one merged dimension to the data.

4 *Design and application of ETL processes*: Due to the variety of data sources, ETL design is expected to take a significant proportion of the project's workload. If the data was originally produced for operational use [an example can be found in Schulze et al. (2007)], it is very possible that the time definition will be missing (i.e., only updated data will be present) and shall be explicitly inserted in this step. Also, the application of the ETL process may reveal weaknesses on the data quality of certain organisational units (e.g., maintained in plain excel files, no remote retrieval methods, etc.) and the provided feedback would improve the overall IT infrastructure of the policy structure.

5 *Design and creation of OLAP cubes*: Due to the existence of different users with different needs, the access to the OLAP cube, ideally should be provided through a web interface with the ability to export in various formats and also through automated retrieval offering web service access.

Figure 4 A graph facilitating the conformation of hierarchical dimensions connected to different data sources (see online version for colours)

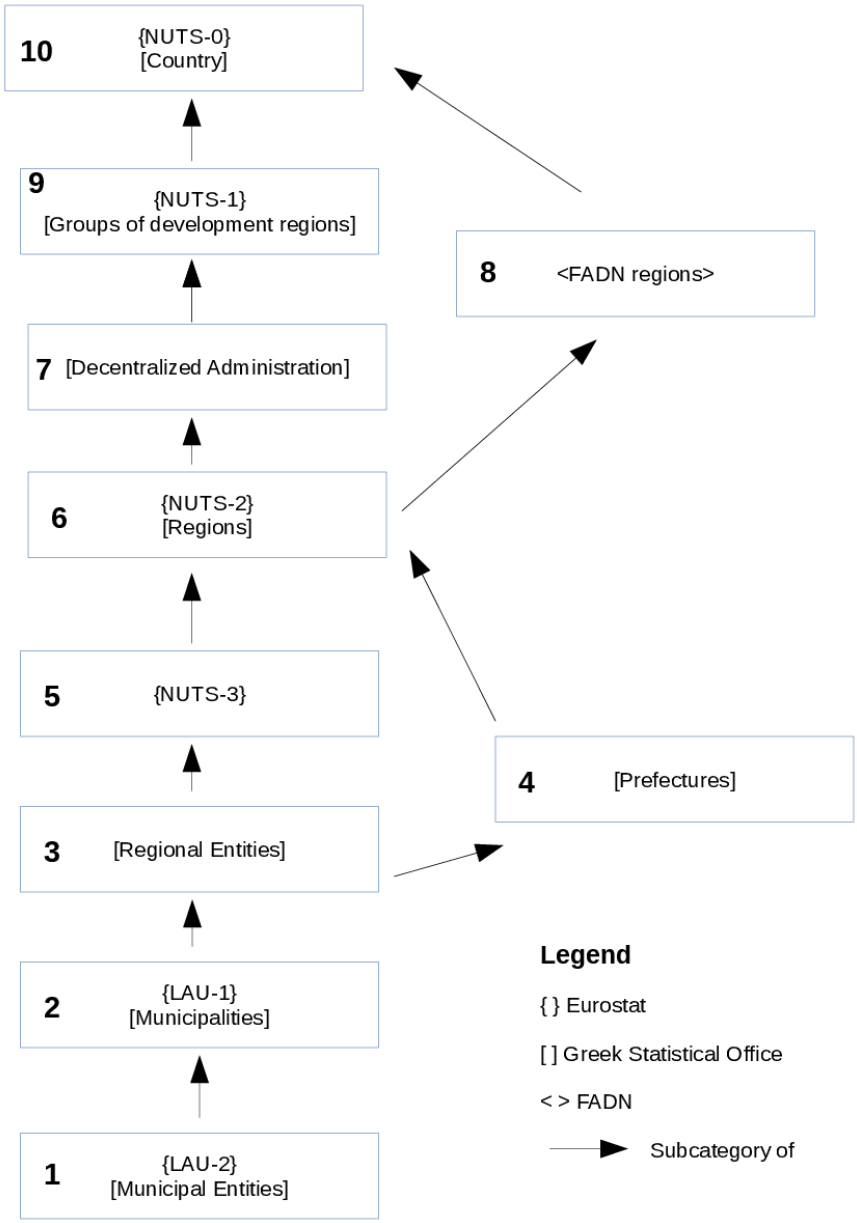


Table 2 List of case study dimensions and measures

Annual Agricultural Statistical Survey (Greek Statistical Office)	*Meteorological data (National Observatory of Athens)*
Dimensions:	Dimensions:
1 Time (in year resolution)	1 Time (in day resolution)
2 Administrative unit (in municipal entity resolution. 2000–2010 in 'kapodistrias coding scheme', 2011–2012 in 'kallikratis coding scheme'	2 Spatial (point in space)
	Measures:
3 Product code (as defined by Greek Statistical Office)	1 High Temperature (in C)
Measures:	2 Low Temperature (in C)
1 Area (in 0.1 ha)	3 Average Temperature (in C)
2 Production volume (in kgs)	4 Rain (in mm)

Figure 5 Star schema for the combination of meteorological and production data (see online version for colours)

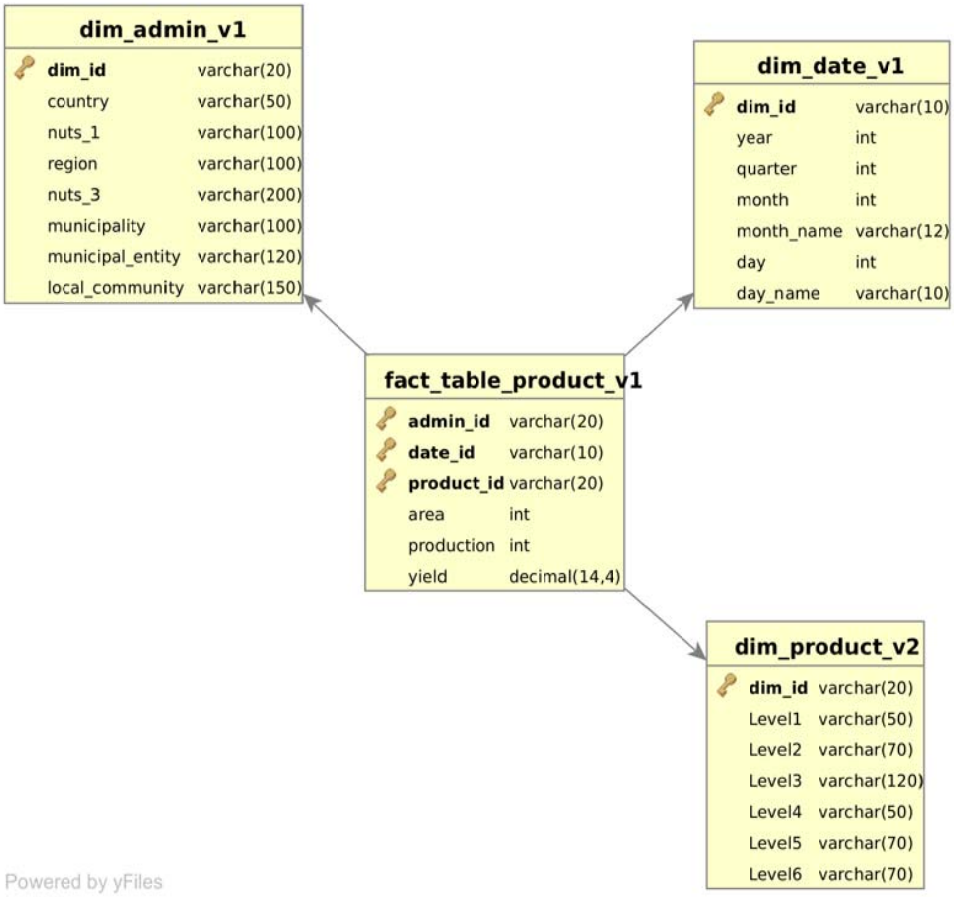
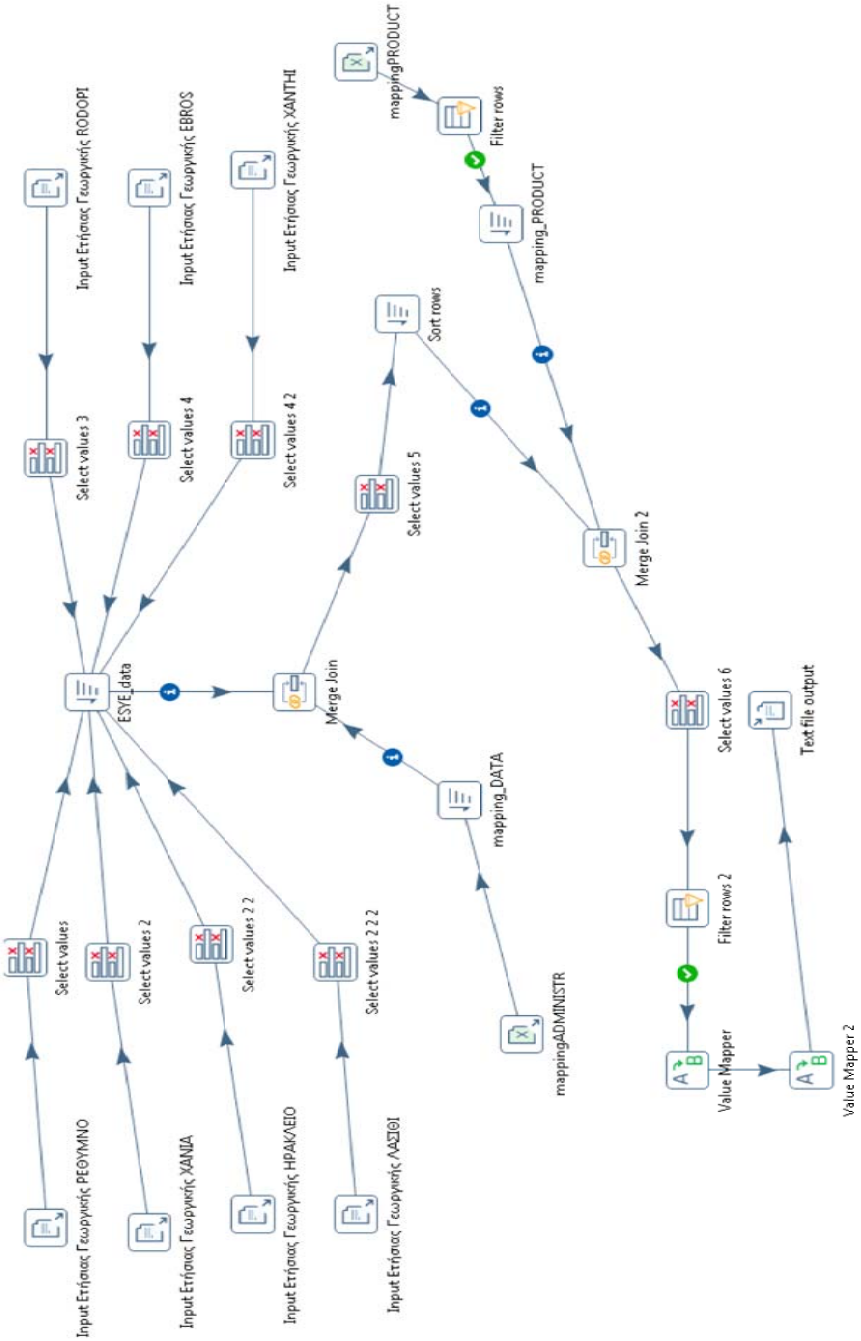


Figure 6 Kettle ETL procedure example (see online version for colours)



3 Case study: a data mart for monitoring yields and climate data

In order to demonstrate the power of using DW technology, we apply the previously discussed design workflow in a climate change case study. Our implementation uses free licensed tools: MySQL (https://www.mysql.com/) as DW storage database, Kettle (http://community.pentaho.com/projects/data-integration/) to facilitate collection, transformation and loading of data and Mondrian (http://community.pentaho.com/projects/mondrian/) to create the OLAP cube and apply data analysis and SpagoBI (http://www.spagobi.org) server to enable the execution of the OLAP cube.

We discuss in detail the implementation of the already presented design workflow:

Step 1 *Definition of subject and scope*: We want to monitor crop yields, weather conditions and their relation in order for policy makers to anticipate any climate change effects. Due to budget and time constraints, we select to monitor and evaluate two regions: Thrace (8,578 km^2) which is the northernmost Greek region and Crete (8,303 km^2) which is the southernmost one. Both areas cover about 20% of the national UAA. We also decided to focus on certain crops: cereals, cotton, tobacco and olives. Our selection diversifies the geographical and product scope so that the results are representative enough.

Step 2 *Data sources identification*: A major criterion for selecting a data source was, apart from being relevant to our subject definition and scope, to be publicly accessible. See Table 2 for a list of candidate data sources. We selected the 'Annual Agricultural Statistical Survey' since it provides annual data in fine grained administrative resolution (after our special request) for production volume and crop areas and thus detailed crop yields can be derived. We also use meteorological data (rain height, temperatures, etc.) recorded from the National Observatory of Athens that maintains a network of 347 scientific meteorological stations (http://meteosearch.meteo.gr/) all over Greece, since it is the most complete source of this kind of information.

Step 3 *Star schema creation*: In this step, it is crucial to identify and deal with any data peculiarities (e.g., how missing values are treated, is there any implicit dimension hierarchy, etc.), either through studying metadata or contacting the authority that supplied them. A list of the relevant dimensions and measures of the data sources will facilitate the creation of the star schema and the application of the previously described methodology will also be helpful. In our case, this list is provided in Table 2. Considering missing values of certain dimensional data, the policy was to ignore the whole row of data. In the case of missing fact data adopted a null value which resulted in exclusion by aggregation calculations.

Dimension conformation was not so difficult for time and administrative dimensions. Regarding time dimension, there is a clear hierarchical relationship (day, month, and year). As far as spatial (in meteorological data) and administrative unit (in agricultural statistics) dimensions, again a specific point in space can be clearly attributed to a municipal entity. On the other hand, the dimension hierarchy shall be carefully crafted because any future data mart development will overall be more efficient if it is based on the existing dimensions. Thus, before concluding, we investigated other potential future data

sources (see Table 1). This is more evident in the production activity dimension, where there are at least three different nomenclatures (Eurostat NACE-2, Eurostat LUCAS, and Greek Statistical Office). We based our hierarchy to the Greek Statistical Office but inserted latent levels so that in the future other nomenclatures can be merged where possible.

Finally, three star schemas were created: standalone production; standalone meteorological data; and another one for combining them. For the latter (Figure 6), we faced a certain challenges and thus provide more details: Yearly area, production volume and yield can be directly derived from production data. Mean and low temperature, rain mm per day is also directly available from weather data. The challenge was that the time dimension granularity was incompatible between the two data sources (weather on daily and production on yearly basis). An OLAP cube can successfully deal with this by aggregating (averaging, summing or giving the minimum) the finer detailed data (weather) to the least detailed one (production), i.e., present meteorological aggregated data in year resolution. But due to the impact of within year weather conditions to the overall behaviour of crops (e.g., a very rainy summer could dramatically decrease/increase yields even if the year average was close to normal), this would result in a significant loss of information. Thus, 36 additional measures were calculated: one for each month of a year and for each of the direct measures (mean and low temperature, rain mm), i.e., 12 months times three measures. To clarify even further a subset of those 36 measures is: mean temperature of January; mean temperature of February; …; mean temperature of December; lowest temperature of January; …; rain mm of December. Consequently, in the combined star schema a total of 39 facts are included.

Step 4 *Design and application of ETL processes*: Regarding the annual agricultural statistical survey, the problems we had to deal with was the relatively poor data connection interface (data was provided in excel files) and the fact that administrative coding schemes between 2000 to 2010 and 2011 to 2012 periods were different. Regarding meteorological data connectivity was also an issue, as they were provided through plain text files with fixed format but inconsistent across years and recording stations (data on some text files was starting on 11th row while for others on 12th and a relevant problem for data column start). Kettle was a valuable tool and handled efficiently the whole process, although meteorological files were more convenient to be downloaded through an http client and pre-processed with AWK scripts. As can be seen in Figure 5, the Kettle transformation takes as input several Excel files with the raw data, combines them with other Excel files containing dimensional data from the previous step, and after numerous transformation steps gives as a result the final dataset which can be uploaded into a database table.

After loading data we observed that aggregated production volume for certain crops was absurd, revealing a false step in transforming the original data. In any case, a last validation step is required so as to ensure that loaded data are error-free. Comparing official aggregated data with aggregation from the loaded data is an easy way of validating the ETL process. Overall, ETL was designed

so as any future data additions (e.g., additional regions or years) to be conveniently imported in the DW engine.

Step 5 *Design and creation of OLAP cubes*: The final step is the creation of the OLAP cube providing further data analysis capabilities. For each star schema, an OLAP cube has been created using the Mondrian server (a relational OLAP engine) connected to a MySQL database using the JDBC protocol. The OLAP cube contains the necessary metadata so that users can efficiently navigate through the different dimensions and aggregate data at different granularity levels.

The execution of the OLAP cubes has been carried out via the SpagoBI analytical server. For each cube one schema instance has been created containing representative dimensions with characteristic granularities and some of the measures. The analytical server is flexible enough to provide dimensions in any combination and granularity with the available measures. The user can easily select the required granularity of any dimension and the measures related to his/her own view of the data and apply to the cube the analytical capabilities described in section 2.1 (slice, dice, etc.). Moreover, any user, having the appropriate knowledge of the MDX query language, is able to create new calculated measures or to apply specific filters to the data. Furthermore, SpagoBI provides the tools to represent the data graphically or export specific instances of analysis. An architectural overview of the OLAP creation components is provided in Figure 7.

Within the analytical server, there is a tradeoff between the dimensions granularity and the speed of analysis. For each new combination selected, the data should be aggregated and represented after scanning the whole dataset. When the fact table contains huge number of rows, then the query, to aggregate data at finer dimension levels, takes longer time, because of the need to scan full dataset. The SpagoBI analytical server is able to store these aggregations in a memory cache, to facilitate subsequent analyses. These data are useful during one analysis session and for this reason the cache cleared after any restart of the server. The performance issue, with huge fact tables, can be solved by building aggregate tables, which contain pre-calculated summary data. The build of aggregate tables should be a part of the ETL process that populate/refresh the DW.

Figure 8 provides an instance analysis report of several meteorological measures aggregated at the year level of the time dimension. The report is coming from the analytical server and offers a high level view, comparing the measures between two specific regions of Greece.

Figure 7 Architectural overview of the OLAP implementation components

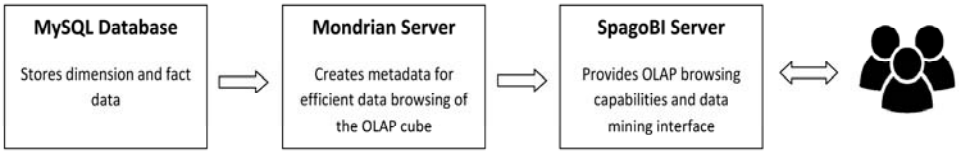
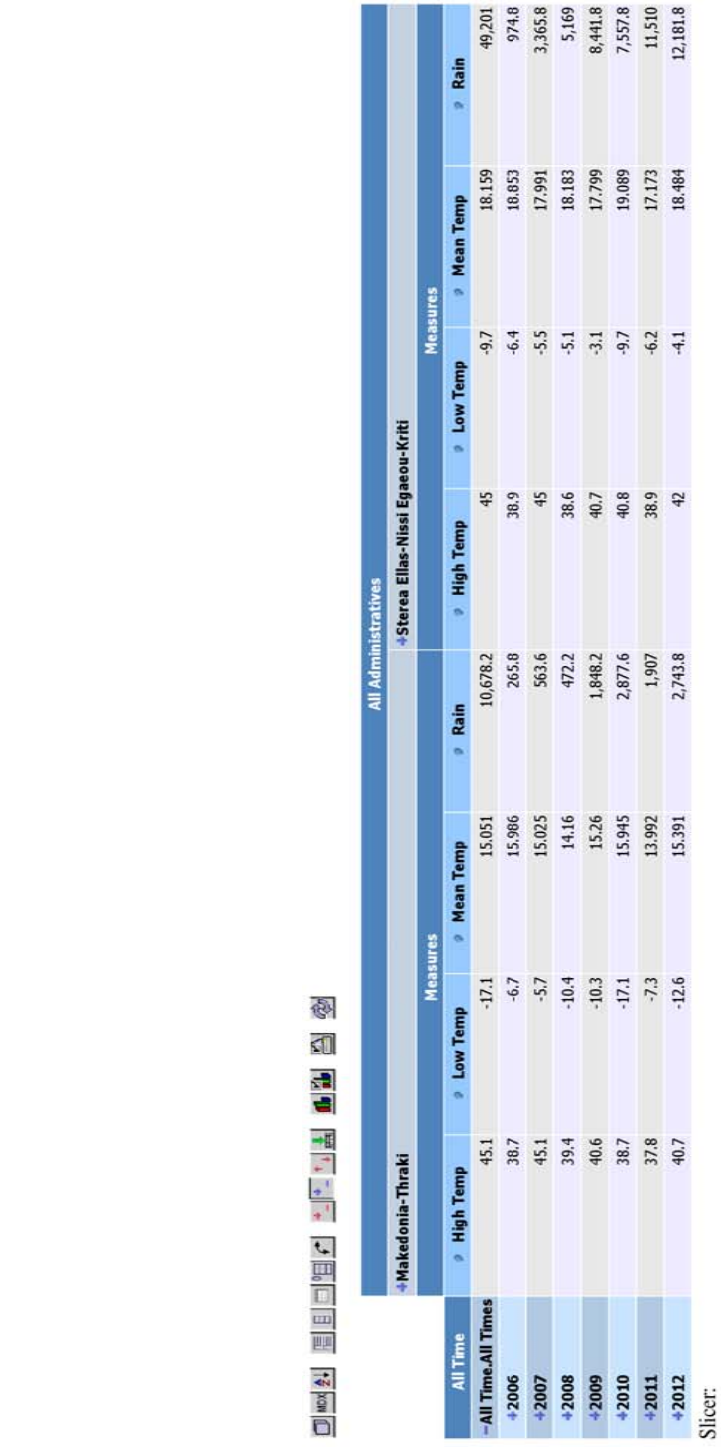


Figure 8 An instance analysis report of meteorological data OLAP cube (see online version for colours)



All Time	All Administratives									
	Makedonia-Thraki				Sterea Ellas-Nissi Egaeou-Kriti					
	Measures				Measures					
All Time	High Temp	Low Temp	Mean Temp	Rain	High Temp	Low Temp	Mean Temp	Rain		
All Time.All Times	45.1	-17.1	15.051	10,678.2	45	-9.7	18.159	49,201		
2006	38.7	-6.7	15.986	265.8	38.9	-6.4	18.853	974.8		
2007	45.1	-5.7	15.025	563.6	45	-5.5	17.991	3,365.8		
2008	39.4	-10.4	14.16	472.2	38.6	-5.1	18.183	5,169		
2009	40.6	-10.3	15.26	1,848.2	40.7	-3.1	17.799	8,441.8		
2010	38.7	-17.1	15.945	2,877.6	40.8	-9.7	19.089	7,557.8		
2011	37.8	-7.3	13.992	1,907	38.9	-6.2	17.173	11,510		
2012	40.7	-12.6	15.391	2,743.8	42	-4.1	18.484	12,181.8		

Slicer:

4 Conclusions

The utilisation of DW technology is valuable for tackling certain agricultural related data characteristics: a multitude of sources of very different nature and of independent origin; dimensions are mostly hierarchical; temporal and spatial aspects are relevant; fine granularity of data is useful; different end user requirements.

Regarding the design of policy related DW, we argue that the bottom-up approach is far more convenient compared to the top-down. In the top-down approach, all DW development costs are incurred at the beginning of the project while the bottom up approach will produce usable results in short time, increasing the motivation for adoption from policy stakeholders. Thus for implementing a policy related DW single subject data marts can be incrementally setup and implemented.

For creating a data mart, we propose the following steps: Define a single subject to investigate based on the question 'what do I want to analyse and why'; identify all relevant data sources; design and implement a star schema that will cater for connecting the facts and their dimensions dealing with possible hierarchy discrepancies using the proposed technique; design and implement the necessary ETL procedures that will fetch data from the selected sources, transform them accordingly and load them into the DW engine; design and create the necessary OLAP cubes for browsing the data.

In order to demonstrate the power of using DW technology we presented a climate change case study, where we pursue an answer to the question 'is the yield of certain crops affected by any climate change effects'. Certain important conclusions can be drawn:

- The use of open source tools for providing a whole data mart solution was adequate in terms of configuration, performance and user experience efficiency.
- The use web services by data providers can facilitate the retrieval of the data. Otherwise an overhead data manipulation cost shall be expected.
- Data validation shall always be part of the ETL process especially if fine detailed granularity data are handled.
- Dimension hierarchy shall be crafted carefully, catering for any future data additions. Thus a good strategy is to review possible future data sources although not used in the current data mart creation.
- There is a trade-off between performance and dimension granularity and we propose a certain solution for dealing with it

Conclusively, in this paper, we have successfully crafted a data mart for an agricultural policy related subject (monitoring relation of yields to climatic conditions), although various shortcomings were encountered (data quality and validation, performance, etc.). As far as the future work is concerned, the need for adding more policy related data marts will arise and any additional problems shall be confronted. A consolidation of data from farm structural surveys, FADN micro-data will be very useful to agricultural policy modellers. Also, the provided OLAP interface shall cater for automated data retrieval through web services. Finally, publicly available spatial data calls for a better integration with quantitative data in the DW context.

References

Abdullah, A. and Hussain, A. (2006) 'Data mining a new pilot agriculture extension data', *Warehouse Journal of Research and Practice in Information Technology*, Vol. 38, No. 3, pp.229–249.

Adamson, C. (2010) *Star Schema. The Complete Reference*, McGraw Hill Professional.

Alston, J.M. and James, J.S. (2002) 'The incidence of agricultural policy', in *Handbook of Agricultural Economics*, Vol. 2, Part B, pp.1689–1749, Elsevier, ISSN: 1574-0072, ISBN: 9780444510792.

Ballard, C., Herreman, D., Schau, D., Bell, R., Kim, E. and Valencic, A. (1998) *Data Modeling Techniques for Data Warehousing*, IBM Corporation International Technical Support Organization.

Boulil, K., Le Ber, F., Bimont, S., Grac, C. and Cernesson, F. (2014) 'Multidimensional modeling and analysis of large and complex watercourse data: an OLAP-based solution', *Ecological Informatics*, Vol. 24, pp.90–106.

Casters, M., Bouman, R. and van Dongen, J. (2010) *Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration*, Wiley Publishing, Inc., Indianapolis, Indiana, USA.

Chaturvedi, K.K., Rai, A., Dubey, V.K. and Malhotra, P.K. (2008) 'On-line analytical processing in agriculture using multidimensional cubes', *Journal of the Indian Society of Agricultural Statistics*, Vol. 62, No. 1, pp.56–64.

Elizabeth, H.A. and Mattison, K.N. (2005) 'Bridging the gaps between agricultural policy, land-use and biodiversity', *Trends in Ecology and Evolution*, Vol. 20, No. 11, pp.610–616, ISSN: 0169-5347.

Inmon, W. (2002) *Building the Data Warehouse*, 3rd ed., Wiley Publishing, Inc., Indianapolis, Indiana, USA.

Janssen, S., Kraalingen, D., Van Boogaard, H., De Wit, A., Franke, J., Porter, C. and Athanasiadis, I.N. (2012) 'A generic data schema for crop experiment data in food security research', in *International Environmental Modelling and Software Society (iEMSs) 2012 International Congress on Environmental Modelling and Software*.

Jukic, N. (2006) 'Modeling strategies and alternatives for data warehousing projects', *Communications of the ACM*, Vol. 49, No. 4, pp.83–88.

Kimball, R. and Ross, M. (2013) *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 3rd ed., John Wiley & Sons, Inc., Indianapolis, Indiana, USA.

Lane, P., Schupmann, V. and Stuart, I. (2005) *Oracle Database Data Warehousing Guide*, 10g Release, Vol. 2, No. 10.2.

Malinowski, E. and Zimányi, E. (2008) *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications*, Springer eBooks, Berlin, Germany.

Nilakanta, S., Scheibe, K. and Rai, A. (2008) 'Dimensional issues in agricultural data warehouse designs', *Computer Electronics in Agriculture*, Vol. 60, No. 2, pp.263–278, DOI: 10.1016/j.compag.2007.09.009.

OECD (2016) *Agricultural Policy Monitoring and Evaluation 2016*, OECD Publishing, Paris, DOI: http://dx.doi.org/10.1787/agr_pol-2016-en.

Pinet, F. and Schneider, M. (2010) 'Precise design of environmental data warehouses', *Operational Research*, Vol. 10, No. 3, pp.349–369.

Ponniah, P. (2011) *Data Warehousing Fundamentals for IT Professionals*, John Wiley & Sons, Inc., Indianapolis, Indiana, USA.

Rai, A., Dubey, V., Chaturvedi, K.K. and Malhotra, P.K. (2008) 'Design and development of data mart for animal resources', *Computer Electronics in Agriculture*, Vol. 64, pp.111–119.

Rai, A., Malhotra, P.K., Sharma, S.D. and Chaturvedi, K.K. (2007) 'Data warehousing for agricultural research – an integrated approach for decision making', *Indian Journal of Agricultural Statistics*, Vol. 61, No. 2, pp.265–274.

Runge, C.F. (2006) *Agricultural Economics: A Brief Intellectual History*, No. 13649, University of Minnesota, Department of Applied Economics, Minneapolis, Minnesota, USA.

Schulze, C., Spilke, J. and Lehner, W. (2007) 'Data modeling for precision dairy farming within the competitive field of operational and analytical tasks', *Computers and Electronics in Agriculture*, Vol. 59, Nos. 1–2, pp.39–55.

Van Broekhoven, E. (2007) 'A new farm accountancy data network for Flanders (Belgium)', in Poppe, K.J. et al. (Eds.): *PACIOLI 15 Workshop Presentations: Integration of Farm Accounting in Research and Statistics*.

Yost, M. (2000) 'Data warehousing and decision support at the National Agricultural Statistics Service', *Social Sciences Computer Review*, Vol. 18, No. 4, pp.434–441.