JRC Scientific and Technical Reports

Potentials of a Harmonised Database for Agricultural Market Modelling

David Verhoog, Michael Heiden, Petra Salamon, Wietse Dol and Frans Godeschalk (authors)
Stephan Hubertus Gay, Marc Müller and Federica Santuccio (editors)

EUR 43298 EN - 2008

JRC
EUROPEAN COMMISSION

ipts
Institute for Prospective Technological Studies

The mission of the IPTS is to provide customer-driven support to the EU policy-making process by researching science-based responses to policy challenges that have both a socio-economic and a scientific or technological dimension.

European Commission
Joint Research Centre
Institute for Prospective Technological Studies

Contact information
Address: Edificio Expo. c/ Inca Garcilaso, s/n. E-41092 Seville (Spain)
E-mail: jrc-ipts-secretariat@ec.europa.eu
Tel.: +34 954488318
Fax: +34 954488300

http://ipts.jrc.ec.europa.eu
http://www.jrc.ec.europa.eu

JRC 43298

EUR 23417 EN
ISBN: 978-92-79-09459-0
ISSN 1018-5593
DOI 10.2791/33791

Luxembourg: Office for Official Publications of the European Communities

Printed in Spain

Table of Contents

List of Abbreviations:

AAFC	Agriculture and Agri-Food Canada
AGLINK	a recursive-dynamic multi-region multi-commodity partial equilibrium model of regional and world markets for temperate-zone agricultural products
AGMEMOD	Agricultural Member State Modelling
AgrIS	Agricultural Information System (Eurostat)
AO	Additive Outliers
ARIMA	Autoregressive Integrated Moving Average
BACI	Base pour l'Analyse du Commerce International (CEPII)
BEC	Broad Economic Categories
BIS	Bank for International Settlements
BMELV	Federal Ministry of Food, Agriculture and Consumer Protection
CAP	Common Agricultural Policy
CAPRI	Common Agricultural Policy Regional Impact Analysis
CAPSIM	Common Agricultural Policy Simulation Model
CARD	Centre for Agricultural and Rural Development at Iowa State University
CCLS	Country-Commodity Linked System (USDA)
CE	Cross Entropy
CEPII	Centre d'Etudes Prospectives et d'Informations Internationales
CGE	Computable general equilibrium
cif	cost, insurance, freight
CN	Combined Nomenclature
CNFAP	Centre for National Food and Agricultural Policy at the University of Missouri-Columbia
COCO	Complete and Consistent Database
COMTRADE	Commodity Trade database (United Nations)
COSIMO	Commodity Simulation Model (FAO)
CPA	Statistical Classification of Products by Activity in the European Economic Community (Eurostat)
CPC	Central Product Classification
DBMS	Database Management System
DG AGRI	Directorate-General for Agriculture and Rural Development
DREAM	Dynamic Research Evaluation for Management (IFPRI)
EAA	Economic Accounts for Agriculture (Eurostat)
ECB	European Central Bank
ERS	Economic Research Service (USDA)
ESC	Economic and Social Development Department (FAO)
ESIM	European Simulation Model
EU	European Union
EuroCARE	European Centre for Agricultural Regional and Environmental Policy Research
Eurostat	Statistical Office of the European Communities
FADN	Farm Accountancy Data Network (Eurostat)
FAO	Food and Agriculture Organisation of the United Nations
FAOSTAT	FAO Statistical Database
FAPRI	Food and Agricultural Policy Research Institute
FARM	Food and Agriculture Regional Model (AAFC)
FAS	Foreign Agricultural Service (USDA)
FBS	Food Balance Sheet (FAOSTAT)
FIPS	Federal Information Processing Standard
fob	free on board
FPS	Fixed Point Smoother
FSS	Farm Structure Surveys (Eurostat)
GCE	Generalised Cross Entropy
GIP	Gross Indigenous Production
GME	Generalised Maximum Entropy
GTAP	Global Trade Analysis Project
GUI	Graphical User Interface
HP	Hodrick-Prescott
HPD	Highest Posterior Density
HS	Harmonised Commodity Description and Coding System
ICB	International Crisis Behaviour

ICEC	Interagency Commodity Estimates Committees (USDA)
IDB	Integrated Data Base (WTO)
IFPRI	International Food Policy Research Institute
IMF	International Monetary Fund
IMPACT	International Model for Policy Analysis of Agricultural Commodities and Trade (IFPRI)
I-O	Input-Output
IPTS	Institute for Prospective Technological Studies, one of the JRC institutes
ISIC	International Standard Industrial Classification of All Economic Activities
ISO	International Organisation of Standardisation
JRC	European Commission's Joint Research Centre
KF	Kalman Filter
LEI	Agricultural Economics Research Institute
MAcMap	Market Access Map (CEPII / UNCTAD)
MS	Member State
NACE	Nomenclature statistique des Activités économiques dans la Communauté Européenne (Eurostat)
NAICS	North American Industry Classification System
NBB	National Bank of Belgium
NUTS	Nomenclature of Territorial Units for Statistics (Eurostat)
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary Least Square
PE	Partial equilibrium
PRODCOM	EU system of production statistics for mining and manufacturing (Eurostat)
PS&D	Production, Supply and Distribution (USDA)
RAS	SAM balancing method (abbreviation derived from matrix notation)
REGIO	Regional Statistics (Eurostat)
S&U	Supply and Use
SAM	Social Accounting Matrix
SBS	Supply Balance Sheet (Eurostat)
SDMX	Statistical Data and Metadata Exchange
SEATS	Signal Extraction in ARIMA Time Series
SITC	Standard International Trade Classification
SPAM	Spatial Allocation Model
SPEL/EU	Sectoral Production and Income Model for Agriculture (EU version)
SUA	Supply Utilisation Accounts (FAOSTAT)
TRAMO	Time series Regression with ARIMA noise, Missing observations and Outliers
UN	United Nations
UNCTAD	United Nations Conference on Trade and Development
UNESCO	United Nations Educational, Scientific and Cultural Organisation
UNSC	United Nations Statistical Commission
US	United States
USDA	United States Department of Agriculture
vTI	Johann Heinrich von Thünen Institut
WASDE	World Agricultural Supply and Demand Estimates (USDA)
WB	World Bank
WCO	World Custom Organisation
WDI	World Development Indicators (WB)
WTO	World Trade Organisation
ZMP	Zentrale Markt- und Preisberichtstelle für Erzeugnisse der Land-, Forst- und Ernährungswirtschaft GmbH

Data-processing (languages, general programmes, storage):
GDX (Gams Data eXchange), TS, XML, .NET, HTML, DOS, Lotus, TROLL, FoxPro, MS-Word, MS-Access, MS-Excel, MS-Windows, FORTRAN, GAMS (General Algebraic Modelling System), Gtree, Stata, Outlook, GNU, GEMPACK, CD-ROM, DVD, URL
File extension mentioned: ascii, tsv, csv, gdx, gref, txt, wk1, bn1, prn

Executive Summary

In a study carried out by the Joint Research Centre, Institute for Prospective Technological Studies (JRC-IPTS), the potentials of harmonising and improving databases that are currently used for agricultural market modelling have been analysed. In this context, it appeared as necessary to investigate also different methods to ensure the consistency of data from different sources. Currently, a large quantity of statistics on agricultural markets and trade is readily available to researchers, policy-makers and the general public. Yet this data has limited coverage (geographical, commodities, time series) and shows differences when compared to other data sources (definitions, methodology, errors). In order to overcome these shortcomings and be able to present a harmonised dataset, additional work is required.

The first step of this study was to analyse existing international agricultural market and trade databases. Depending on the respective database, a number of commodity balance items were considered: production, demand, export and import, stocks, food use, feed, seed, waste, processing, residuals, and beginning and ending stocks. Emphasis was placed on the following agricultural commodities:

- soft wheat, durum wheat, barley, maize, rye, triticale and other grains;
- rapeseed, sunflower seed, soybeans, oils and meals;
- sugar, cotton, tobacco;
- milk, butter, skimmed milk powder, cheese, whole milk powder;
- beef and veal, pork, poultry, sheep and goats.

Two general types of databases may be distinguished:

- Statistical databases provide information that is received from reporters, whether consolidated or not;
- Scientific databases require harmonised, complete, consistent, and if possible, timely data series for establishing models or other quantitative analysis methods.

A literature review on existing harmonised databases in the field of commodity balances was conducted to study procedures concerning the consolidation of agricultural product balances for market modelling. This was complemented by a questionnaire sent out in cases where limited documentation was available. The questions focused on the following topics:

- Data sources: whether the model has its own database and if so, what are its sources; how the data is assembled and which classification schemes are used; if the data is published; which model approach is used; if the process is automated;
- Data validation: what methods are used for harmonisation, completion and balancing, with special consideration given to missing or unreliable data, breaks, and regional coverage; deviating or misleading nomenclature; and
- Technical issues such as data formats, data imports and availability to external users.

In general, the databases for agricultural market models are related to certain statistical databases from the statistical departments of selected international organisations. This is not only because data from one single source considerably alleviates data management. Indeed, modellers also tend to prefer unique international statistical databases, as they are also thought to be consolidated. Normally, additional information is only retrieved if the required data is missing or of bad quality. In this frequently-occurring case, and in order to overcome

these problems modellers may apply different practices which range from expert knowledge and alternative sources via estimation techniques to complex mathematical procedures.

In general, the models' documentation hardly mentions the classifications that are explicitly or implicitly used. Instead, in most cases emphasis is placed on the sources used and the model's code. Additionally, data are mostly available only as pre-formatted sets of tables and not provided as a database, or might not be easily accessible for bulk data download.

To allow for the regular comparison of results from statistical and model databases, the diverting statistical concepts as well as the data consolidation procedures applied to the databases themselves have to be taken into account. As the classification codes are not always documented, it is difficult to set up mappings or concordance tables for different model results. Of further concern are the differing time scales employed in updating model databases.

Consolidating model databases is difficult to achieve when the data in the underlying statistical databases vary due to classification, consolidation, or data collection methods. A higher degree of compliance between statistical databases will induce a similar effect on model databases. Existing model parameters are estimated and calibrated to the respective model databases which might lead to adjustments in current simulation models.

In addition to conceptualise a uniform classification scheme of a harmonised database for market modelling purposes, efforts need to be applied to the data consolidation process itself. Such a procedure has to be supplemented by methods dealing with completion and balancing. Results of the questionnaires, as well as the literature review, reveal that scattered methods have been applied. These range from expert knowledge and simple averages up to regression and time series analysis methods weighted by complex filters and a priori information. However, it appears appropriate to distinguish between a completion and a balancing process. Nevertheless, with regard to consolidation, data generated in this process should not deviate much from the original data supplied by national countries, as these are well-known to national experts.

Following the analysis of existing databases, work focussed on the structural design of a MetaBase. The aims were to describe how a MetaBase database system should look like and also to develop a pilot version. By prototyping the database and software, the required efforts and important functionalities became more transparent. When constructing the prototype, a five spheres approach to software building developed by the Agricultural Economics Research Institute (LEI) was used. In this approach, each sphere has its own task: manager, user interface, database, business rules and output generator. The basic idea is that each sphere can be replaced without having to change any of the other spheres. An essential part of the structural design is classification (harmonisation), i.e. those techniques for completing time series (completion) and the procedures for balancing the commodity balances (consistency).

Concerning the conceptualisation of a potential MetaBase system, it can be concluded that it should at least allow tracking the original data from model and statistical databases and therefore needs to provide a comparison tool. This will require linking supply balance sheets' respective supply utilisation accounts through multidimensional mapping tables for the various classifications. However, properly setting-up these mappings requires additional input, as not all questions could have been satisfyingly answered within this study. Comparability might be improved by completing and balancing statistical databases.

The present report has been prepared by David Verhoog, Wietse Dol and Frans Godeschalk from the Agricultural Economics Research Institute (LEI), Michael Heiden and Petra Salamon from the Johann Heinrich von Thünen-Institut (vTI), and coordinated and edited by Stephan Hubertus Gay, Marc Müller and Federica Santuccio (JRC-IPTS).

1 Introduction

National and international statistical organisations devote a lot of effort to creating valuable statistical data according to well-established definitions that must be respected by all organisations delivering the basic data. This statistical data is generally collected to support policy-makers' decision–making, or to evaluate policies already in place. However, there seems to be a gap between what statistical organisations provide and what researchers and modellers require to provide policy–makers with valuable analysis. In this report, a distinction is made between *statistical databases* and *scientific databases*. Statistical databases provide information that statistical organisations received from their reporters, while the aim of scientific databases is to provide harmonised, complete and consistent data series that can be used for research and modelling. The purpose of this project is to analyse existing agricultural databases and to investigate the potentials of establishing a harmonised scientific database for agricultural commodity markets. The present report has been prepared by David Verhoog, Wietse Dol and Frans Godeschalk from the Agricultural Economics Research Institute (LEI), Michael Heiden and Petra Salamon from the Johann Heinrich von Thünen-Institut and coordinated by Stephan Hubertus Gay, Marc Müller and Federica Santuccio (JRC-IPTS).

The study focuses on both statistical databases (FAOSTAT, USDA, Eurostat) and model databases (AGLINK/COSIMO, AGMEMOD, CAPRI/CAPSIM, ESIM, FAPRI, GTAP, FARM) in order to render data discrepancies understandable and to determine ways of harmonising definitions in order to render data comparability. Due to its focus on the agricultural market, this report especially deals with agricultural commodity or product balances. However, some methods of consolidating bilateral trade statistics will be considered. Depending on the respective database, commodity balance items such as production, demand, export and import, stocks, food use, feed, seed, waste, processing, residuals, and beginning and ending stocks are considered. Here, the emphasis is on the following agricultural commodities:

- soft wheat, durum wheat, barley, maize, rye, triticale and other grains;
- rapeseed, sunflower seed, soybeans, oils and meals;
- sugar, cotton, tobacco;
- milk, butter, skimmed milk powder, cheese, whole milk powder;
- beef and veal, pork, poultry, sheep and goats.

This report will examine the question of what has been done by either providers or users in the field of consolidating international agricultural databases. Thus, an overview is provided on the methods used to generate harmonisation, completion and consistency. Such databases are necessary for setting-up agricultural economic models, and the processes employed may reflect the requirements of the respective model. Nevertheless, issues such as data sources used, methods applied for data assembly, data completion, harmonisation and balancing, as well as shortcomings, have had to be dealt with in all databases. Various classifications (harmonisation), techniques for completing time series (completion) and procedures for balancing the commodity balances (consistency) will be presented.

As part of the methodology, a literature review on existing harmonised databases in the field of commodity balances was carried out to study procedures concerning the consolidation of agricultural product balances for market modelling. However, during the project it became apparent that a literature review would provide only very limited insight into the data consolidation of statistical and model databases. Hence, a questionnaire was developed and

sent out for cases in which no written or limited documentation was available. The questions focused on the following topics:

- Data sources: whether the model has its own database and if yes, what are its sources; how the data are assembled and which classification schemes are used; if the data is published; which model approach is used; is the process automated;
- Data validation: what methods are used for harmonisation, completion and balancing, with special consideration given to missing or unreliable data, breaks, and regional coverage; deviating or misleading nomenclature; and
- Technical issues such as data formats, data imports and availability to external users.

Returned questionnaires were assessed and in cases where additional information was still required, the working groups were contacted by phone.

In an additional step, the possibility of linking data from different international databases in one harmonised framework are elaborated upon. To obtain a harmonised database that can capture available data from various international databases, it is essential to have a unique classification scheme for commodities, territories, items, years and units. Further, it is important that the time series are complete in the sense that they do not contain un-explained breaks or missing data for certain years. Another important feature of a scientific database is consistency. This means, for example, that the total value of the database for the EU-27 must be equal to the sum of the 27 underlying individual countries. The same is true for balance sheet items.

This report is structured as follows: Chapter 2 provides a general overview of the various problems and shortcomings that result when different agricultural market models or statistical databases are compared. In Chapter 3, various classification schemes are described which cover, respectively, countries or regions, products, items, years, as well as units. Different data sources and data handling are dealt with in Chapter 4, whereas consolidation methods like harmonisation, completion and balancing are explored in Chapter 5. In the first part of Chapter 5, consolidation methods found in the literature are discussed, while in the second part, actual methods employed in the databases are described. Considerations regarding a harmonised, consistent and complete MetaBase are put forward in Chapter 6. Finally, Chapter 7 puts forward some conclusions.

2 Agricultural market models and commodity balances

Changes in world agricultural commodity markets play an important role in politicians' decision-making in organisations such as national ministries and the EU. A great deal of effort is given over to analysing shifting agricultural production patterns on world commodity markets, which, e.g. are carried out in the context of ongoing multilateral trade negotiations, but also in association with more regional decisions like the expected CAP health check. To underscore the importance of these analyses, different organisations regularly publish Agricultural Outlook reports analysing and projecting long-term developments on global agricultural markets, as well as concerning major suppliers and consumers of main agricultural and food commodities. Cereals, oilseeds, sugar, cotton, beef, sheep meat, pig meat, poultry meat, milk, cheese, butter and milk powder are mostly covered by these publications. For example, regular outlooks are provided by the EU Commission (COM EU, 2007), the Food and Agricultural Policy Research Institute (FAPRI, 2006), and also by a joint effort between the Organisation for Economic Co-operation and Development (OECD) and the Food and Agriculture Organisation of the United Nations (OECD/FAO, 2006). These outlooks draw on information from widely used sources of historical data such as FAOSTAT

(2007), published by the FAO, which contains the most detailed regional and product coverage in the food and agriculture area; the PS&D (2007) (production, supply and distribution online), published by the United States Department of Agriculture (USDA) Foreign Agricultural Service (FAS), provides regular updates and short-term forecasts for the current year; and Eurostat (2007a), from the European Union (EU), supplies very detailed information on EU Member States (MS) and Candidate Countries. Besides the regularly provided baselines, policy requires the use of policy models as well-accepted instruments to support decision making.

Box 1.1: Agricultural Market Models

The agricultural market models considered in this study can basically be divided into two groups: **computable general equilibrium** (**CGE**) models and **partial equilibrium** (**PE**) models. In brief, a CGE model is a system of nonlinear simultaneous equations representing the constrained optimizing behaviour of all agents as producers, consumers, factor suppliers, exporters, importers, taxpayers, savers, investors, or government. This means that it maps the production, consumption, intra-sectoral input and trade of all economies for one country, a region or even all countries worldwide. In contrast, PE models include only those markets most immediately relevant to a problem and exclude everything else. These models investigate the impact of changes on certain sectors of the economy. The CGE and the multi-market PE models within the agricultural market model family generated, endogenously simultaneous price and quantity equilibriums for several commodity markets, given a set of assumptions including agricultural policy variables. The supply sector models generate endogenously regional or farm level supply reactions to given market prices and policy assumptions.

In its standard version, the CGE model **Global Trade Analysis Project** (**GTAP**) is a comparative statistical, multi-regional general equilibrium model which depicts the interaction between all economic sectors. Separability of primary factors and intermediate inputs, perfect competition and profit-maximizing behaviour is assumed and trade is depicted with bilateral trade matrices with consideration of the Armington approach. Political measures are included using price wedges. A detailed presentation of the standard model can be found at Hertel (1997), or at http://www.agecon.purdue.edu/gtap. Currently, the GTAP model database has 87 regions and covers 57 sectors, of which 20 sectors include agricultural and food products. The standard GTAP model differs from others considered in this report, as it tracks values instead of quantities. Thus, only quantity and price changes instead of absolute values can be derived. (Dimaranan, 2006).

AGLINK is a recursive-dynamic, partial equilibrium, supply-demand model of world agriculture, developed by the Organisation for Economic Cooperation and Development (OECD) Secretariat. It started in 1992 (OECD, 2006). **COSIMO** is a partial equilibrium dynamic agricultural model built as a complement to the current version of the AGLINK model of the OECD. The two models are solved simultaneously, and COSIMO contains both the countries included in AGLINK, plus the details for a further 29 countries and regions which are now included in the "Rest of the World" region of AGLINK. Originally limited to temperate zone commodities, COSIMO was extended in 2005 to include sugar, cotton and coffee. The COSIMO integrates data from FAOSTAT and the FAO Economic and Social Development Department (ESC) commodity balance system database that are used in the preparation of FAO's Food Outlook (FAOSTAT, 2005).

AGMEMOD is an econometric, dynamic, multi-product partial equilibrium model wherein a bottom-up approach is used. Based on a common country model template, the respective adjusted country level models were developed to reflect the specific situation of their agriculture and to be subsequently combined in a composite EU model. This approach seeks to better capture the inherent heterogeneity of existing EU agricultural systems while still maintaining analytical consistency across the country models by adhering as closely as possible to the templates (AGMEMOD, 2007).

CAPSIM is a calibrated, comparative static partial equilibrium model. It was developed on behalf of Eurostat and applied to policy-relevant analysis of the CAP. It is based on earlier model implementations such as the SPEL/EU Base System or the SPEL/EU Medium-term Forecasting and Simulation System, which were applied on various occasions for the EU Commission (Eurostat, 2003).

CAPRI is a regionalised, agricultural sector programming model for the EU that is embedded in a global spatial commodity PE model. CAPRI covers the whole of the EU-27 and Norway at the regional level (250 regions), as well as global agricultural markets. The first operational version was released in 1999. CAPRI simulates and compares ex-ante impacts of different sets of agricultural policies for a medium-term horizon (Britz, 2005).

ESIM was initially developed by the Economic Research Service (ERS) of the United States Department of Agriculture (USDA) and Josling and Tangermann in 1994, and started as a comparative static partial equilibrium model. In the meantime, it has been converted into a recursive-dynamic partial equilibrium multi-country model of agricultural production that includes consumption of agricultural products as well as some first-stage processing activities (Banse et al., 2004).

Established in 1984 by a grant from the U.S. Congress, the **Food and Agricultural Policy Research Institute** (**FAPRI**) combines the efforts of the Center for Agricultural and Rural Development (CARD) at Iowa State University and the Center for National Food and Agricultural Policy (CNFAP) at the University of Missouri-Columbia. FAPRI uses different databases containing worldwide product balances, elasticities, and policy information, as well as a recursive multi-market partial equilibrium model to analyse and project the complex economic interrelationships present in the food and agriculture industry of the considered regions (FAPRI, 2007).

The **Food and Agriculture Regional Model** (**FARM**) is a dynamic partial equilibrium policy-specific model of Canada's key agricultural markets. The model was initially developed in stand-alone commodity components at the end of the 1970s using standard econometric techniques. At the beginning of the 1980s, all components were merged into one model, called FARM. The model is used to produce a range of forecasts, from short-term quarterly forecasts to medium-term forecasts (5 years), with market analysis published in Agriculture Canada's Market Commentary. The model has also been used to produce the medium-term policy baseline published by Agriculture and Agri-Food Canada (AAFC, 2003).

The International Food Policy Research Institute (IFPRI) currently uses two primary models to analyse the economic impact of agricultural research and development and investigate possible scenarios for future global food supply and demand trends. Since the early 1990s, the Special Project Global Trends in Food Supply and Demand has used the **International Model for Policy Analysis of Agricultural Commodities and Trade** (**IMPACT**) to examine alternative futures for global food supply, demand, trade, prices, and food security. IMPACT covers 32 commodities (which account for virtually all of the world's food production and consumption) in a partial equilibrium framework. It is specified as a set of country-level supply and demand equations where each country model is linked to the rest of the world through trade (IFPRI, 2007a). The second model used by IFPRI is the **Dynamic Research EvaluAtion for Management** (**DREAM**), which is a menu-driven software package for evaluating the economic impacts of agricultural research and development. Users can simulate a range of market, technology adoption, research spillover, and trade policy scenarios based on a flexible multi-market partial equilibrium model (IFPRI, 2007b). Until now, these two models have been developed in separate modelling groups, but IFPRI is now working towards merging these two models.

All of the preceding models are based on data, thus the quality of their outcome very much depends on the quality of the underlying databases. In general, the basic statistical data for these international databases are provided by national statistical offices or ministerial agencies, and are normally of high quality. However, the problem is that data are mostly consistent in their own domain and region, but are in many cases inconsistent over the period covered (crop year versus year), between geographical regions, concerning definitions of the commodities, and with specifying the activities or items included. Even more, the coverage of a region or definition of products could have changed over time. In addition, the data may originate from different institutions within one country which use different techniques for data

processing (up and down scaling, handling of missing data, aggregation, etc.). Often, not all data required for one particular modelling approach can be found in one source. Thus, modellers need to merge different databases. Furthermore, one has to keep in mind that even if data originates from one source, they may differ when they are retrieved at various points in time. This is brought about by the fact that databases will be regularly updated if new or revised data from national sources become available.

Various model approaches and different assumptions may regularly lead to deviating projection results. But due to the data problems mentioned above, comparing model projection results will further confuse model result users as well as policy-makers. Thus, it is important to use validated data and to avoid misunderstandings concerning data labelling and concepts. Therefore, creating a harmonised, consistent, complete and timely database for agricultural market modelling offers the potential to overcome at least one of the problems associated with the various modelling results.

From the viewpoint of many modellers, the ultimate goal is to have one consistent, harmonised, complete and timely database that brings together the databases of various international institutions. This database could then serve as an analytical framework on its own, which should make it possible to validate the data and to analyse and describe those differences in the data which originate from the use of different sources. In the future, such a database could encourage increased co-operation between international institutions and lead them to come forward with a single questionnaire and only one database. The big advantage for modellers and other users would be that the discussion on the model outcomes would really focus on the results and the quality of the models and not on the underlying data. But it would also provide an advantage for other purposes and to other users, as it should enable easy crosschecks between different sources and allow discrepancies to be spotted which would otherwise go unnoticed.

2.1 Overview of problems with comparing agricultural market model results

Statistical data provided by international and national organisations and institutes is subject to close inspection, data validation and quality checks (e.g. the OECD statistical quality framework, and FAOSTAT data quality stamp). Nonetheless, during the process of extracting and analysing datasets from different sources, such as FAOSTAT, PS&D, and Eurostat, a number of differences in published data can be identified. These differences between statistical sources may often be caused by divergent product classification, time period covered, commodity aggregation, geographical region, or simply typing and processing errors. As agricultural market models greatly rely on agricultural statistical databases to create their own (scientific) model databases, they are likely to inherit these errors. Furthermore, most models require source data from more than one agricultural database and are therefore even more susceptible to such data problems.

Data availability tends to be the major problem. For example, production statistics may not be available for all required commodities. If they exist, then production statistics are mostly confined only to commercialised major food crops. Non-commercial or subsistence production, which might be a considerable part of total production in some countries, is usually not included. Also, information on commercial stocks may be available from official or marketing authorities, factories, wholesalers and retailers, but inventories of catering establishments, institutions and households may not be available. Further, certain kinds of food may not be covered because they are not included in national production statistics. For this reason, meat from game, wild animals and insects may be excluded (FAO, 2001).

Classifying, defining and naming a commodity are crucial issues. For instance, the product poultry meat has three different definitions in the FAOSTAT database, appearing under

production, trade and commodity balances. In contrast, the product skim milk powder is named differently in various databases (e.g. skim milk evaporated, SMP, nonfat dry milk, or milk nonfat dry) even though it is actually the same product. Another example is maize, which is named corn in the PS&D database, but maize in other databases (e.g. FAOSTAT, Eurostat). This may create confusion for users that would like to retrieve information and are not aware of these differences between data. Table 2.1 provides further examples of product nomenclature discrepancies.

Table 2.1: Examples of nomenclature discrepancies between international databases

Eurostat name	FAOSTAT name	OECD name	USDA-PS&D name
Rye and maslin	Rye	Rye	Rye
Flax	Linseed	-	-
Cattle	Bovine meat	Beef and veal	Meat, beef and veal
Pigs	Pig meat	Pigmeat	Meat, swine
Whole milk raw	Milk, whole, fresh	Milk	Dairy, milk, fluid

Variations in classifying countries and changing regional aggregates are another source of possible problems. In particular, the EU (EU-9, EU-10, EU-12, EU-15, EU-25, EU-27), as well as the break up of the former Soviet Union and other countries have recently impacted on national/aggregated definitions. For example, USDA data published in FAS, livestock & poultry: world markets & trade, may in some cases not yet include Finland, Sweden, and Austria, which were added in 1996 to the EC-12, resulting in the EU-15. The EU-15 will have data through 1998, while the new European Union with 25 Member States(EU-25) will include data beginning in 1999. PS&D online includes all data in the USDA database for these countries in a complete EU-25 total (PS&D, 2007).

Some databases list the more important countries for a given commodity; they do not necessarily cover the whole world. Although no actual world aggregate is provided, a user might be inclined to sum up the different countries, resulting in a value that is smaller than the overall world value. For example, for the livestock complex, the "world total" in PS&D online is the sum of only those countries included in the USDA database, which are representative countries. The representative countries in the USDA database thus capture approximately 90 percent of world livestock trade. In USDA livestock publications, this sum is more correctly labelled as "selected country total", rather than "world total" (PS&D, 2007).

Some models deliver different data sets for different countries and commodities, e.g. for one country production and consumption data of certain dairy products are available, while for another country no data on dairy products exist at all (e.g. FAPRI, OECD). These differences are sometimes difficult to comprehend, but might be explained by the nature of a specific model or because of unavailable statistics.

The time-reference period can be of additional concern. There are a number of combinations of twelve month periods available which are used to compile balance sheets (such as calendar year, crop year, market year, financial year), but depending on the database and commodity, these time periods could be defined quite differently (e.g. PS&D defines the crop year for wheat as July/June, but for rice as January/December, while FAOSTAT defines the crop year for both commodities as January/December). In some cases there are even different marketing years for the Northern Hemisphere and the Southern Hemisphere. This makes comparison of model results even more complex.

Differences are even to be found within one database. For example, a user may find that what country A officially declares as imports from country B will not correspond to what country B officially, and reciprocally, declares as its exports to country A for a given

commodity in a given year (in terms of quantity and/or value). There are a number of reasons which explain this:

- time lag: an export reported in December of a given year could reach its destination in January of the following year (and would only then be reported as an import);
- commodities could be misclassified between the exporter and importer;
- exported quantities could be destroyed or lost en route due to accidents, weather conditions, etc.;
- simple typing/calculation errors can be made by the reporting country;
- data confidentiality is required by one of the reporters;
- place of origin/final destination inconsistencies may exist;
- customs tax avoidance, e.g. exports are not declared to circumvent an embargo.

With regard to discrepancies in the reported values of trade (assuming that the corresponding reported quantities are identical): most countries report export values as Free-On-Board (fob, i.e. insurance/transport costs are not included), while import values are mostly reported as Cost-Insurance-Freight (cif, i.e. insurance/transport costs are included). Therefore, for a given agricultural commodity, the reported export value should be lower than the corresponding reported import value. The adjustment factor varies according to commodity, distance, packaging, etc. (FAOSTAT, 2007).

Concerning data providers, one could presume that data from national sources might differ depending on the methodology applied and the time frame for revisions implemented. In the case of the EU, no changes are made by MS in passing data to FAO, OECD, or Eurostat. The differences that exist between data published by Eurostat and those published by MS will therefore exist between Eurostat data and that published by these other international organisations. There are two practical reasons for discrepancies between the data published by the various international organisations:

- The issue of revisions. National practices in revising data to correct past estimates are complex and vary between MS, as does their practice of providing revisions to Eurostat and other international organisations. For this reason it is likely that the data published by different organisations and related to different generations of data may differ.
- Methods of converting national data into a common currency. For example, if the flow of data to an international organisation is only annual, it seems the conversion must be carried out with an annual factor. This will produce different results from conversion that is carried out monthly (Eurostat, 2006).

The challenge for agricultural market modelling groups is to be aware of such data discrepancies and problems while creating their own scientific databases. Moreover, additional data problems may surface during the process of compiling the data:

- breaks in time series (e.g. changes in definition);
- missing data in single years (e.g. ending stocks);
- unreliable data in single years;
- unclosed balances in a single year;
- missing time series (e.g. no data on a specific product such as skim milk powder);
- incomparable time series (e.g. production statistics versus product balance accounts).

All these possible data problems must be addressed by modelling groups. For this purpose, the groups have developed and adapted various consolidation mechanisms to guarantee complete and consistent data series. However, the consolidation techniques applied vary considerably for the models reviewed in this study and may even differ across products and countries. Thus, a comparison of baseline results from the different agricultural market models may be difficult.

2.2 Agricultural commodity balances

As described in the introduction, emphasis is placed on agricultural commodity balances and related topics. For each database, be it for statistical or modelling purposes, an inventory of the available commodity balances will first be provided, followed by a description of the available balance sheet items. The items available in principle will be stated here, but actual data may vary with regional coverage.

2.2.1 FAO

For FAOSTAT one can find the following commodity balances that can be used for both linking and comparison in this project:

- For cereals (8 balances): wheat, rye, barley, oats, maize, sorghum, millet and cereals nec (not elsewhere considered).
- For oilseeds (11 balances): soybeans, groundnuts, linseed, rape & mustard, sunflower, cotton, coconuts (including copra), palm kernel equivalents, olives, sesame and oilseeds nec.
- Sugar.
- For dairy products (1 balance): whole fresh milk.
- For meat (10 balances): bovine, pig, sheep & goat, equine, meat nec (including camel and game meat), chicken, turkey, duck & goose or guinea fowl, rabbit and edible offal.

The commodities currently distinguished in FAOSTAT are less detailed than in a previous version (old bulk downloads). In the previous version, oil and meal of oilseeds, tobacco and cotton were also distinguished.

The balance items provided not only contain less detail (production, import, export, feed & seed, other net uses and food consumption), but stocks are also missing. Similar to its commodities, the balance items of the current version of FAOSTAT are also less detailed than in a previous version.

2.2.2 EUROSTAT

For EUROSTAT (AgrIS database) one can find the following commodity balances that can be used for both linking and comparison in this project:

- For cereals (11 balances): cereals total, wheat total, common wheat, durum wheat, cereals other than wheat, rye and maslin, barley, oats and mixed grains other than maslin, maize, triticale and cereals n.e.s. ("not elsewhere specified", including sorghum).
- For oilseeds (24 balances): total oilseeds, rape and turnip rape, sunflower, soya bean, flax, cotton, olives and others. For each of these 8 commodities there are also balances available for vegetable oils and fats and oilcake.
- Sugar.
- For dairy products (10 balances): whole milk, fresh milk products except cream, drinking milk, cream, concentrated milk, skimmed milk powder, butter, cheese and processed cheese.

- For meat (8 balances): meat total, cattle, pigs, sheep and goat, Equidae, poultry, other meat and offal.

The balance items provided differ from one commodity balance to the other. The overall picture of Eurostat balances is that they are very detailed for the items (e.g. imports and exports from/to EU-9, EU-10, EU-12 and EU-15). However, this report will only focus on the main Eurostat balance items: gross indigenous production (GIP), import and export of live animals, usable production, total imports and exports, beginning and ending stocks, stock changes, total domestic uses, seed, losses, animal feed, human consumption, processing and industrial uses.

2.2.3 USDA

For the USDA, one can find the following commodity balances that can be used for both linking and comparison in this project:

- For cereals (9 balances): wheat, durum wheat, barley, rye, oats, mixed grains, corn, millet, sorghum.
- For oilseeds (24 balances): peanut, rape, sunflower, soybean, soybean (local), copra, cotton, palm kernel. For each of these 8 commodities there are also balances available for vegetable oils and meal. Further, there are two more balances on vegetable oil, one for olive oil and one for coconut oil.
- Sugar, cotton, tobacco.
- For dairy products (5 balances): fluid milk, non fat dry, dry whole milk powder, butter and cheese.
- For meat (4 balances): beef and veal, swine, poultry (broiler) and turkey.

The overall picture of the USDA balances is that they are very detailed. It is noteworthy that the USDA splits, for example, production, imports and exports in fresh, canned, green and main. This report will only focus on the main USDA balance items: production, total imports and exports, beginning and ending stocks, domestic consumption, loss, feed, human consumption, processing and industrial uses. This means that compared to Eurostat, only the balance item “seed” is missing.

2.2.4 OECD

For the OECD (Outlook) one can find the following commodity balances that can be used for both linking and comparison in this project:

- For cereals (6 balances): wheat, barley, oats, maize, sorghum and coarse grains.
- For oilseeds (7 balances): oilseeds, soybeans, sunflower seed, rapeseed, oilseed oils, palm oil and oilseed meals.
- Sugar, raw sugar and refined sugar (all in raw sugar equivalent).
- For dairy products (8 balances): milk, milk fat (pw), wholemilk powder (pw), skim milk powder (pw), butter (pw), cheese (pw), whey powder (pw) and casein (pw).
- For meat (4 balances): beef and veal (cwt), sheep meat (cwt), pigmeat (cwt) and poultry meat (rtc).

There are no commodity balances for tobacco and cotton. The balance items provided differ from one commodity balance to the other. The overall picture of the OECD balance items is

that they are very detailed. From the OECD balances, the following items will be used: production, imports, exports, beginning and ending stocks, variation in stocks, consumption, feed, food, on-farm use, deliveries, industrial use, other use, crush and waste or statistical difference.

2.2.5 FAPRI

For FAPRI one can find the following commodity balances that can be used for both linking and comparison in this project:

- For cereals (15 balances): all grain, food grains, feed grains & hay, hay, coarse grains, wheat (all, durum, spring, feed), rye, barley, oats, corn, sorghum, millet.
- For oilseeds (28 balances): total oilseeds, soybeans, peanut, canola, rape, sunflower, palm kernel, cotton, other oilseeds. Additional balances are available for oils (same commodities; instead of other oilseeds, palm and corn are distinguished) and meals (same commodities; instead of other oilseeds, corn gluten meal is distinguished).
- Sugar, cotton.
- For dairy products (40 balances): different kinds of milk, fluid milk, dry milk, milk powder, whey, milk solids, milk fat, evaporated milk, butter, cheese; only some of these balances are available internationally.
- For meat (18 balances): different kinds of beef & veal, pork, lamb & mutton, and poultry; as in the case of dairy products, only some of these balances are available internationally.

The overall picture of the FAPRI balances is that the commodities distinguished are very detailed. However, the view of the balances is both activity- and product-oriented (e.g. food grains and feed grains), so the commodities and balance items are not strictly separated. In the FAPRI balance items, much detail is given (production, imports, exports, ending stock, beginning stock, stock change, use items (domestic consumption; feed & seed & residual; loss; feed; total industrial use. For manufactured products: food; other; per capita consumption). Only seed is not distinguished as a separate balance item.

3 Classification schemes

All information and data contained in statistical and agricultural market model databases is subject to thorough classification and definition procedures. The aims of such classification measures are to:

- create a strict and detailed hierarchical organisation of categories to collect and present information at various levels of aggregation;
- enable extensive coverage of the observations;
- classify each phenomenon or object in only one category or classification (mutually exclusive categories);
- group the various categories of the classification by consistent methodological principles.

Different classification schemes exist for various forms of data such as countries, commodities, activities, or trade. Classifications can differ among countries, organisations and governmental administrations. For example, within Europe over 202 different classification systems exist for activities and products (UN, 2007).

When dealing with classifications, a distinction is made between reference, related and derived classifications. Reference classifications are established through international agreements approved by the United Nations Statistical Commission (UNSC) or other qualified intergovernmental boards (e.g. International Monetary Fund (IMF), UNESCO). Indeed, these classifications have achieved broad acceptance and official agreement, and are approved and recommended as guidelines for the preparation of classifications. Moreover, derived classifications are based upon these classifications . In most cases, they adopt the structure and categories of the reference system, but may add additional details or rearrange or aggregate items from one or more reference classifications. Derived classifications are often tailored for use at the national or multi-national level. If a system only partially refers to the reference classification, it is called a related classification. Generally, such classifications are only associated with the reference system at specific levels of the structure (Eurostat, 2002).

3.1 Countries

The current standard definition for most of the countries in the world was established by the International Organisation of Standardisation (ISO). This standard is called ISO 3166 and has established two-letter (*2-alpha*) and three-letter (*3-alpha*) codes for the various countries of the world, including independent states, dependent areas, and certain areas of contested jurisdiction or special status. Established in 1974, ISO 3166 was the result of cooperation between ISO and UN experts. The present list comprises 244 countries, with the standard codes published and revised as needed by the ISO 3166 Maintenance Agency (ISO, 2007).

The Classifications Section of the United Nations (UN) Statistics Division has also created a three-digit numerical code, included in the ISO 3166, that is used for statistical processing purposes. The current list includes those countries or areas for which statistical data have been compiled by the Statistics Division of the UN Secretariat since 30 June 1999 (UN, 2007).

The present nomenclature used in the EU for countries and territories is derived from the ISO 3166 classification. As such, it uses a two-letter (*2-alpha*) code for countries and dependent areas taken from the international ISO standard and is mostly identical to that of the ISO classification[1]. The code also includes a three-digit numerical code which differs from the ISO 3166 (Eurostat, 2007b). Note that Eurostat contains one important difference from ISO 3166: for the United Kingdom, it uses the code 'UK', while the ISO 3166 code is 'GB'. Other organisations may still keep 'GB' if they follow the ISO 3166. The Federal Information Processing Standard (FIPS) classification is an official U.S. government listing of codes which is compulsory for all U.S. agencies to use. The FIPS country list differs from ISO 3166 in some minor ways.

Several additional country aggregates (e.g. EU-15, EU-27) have been added to this UN classification to establish a link between the different databases involved.

A number of other country classification systems exist, but are not important in the context of building agricultural databases (e.g. Internet country codes, telephone country codes, international vehicle codes).

3.2 Activities and products

Economic activities in the fields of economic statistics, population, production, employment, national income and others have been classified by the International Standard Industrial Classification of All Economic Activities (ISIC) since 1948. A number of countries have

[1] The letter codes differ for Kosovo, Montenegro, Serbia, Finland, France, Morocco, and Norway.

utilised ISIC, which is maintained by the UNSC, to develop their national industrial classification. The present revised edition is no. 3.1, but draft no. 4 is currently being developed. Despite containing the word "industrial" in its name, it is not just a classification of industries (UN, 2007).

ISIC is a classification according to type of economic activity, not a classification of goods and services. The activity carried out by a unit is the type of production in which it engages. This is the characteristic of the unit according to which it will be grouped with other units to form industries. An industry is defined as the set of all production units engaged primarily in the same or similar kinds of productive economic activity (UN, 2007).

The complementary classification of the EU is called NACE (Nomenclature statistique des Activités économiques dans la Communauté Européenne). It has been in use since 1970 and is used to designate the various statistical classifications of economic activities within the EU. The original NACE classification was developed separately from the ISIC classification and was not covered by community legislation. It therefore suffered from major drawbacks. In particular, it offered poor comparability with national EU classifications and the recognised international framework. A joint initiative in the late 1980s between the UN Statistics Office and Eurostat developed a third revision of the ISIC classification. Subsequently, starting from the structure of ISIC revision 3, the NACE structure was modified and special features of national classifications were introduced in the process. NACE revision 1.1 is the present classification of economic activities corresponding to ISIC revision 3.1 at the European level. All MS of the EU have to use this classification to compile their economic activities or must use a national classification derived from it (Eurostat, 2002).

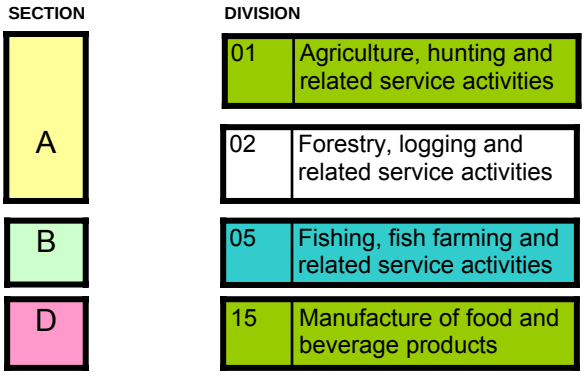
The NACE revision 1.1 is a simple subdivision of the ISIC revision 3.1. The coding comprises:

- a first level that consists of headings identified by an alphabetical code (sections), an intermediate level consisting of headings identified by two-character alphabetical codes (subsections) (refer to Figure 3.1);
- a second level consisting of headings identified by a two-digit numerical code (divisions);
- a third level consisting of headings identified by a three-digit numerical code (groups);
- a fourth level consisting of headings identified by a four-digit numerical code (classes).

The first level of ISIC revision 3.1 (sections) was taken over unchanged in NACE revision 1.1, but disaggregated into subsections in some areas. The second level of ISIC (divisions) was taken over without any modifications. The third and fourth levels (groups and classes) of ISIC were subdivided according to European requirements. However, the groups and classes of NACE revision 1.1 can always be aggregated into the groups and classes of ISIC revision 3.1 from which they were derived (Eurostat 2002).

NAICS, the North American activity classification, was developed to provide common industry definitions for Canada, Mexico and the United States in order to facilitate economic analyses of these countries' economies. Easily converting NAICS data into ISIC/NACE is not possible, as NAICS is constructed on a production-orientated or supply-based conceptual framework, which creates industries that do not translate into ISIC two-digit codes (Eurostat, 2002).

Figure 3.1: Structure of NACE with respect to agriculture and food products



Source: Angelini (2003)

As it is not possible to establish a one-to-one correspondence between activities and products, ISIC or NACE are not designed to measure product data at any detailed level. A separate classification was developed for this purpose, namely, the Central Product Classification (CPC) (UN, 2007). Before the CPC was developed, the international system did not have any classification that encompassed both goods and services. The CPC was devised by the UN in 1989 and was created to provide a structure for comparing many different kinds of statistics concerning goods and services. Its aim therefore is not to replace other product classifications, but rather to enable the latter to be harmonised in such a way that data can be transposed into the relevant CPC categories. The CPC can thus be seen as a means of harmonisation at both the international and national level. Regarding goods, the CPC uses the headings and subheadings of the Harmonised Commodity Description and Coding System (HS) as building blocks, i.e. every heading at the lowest level of the CPC corresponds exactly to a heading or subheading of the HS, or to an aggregation of two or more HS headings or subheadings. With respect to goods, therefore, the definition of categories in the HS is used as the basis for classification in the CPC. The CPC has its own coding system which is independent of ISIC revision 3.1. The most recent version is the CPC revision 1.1. The criterion according to which the CPC arranges products is their "material composition and nature (properties)". This includes, for example, the type of raw material used, the production process involved, the purpose for which the goods are intended, etc. (Eurostat, 2002).

The Statistical Classification of Products by Activity in the European Economic Community (CPA) is the European version of the CPC, and the purposes it serves are in line with those of the CPC. Whilst the CPC is merely a recommended classification, however, the CPA is legally binding in the European Community. In addition, specific survey classifications are linked to the CPA unless the CPA is itself used as a survey classification. Although the CPA is the European counterpart of the CPC, it differs from the latter not only in that it is more detailed but also regarding its structuring. The view at the European level is that a central product classification should be structured according to the criterion of economic origin, with the framework (and thus the definition of the economic activities) being based on NACE revision 1.1. This recourse to NACE revision 1.1 with respect to the definitions of economic activity means that the CPA's structure corresponds at all levels to that of NACE revision 1.1. Finally, further subdivisions are included in line with the specific requirements of the Community and the individual MS. As a result of this breakdown, the CPA has more subcategories than the CPC has subclasses. Since the elements of the CPA are based on those of the CPC, links between the CPA and the HS exist in the same way as those which exist between the CPC and the HS, as referred to above (Eurostat, 2002).

NACE and CPA nomenclatures are not commonly used in Eurostat agricultural statistics for two main reasons, namely:

- NACE Revision1.1 does not fully reflect the reality of agricultural holdings in the EU. Economic and structural statistics for agriculture are therefore based on a specific classification, the "Community Typology for Agricultural Holdings". This results in a lack of comparability between agriculture and other sectors.
- Due to the weakness of NACE, the CPA Class 01 (agricultural products) does not match the requirements of agricultural applications. More precisely, it is very difficult, in the framework of the current CPA, to represent most product aggregations used in agricultural production statistics.

HS, the international customs product classification drawn up by the World Customs Organisation (WCO) for foreign trade, plays a key role in developing the revised international system of economic classifications by providing the building blocks for the central product classifications. Since 1988, a number of countries have been using the HS for both customs tariffs and foreign trade statistics purposes. The HS is a hierarchically structured goods classification divided into 96 chapters, each of which are identified by means of a two-digit numerical code. The chapters are subdivided into headings, which are in turn subdivided into approximately 5000 subheadings. The headings are identified by a four-digit code, while the subheadings utilise a six-digit numerical code (Eurostat, 2002).

The Combined Nomenclature (CN), introduced in 1988, is the classification used within the EU for foreign trade purposes and provides a degree of detail beyond that of the HS. Headings in the CN are identified by means of an eight-digit numerical code. Additional subdivisions within the CN are introduced with the EU's specific customs and foreign trade statistics requirements in mind. The CN is revised every year and, as a Council Regulation, is binding on the MS (Eurostat, 2002).

The HS and CN are in sense multi-purpose classifications for both customs and statistical applications. They are therefore heavily concerned with the nature or material of the products. For analytical purposes, alternative classifications may be used. Certain results are presented in accordance with the Standard International (SITC), which is managed by the United Nations Trade Classification. SITC was originally created to compile international trade statistics on all merchandise entering international trade, and to promote the international comparability of international trade statistics. The commodity groupings of SITC reflect (a) the materials used in production; (b) the processing stage; (c) market practices and uses of the products; (d) the importance of the commodities in terms of world trade; and (e) technological changes. Aggregated data on trade are often presented by the one- and two-digit SITC categories. The SITC corresponds with classifications of the Broad Economic Categories (BEC), the HS, the CPC and the ISIC (UN, 2007).

Before the adoption of HS and CN, external trade statistics used a product classification called Nimexe. This is no longer used, but can still be found in some historic data series. Data on foreign trade may sometimes be published and analysed by a number of other classifications, all of which can be related to the finest HS/CN headings that are used to collect basic data. BEC classification permits the conversion of international trade data compiled by the SITC into end-use categories that are more meaningful for economic analysis and which fall within the framework of the System of National Accounts (capital, intermediate and consumer goods) (Eurostat, 2006).

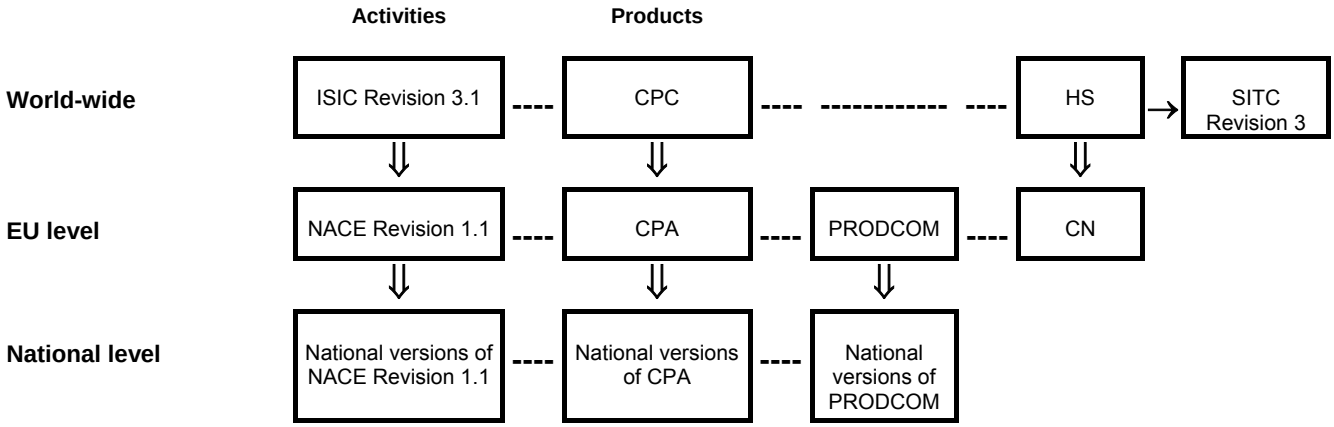
"PRODCOM" is the abbreviation for the EU system of production statistics on mining and manufacturing (i.e. excluding services). The product classification (PRODCOM list), upon which production statistics are based, is drawn up each year by the PRODCOM committee. The headings of the PRODCOM list are derived from the HS or the CN, which thus enables

comparisons to be made between production statistics and foreign trade statistics. PRODCOM headings are coded using an eight-digit numerical code, the first six digits of which are identical to those of the CPA code. The PRODCOM list is therefore linked to, and consistent with, CPA (Eurostat, 2002).

Figure 3.2 displays an overview of the relations between international and national product classifications from an EU perspective. For activities, the standard international classification is the ISIC revision 1.1. From this, the European classification NACE revision 1.1 is derived. All MS of the EU must use this classification to compile their economic activities or must use a national classification derived from it. For products, the European production classification CPA is derived from the international CPC. Concordance tables are available between ISIC and CPC, as well as between NACE and CPA. The HS classification is the standard for international trade. Eurostat has derived the CN classification from HS and this is applied by all EU MS.

In general, all standard classifications of the same level displayed in Figure 3.2 (international, EU, national) correspond with each other, i.e. extensive concordance tables are available.

Figure 3.2: Relations of international classifications for activities, products and trade



Source: Eurostat (2002)

In contrast to the product classifications described above, supply balance sheets (SBS) are not directly linked to any classification scheme. In fact, SBS are compiled with data from different product classifications (such as production and trade classifications). Since data from production and trade statistics are not directly comparable, statistical databases make use of conversion factors to adjust for those differences. In general, conversion is done to pass from a processed product to the raw product, or to pass from a compound product to a part of its raw product expressed in the unit of the balance. Eurostat, for example, establishes foreign trade for the SBS by following a list of selected products corresponding to the tariff and statistical nomenclature and to the Common Customs Tariff in force. The product list refers to the CN from which it is extracted and includes all the products related to the concerned SBS: the raw product, the processed products and the manufactured products. For instance, meat balance sheets are compiled in 1000 tonnes of carcass weight. For trade between countries (MS, third countries), the weight of all meat products is converted in tonnes of carcass weight using a coefficient which considers the share of meat in the product and the conversion of boneless meat to carcass weight (with bones) (Eurostat, 2007c). In the case of the EU, reference coefficients have been set by Eurostat, but the MS are entitled to use their own coefficients (Eurostat, 2005b).

To sum up, there are basically only three directions in which to establish a classification for commodities.

- One direction is a classification for commodities based on activities. Following this direction would end up with industry classifications such as the International Standard Industrial Classification of all Economic Activities (ISIC) and the complementary classification of the EU, which is called NACE. A major drawback of these two classifications is that there is no one-to-one correspondence between activities and commodities. In many cases one activity produces more than one commodity.
- Another direction is to create a classification for commodities based on products. This would lead to the internationally used Central Product Classification (CPC) from the United Nations and its European counterpart, CPA. The difference between the CPC and the CPA is that the CPA is structured according to the criterion of economic origin. This means that the CPA product framework is strongly linked to the NACE revision 1.1 activity framework. Since this project deals with commodity balances, it seems logical to base the commodity classification on one of these classifications.
- The third direction would lead to classifications for trade commodities. For example, HS is the international customs product classification drawn up by the World Customs Organisation (WCO) for foreign trade and is a hierarchically structured goods classification. The Combined Classification (CN) is the classification for trade commodities used by the EU; it contains more detail than the HS.

3.3 Items

This review covers agricultural commodity balances but does not include bilateral trade statistics. Therefore, one of the objectives of this study is to determine which agricultural SBS item classifications are available and to what extent they have been utilised by the reviewed model groups.

A commodity balance presents a comprehensive picture of a country's agricultural product and food supply pattern during a specified reference period (e.g. calendar year, market year). In general, a balance sheet shows the total amount of a specific crop or processed product available (= supply) and how this supply is distributed to various elements of the economy (= utilisation). The different elements of such a balance sheet (e.g. production, import, export, human consumption) are called items, but depending on the source might be named differently, i.e. element, activity or subject. The quantity of a product compiled in a balance sheet is usually expressed in metric units, but this depends on the specific nature of the commodity (e.g. milk in litres, wheat in tonnes, cows per head).

To construct a balance sheet, the following principle problems have to be solved:

- A country should have a comprehensive statistical system which records all current information relating to each component of the balance system.
- The concepts of the adopted information should be similar to those of the balance sheet.
- The information available should be consistent, at least with respect to measurement unit and time reference period.

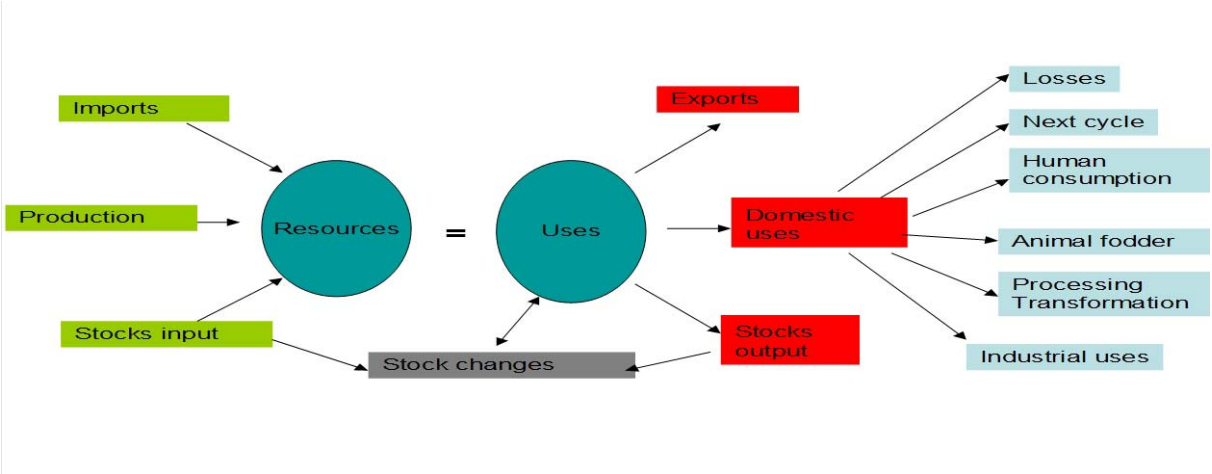
In practice, however, such an ideal statistical system does not exist. Therefore, the basic data are necessarily based on a large variety of sources. As this is the case, the data are subject to inconsistency, and their concepts and time references may not be consistent throughout (FAO, 2001).

First attempts at preparing food balance sheets were undertaken in 1936, but they became increasingly important during and shortly after the Second World War. In 1948, the fourth FAO Conference recommended that governments be encouraged to develop their own food balance sheets. Since then, the FAO has taken a leading role in further developing food balance sheets. As balance sheets are assembled from a variety of sources, their quality and coverage vary considerably among countries and commodities, even in the work of the FAO itself.

The purpose of a balance sheet can be either market management or food management. In the EU, SBSs are used as market management instruments and provide an overall view of an agricultural product's market by country and for the whole EU. Market management means regulating the availabilities between domestic use and the needs of external trade under price and budgetary considerations. Food management is a concept applied by the FAO to supply food to the world's human population first and then to its animals (Eurostat, 2005b). Depending on the type of SBS (e.g. cereals, oilseeds, beef), the balances are compiled at the MS level or by Eurostat.

The following figure shows the different components of a SBS in its general form applied by Eurostat. For a given product or country, all the components are not mandatory. In the case of no production, imports are required; if the raw product is not directly consumable, the main domestic use is processing.

Figure 3.3: The Eurostat supply balance sheet



Source: Eurostat (2005b)

In general, balances are compiled for a reference period (a calendar year or marketing year) and with regard to a geographical entity to compare the resources of a product, which are made up of: production; stocks at the start of the year; imports and its uses for internal use; final stocks; and exports. Internal use can in turn be broken down into losses, human consumption, animal feed, processing, industrial uses and preparations for the next production cycle (seed). Even if the methodology is clear, in practise for some balances it is difficult to collect all the data. Availability and the need for information vary by product (Eurostat, 2005b).

The coverage and complexity of compiling agricultural balance sheets can become quite extensive. A look at statistical data published by Eurostat in 2005 reveals 156 balances for crop products, meat, milk, eggs, sugar and fats (Eurostat, 2005a).

The USDA calls its commodity balances supply and distribution, or supply and use tables. These tables are available through the USDA's production, supply and distribution (PS&D) database. The PS&D database's supply and distribution tables are the standard method of

accounting for the total supply of a particular commodity and the total use of that same commodity. Supply and use tables are usually based on the marketing year for the commodity because this is the period in which the supply and use will balance. In every supply and use table, supply must equal use in each marketing year.

- Supply = beginning stocks + domestic production + imports.
- Use = domestic consumption + exports + ending stocks.
- Domestic consumption = all possible uses of the commodity, i.e. food, feed, seed, waste, and industrial processing.

Stocks include all of the commodity not currently in use, regardless of where it is stored, be it: on the farm; in a warehouse or elevator awaiting marketing; at a port awaiting shipment; at a mill awaiting use; or in an emergency government reserve.

- Ending stocks = the unused commodity remaining at the end of the marketing year to be used in the next year.
- Beginning stocks = the ending stocks carried into the new marketing year from the previous year.

Although supply and use data are balanced for the local marketing year, the PS&D database system also includes, for comparison purposes, standard international trade years for some commodities (PS&D, 2007).

The FAO uses two balance concepts: supply utilisation accounts (SUA) and food balance sheets (FBS). The SUAs are time series data dealing with statistics on supply (production, imports and stock changes) and utilisation (exports, seed, feed, waste, industrial use, food, and other use) which are kept physically together to allow for the matching of food availability with food use. The SUA's statistical framework has been developed to provide a useful statistical tool for the preparation, conduct and appraisal of government action aimed at developing and improving the agricultural and food sectors of national economies.

FBS basically deal with food items, i.e. primary commodities and a number of processed commodities potentially available for human consumption. The total quantity of foodstuffs produced in a country, added to the total quantity imported and adjusted to any change in stocks that may have occurred since the beginning of the reference period, provides the supply available during that period. On the utilisation side, a distinction is made between the quantities: exported; fed to livestock; used for seed; put to manufacture for food use and other uses; that are lost during storage and transportation; and food supplies available for human consumption. The per capita supply of each food item available for human consumption is then obtained by dividing the respective quantity by the related data on the population actually partaking in it. Data on per capita food supplies are expressed in terms of quantity, and by applying appropriate food composition factors for all primary and processed products, also in terms of caloric value, protein and fat content (FAOSTAT, 2007).

The variables in a supply balance are statistical variables, and whilst they provide a broad idea of the actual situation, they approach it in ways which can differ considerably. For example, SBS and SUA are assembled from a variety of sources, and the quality of their balance sheets and coverage vary considerably among countries and commodities. Inaccuracies and errors may be introduced at each stage of a balance sheet's construction. The user of these data must therefore bear in mind their limitations. For while production statistics deliver high quality data for most agricultural products in general, there are various shortcomings concerning other data sources such as: trade statistics (e.g. EU intra-trade, coefficients to convert processed products to the weight of primary equivalents); data on stocks (e.g. wholesale and industry stocks are mostly excluded); and domestic uses (e.g. losses are estimated, consumption of tourists is not included).

As described above, SBS product classifications do not exist, and the same can be said about SBS items. Rather, each international statistical database uses its own set of balance items for which it has developed comprehensive definitions and guidelines. In some cases the item definitions differ slightly among the databases or have become so detailed that almost each commodity or group of commodities has its own set of SBS items. In Table 3.1 a simple concordance of items of the main statistical databases has been put together.

Table 3.1: Item concordance for the main international statistical databases[2]

Eurostat	FAOSTAT	USDA – PS&D
Usable Production	Production	Production
Change in stocks	Changes in stock	Beginning/ending stocks
Total imports	Gross imports	Imports
Resource/Use	Supply	Total supply
Total exports	Gross exports	Exports
		Total distribution
Animal feed	Feed	Feed
Seed – total (including eggs for hatching)	Seed (including eggs for hatching)	Seed (grains, oilseeds)
Processing	Food manufacture	
Losses	Waste	Waste & Losses (oilseeds, horticulture)
Industrial uses		
		Non-feed (industrial purpose, seed, human consumption)
	Other uses	
Gross human consumption	Food	
		Total domestic consumption (feed + non-feed)
Gross human consumption per capita	Per caput supply	
Degree of self-sufficiency		

Source: Own compilation

In the following, the individual balance items and the differences in their definitions will be examined.

i. Production:

Eurostat defines usable production as those usable quantities resulting from the production process during the reference period, with the understanding that the losses suffered during this process and up to delivery do not appear in this item. The FAOSTAT definition explicitly includes non-commercial production and production in kitchen gardens. For both databases, production is reported at the farm level for primary crops and livestock items.

[2] Some of the items displayed in this table might not be available from the FAOSTAT or USDA-PS&D websites.

- Cereals: usable production for Eurostat includes marketed production, misrepresented quantities, self-provided quantities, self-consumed quantities and on-farm losses (handling, waste, pest damage, etc.), but excludes non-harvested quantities, losses before the harvest (on the plot) and losses at the time of transportation from the plot to the seat of the holding. The USDA definition for rice production distinguishes between milled and rough (unprocessed) rice.
- Oilseeds: for the USDA, production for oilseeds covers the weighed quantity of meal, oil, or dry, unprocessed seeds harvested within a specified 12-month period, measured in metric tonnes prior to processing. Unshelled peanuts are included.
- Meat: at Eurostat, usable production (net production) is the overall tonnage of meat found suitable for human consumption by the qualified health services. This meat comes from all the animals slaughtered in the respective country, be they of domestic or foreign origin. The USDA has a similar definition. As a general rule, all data on meat are expressed in all three international databases in terms of carcass weight.
- Dairy: the Eurostat definition of milk involves the total quantities of milk resulting from milking, including own consumption and quantities used for animal feed (milk used to feed calves is excluded). The USDA distinguishes between cow milk and other milk (mainly from sheep, goats, and buffaloes) and combines the two in total milk (but also excludes milk suckled by calves).

ii. Stocks:

The change of stocks corresponds to the development of stocks wherever they are held during the reference period, i.e. closing/ending stock or opening/beginning stock. In general, the following definitions are used:

- opening/beginning stock: the unused but stored quantities, existing on the first day of the reference period and coming from the previous reference period(s);

- closing/ending stock: the stored quantities existing on the last day of the period.

Instead of stocks existing at the beginning and end of the reference period, the changes of stocks during this period are frequently used. For some products, despite existing stocks, the stock variation is not significant. For highly perishable goods (for example, fresh vegetables), it is possible to ignore stocks and stock changes.

The Eurostat definition of stocks includes producers' stocks, public stocks and security stocks, intervention stocks and stocks on markets (including wholesale, importers/exporters, and processing plants). Retail stocks and household stocks are not included. FAOSTAT has a similar definition for stocks, but states that in practise, the information available often relates only to stocks held by governments, and even this, for a variety of reasons, is not available for a number of countries and important commodities.

- Cereals and oilseeds: the USDA defines the beginning stocks for cereals and oilseeds as the quantity of meal and oil (for oilseeds), unprocessed, dry seeds, in metric tonnes, held in all known storage facilities, or in transit to those facilities, at the beginning of the specified, 12-month marketing year period, which normally corresponds to a local crop/marketing year (beginning stocks Y_t = ending stocks Y_{t-1}).
- Meat: the USDA defines the beginning stocks for meat as of January 1 = ending stocks of the previous year. The same definition is applied by Eurostat and FAOSTAT.

iii. Trade, imports and exports:

The item defined as trade can be split into imports and exports. According to Eurostat, the source used for the “Imports” and “Exports” items of the supply balance sheets are the official Foreign Trade statistics which cover goods (gross products or processed products):

- which enter or leave the statistical territory of the Community (extra-Community trade in the case of the EU);

- which circulate between the statistical territories of the MS (intra-Community trade).

Apart from commercial trade, FAOSTAT explicitly includes food aid granted on specific terms, donated quantities, and estimates of unrecorded trade. As a general rule, figures are reported in terms of net weight, i.e. excluding the weight of the container. Eurostat maintains the distinction between intra-EU and extra-EU trade remains in order to get external EU trade with third-party countries, which is reported in the EU-level SBS.

The USDA definition of import and export includes local marketing years, standard international trade years and calendar years. For each commodity or group of commodities, different time-reference periods are defined that do not necessarily have to correspond with each other.

iv. Supply:

There are various ways to define supply. For example, FAOSTAT has three different definitions for the term:

(a) Production + imports + decrease in stocks = total supply

(b) Production + imports + changes in stocks (decrease or increase) = supply available for export and domestic utilisation

(c) Production + imports - exports + changes in stocks (decrease or increase) = supply for domestic utilisation

In recent years concept (c) has been adopted when preparing and publishing balance sheets in order to identify the quantity of the commodity in question which is available for utilisation within the country.

At Eurostat, the absolute stocks at the beginning or end of a period are not always available. In this case, only the variation of stocks can be indicated. Consequently, the lay-out of the statement is as follows:

- Resources = Production + Imports

- Uses = Exports + Variation of stocks (= closing stock - opening stock, therefore ≥ 0 or ≤ 0) + Internal Use.

However, when data regarding stocks are provided by the MS, they are always incorporated into the statements.

- Cereals and oilseeds: the USDA has identical definitions for the two commodity groups:

  Total supply = beginning stocks + production – total imports

  Total distribution = total exports + total domestic consumption + ending stocks

- Milk: for milk, the USDA defines total supply as total production.

v. Domestic uses

This is a balance item that comprises a number of sub-items. In general, domestic use is a country’s overall consumption of a product. However, Eurostat defines it as the quantity that is shared between seeds or eggs for hatching, losses, human consumption, animal feed,

processing and industrial uses. While FAOSTAT applies a similar definition, the USDA distinguishes domestic use (designated as total domestic consumption) basically in feed and non-feed consumption. Depending on the commodity, non-feed consumption may include different items (such as seed, human consumption, industrial purposes).

Definitions for the sub-items are as follows:

- Seed: Eurostat defines this as the quantities of raw product used for the following production cycle. In contrast, FAOSTAT applies a more detailed definition: all amounts of the commodity in question used during the reference period for reproductive purposes, such as seed, sugar cane planted, eggs for hatching and fish for bait, whether domestically produced or imported. Whenever official data are not available, seed figures can be estimated either as a percentage of production or by multiplying a seed rate with the area under the crop for the subsequent year.

- Losses/waste: losses, in general, are fixed in the most realistic way possible, but can be an estimate in percentage terms (according to experience). Losses that occur during pre-harvest and harvest are not included here – they have already been subtracted from production (both for Eurostat and FAOSTAT). This item comprises mainly post-harvest (such as waste, loss in sorting, loss after the wine harvest statement), marketing (e.g. the result of untimely harvesting and improper packing and/or transport) and production losses (e.g. transformation of primary products by extraction). At the USDA, only oilseeds and horticultural products have waste included as an item.

- Human consumption: this is defined by Eurostat as the quantities of foodstuffs self-consumed or produced by the agro-industry for consumption by the inhabitants of the territory during the reference period. Moreover, it involves the quantities delivered in various forms (unprocessed, processed, preserved, etc.) by the wholesalers to the retail trade, to the communities (canteens, restaurants, hospitals, etc.) and the quantities consumed directly by the producers. The losses and changes of stocks at the retail trade level and at the consumer level also appeared in this item. The FAOSTAT definition is much more general, but covers basically the same concepts.

- Feed: this position shows the quantities of raw products used for feeding animals during the reference period. It covers the quantities of raw products produced and consumed on the holdings (direct animal feeding) and those delivered to the animal fodder industry. The FAOSTAT definition includes the statement "whether domestically produced or imported". At the USDA the item of feed is available and defined only for cereals, oilseeds and milk.

- Processing: This item defines the transformation of a raw product into another food product for which a balance sheet is compiled. In Eurostat, the processing item of the balance sheet is created to establish a link between the supply balance sheet of a product (raw) and that of another product resulting from the processing of the first. This position shows the quantity of raw product processed. In FAOSTAT balances, two types of processing are distinguished: food (e.g. sugar, fats and oils, alcoholic beverages) and non-food (e.g. oil for soap). Non-food processing is shown under other uses. At FAOSTAT, processed products do not always appear in the same food group. While oilseeds are shown under the aggregate Oilcrops, the respective oil is shown under the Vegetable Oils group; similarly, skim milk is in the Milk group, while butter is shown under the aggregate Animal Fats. Barley, maize, millet and sorghum are in the Cereals group, while beer made from these cereals is shown under the Alcoholic Beverages group. The same principle applies for grapes and wine.

- Industrial uses: This sub-item only appears at Eurostat and is defined as the quantities used by industry during the reference period, which are intended neither for human consumption nor for animal feed.

- Other uses: This is applied by FAOSTAT and covers mainly food consumed by tourists, as well as the amounts of the commodity in question used during the reference period for manufacture for non-food purposes (e.g. oil for soap). Statistical discrepancies are also included here, defined as an inequality between supply and utilisation statistics.

vi. Derived information:

This includes mainly two items: per capita consumption/per caput supply and degree of self-sufficiency.

- Per capita consumption: in general, human per capita consumption is obtained by dividing total human consumption by the number of inhabitants. In Eurostat balances the calculation of consumption per capita uses population data that appears in official statistics. The following dates are retained:

  - 31 December for balance sheets per marketing year;

  - 30 June for balance sheets per calendar year.

Statistics refer to the residing population of each country: persons normally residing in a country but temporarily absent are included in the total population figure, while foreigners temporarily residing in the country are excluded from it for the same reasons. This is handled differently at FAOSTAT. Per caput supplies in terms of quantity are derived from the total supplies available for human consumption by dividing the quantities of the food element by the total population actually partaking of the food supplies during the reference period, i.e. the present in-area (de facto) population within the present geographical boundaries of the country in question at the mid-point of the reference period. Accordingly, nationals living abroad during the reference period are excluded, but foreigners living in the country are included. Adjustments should be made, wherever possible, for part-time presence or absence, such as temporary migrants and tourists, and for special population groups not partaking of the national food supply such as aborigines living under subsistence conditions and refugees supported by special schemes (please refer to Annexe C for more detailed definitions).

3.4 Years

The time-reference period to be used in preparing balance sheets may create problems. As a rule, product balance sheets are based on twelve-month periods. However, several twelve-month periods, such as July/June, October/September, April/March, are used for different commodities and applied by various organisations. For example, the July/June period may be defined as the "crop year" for a certain commodity in the Northern Hemisphere (e.g. wheat, barley, maize), while a crop year in the Southern Hemisphere may cover a different period. Presently, none of the applied classifications for years satisfactorily and uniformly covers the production of all agricultural commodities, their trade and domestic utilisation. In fact, there is no single twelve-month period which is fully suitable for recording the supply and utilisation of all products.

The FAO uses the calendar year time-reference period (January/December) as a standard for preparing their balance sheets, even though this time-reference might not be a completely satisfactory solution. The application of a calendar year during which the bulk of the harvest takes place also helps in linking agricultural statistics with those of the industrial and other sectors of the economy (FAO, 2001).

The Eurostat standard for compiling SBS is the calendar year for livestock products and the twelve months of a marketing year for crop products. Certain balance sheets (vegetable, fat balance sheets) are worked out by calendar year, as well as by marketing year. The marketing years for the various SBS crops can start at different dates. Council Regulations

on the common organisation of the markets specify the starting and end dates of the marketing year required by the Community (Eurostat, 2005a).

The USDA compiles its supply and use tables for the livestock complex based on a calendar year. The situation for cereals and oilseeds is a bit more complex. For the group of cereals, the following time-reference periods are applied:

- Local marketing year: the quantity of a commodity, usually at specified prices/terms, during a specified 12-month period corresponding to that country's marketing year, and measured in metric tonnes. Although the supply and use data are balanced for the local marketing year, the PS&D Online system, for comparison purposes, also includes standard international trade years for some commodities.
- For Grains, standard international trade years have been created in the USDA database to allow more accurate aggregation of total world trade at a given point in time. Adjustments are based on countries' monthly trade data. These standard international trade years are the years for which the "official" published USDA grain trade data is based. They are:
    - Wheat: July/June.
    - Coarse grains: October/September.
    - Rice: January/December calendar year following the marketing year indicated (e.g. 1990 rice market year shows standardised rice trade for calendar year 1991).
    - Total Grain: an aggregate of the previous 3 types of years, without adjustment.
- For oilseeds, standard trade years vary among commodities. Most seeds, meals, and oils are reported on either an October/September or calendar year basis. The exception is olive oil, which is reported on a November/October year. Calendar year figures are provided for smaller countries and minor commodities. Major oilseed producing countries are on a local marketing year roughly based on harvest dates or an October/September year. For Northern Hemisphere countries, these local marketing years are:
    - Soybeans: September/August.
    - Rapeseed: July/June.
    - Sunflower seed: September/August.
    - Peanuts: August/July.
    - Cottonseed: October/September.
    - Copra: October/September.
    - Palm kernel: October/September.

Southern Hemisphere marketing years generally begin in March or April. For cotton, all data is reported according to an August/July trade year (PS&D, 2007).

It is complicated to choose, from possible harvest years, the right common denominator from each data source for each product. From the literature review it became clear that there are basically two ways to link the different characteristics of a year to the calendar year:

- Simple adding up: monthly data can be aggregated to annual data or to the crop year required. This simple adding up of monthly data can be done for addable variables such as production quantities and values. Weighting is necessary when the monthly information of prices will be added to yearly information.
- Retrieving data for the calendar year from harvest years can be done by using coefficients like 2/3 of harvest year (t) and 1/3 of harvest year (t-1) in the case of April/May. When the months of the harvest year are given, different coefficients can be used.

For the prices of agricultural commodities, the international databases normally provide annual data, though sometimes one can also find monthly data. When this is the case, there is a need for some weighting to calculate prices for calendar years. For commodity balances, the second option for linking the years is most important. However, based on earlier experience with linking different datasets, it became clear that, at least for EUROSTAT, the agricultural commodity balance of year (t) must be mapped to the production from year (t-1) of the production statistics. This means that one has to be careful and not just recalculate all the harvest year information into calendar year information.

3.5 Units

Worked out from various sources of data with their own representation, the SBS of Eurostat are compiled using a common unit with the capability of converting a processed product into a raw product or a compound product into the reference of the SBS. The most frequently used units refer to weight in 1000 tonnes: in the case of meat, 1000 tonnes of carcass weight; for eggs, 1000 shell-eggs tonnes (Eurostat, 2005b).

At FAOSTAT, commodities are presented in total quantity (1000 metric tonnes) per commodity (FAOSTAT, 2007).

The USDA uses metric tonnes for all commodities except cotton (cereals in 1000 metric tonnes, horticultural products in metric tonnes, wine in 1000 hectolitres, sugar in raw sugar equivalents, tobacco in metric tonnes, red meat in 1000 tonnes carcass weight, poultry in 1000 tonnes ready-to-cook weight, and products such as fluid milk, dry milk, cheese, etc., are in 1000 tonnes) (PS&D, 2007).

To convert dimensions, it is important to clarify the relationship between units and dimensions, i.e. a unit indicates a group of dimensions that can be converted into each other. For instance, the unit “Weight” contains the dimensions “Kg”, “Pound”, “Ounces”, etc. Converting weights is very straightforward and is done by fixed constants. For example, “Currency” contains the dimensions “Dollar” and “Euro”. For converting currencies it is important that a time series of the exchange rates is available.

3.6 Classifications used in model databases

The agricultural market models reviewed in this study are linked to various classifications depending on the main source of the data (see Chapter 4). While additional sources may be used to complete certain time series or product balances, in general all agricultural modelling groups use one international statistical database as their major data source. One exception to this rule might be the GTAP database, as it receives the data for its Input-Output (I-O) tables from GTAP members, various national and international organisations and a number of individuals that put together country specific I-O tables; however, GTAP does draw

additional information from international databases such as FAOSTAT and USDA. If necessary, this information is calibrated to fit trade data.

An overview of the different classifications applied in the various agricultural market models is given in Table 3.2, which shows that almost all databases use the ISO-3166 classification and the Eurostat classification on countries, respectively. The FAO based its country classification on the ISO standard, but made changes to the country codes. This applies to the database of FAOSTAT and the model database of COSIMO. The USDA makes use of the FIPS classification, which is also derived from ISO, but uses a different 2-letter-country-code.

It becomes more complex when examining the product and items classifications used in the product balances. As stated before, Eurostat product balances are not linked directly to any product classifications such as NACE, CPA or CN. Definitions exist for products and items compiled and products can only be indirectly linked via trade to the CN classification. The models that source most of their data from Eurostat therefore implement the same definitions (AGMEMOD, CAPRI/CAPSIM, ESIM). The FAO has developed its own definitions based on the compilation of balance sheets. COSIMO, as a market model developed by FAO, therefore uses the FAO definitions. Commodities that are compiled in the balance sheets are mapped to the HS code with own FAO adjustments. The AGLINK model, which shares a considerable amount of model components with the COSIMO model, also links its nomenclature codes to HS codes. For aggregation purposes, this linking is done for tariff data specified by HS code by using an Access macro that generates a concordance table; this table basically attributes the AGLINK codes to tariff lines according to a specified correspondence at the 4- or 6-digit level.

The GTAP model is different from other models in that it does not use product balances as such. Rather, it uses its own classification of products but links this directly to the CPC classification for agricultural and food products.

FAPRI sources almost all of its data from the PS&D database.

Concerning balance items, no standard classification could be found. However, almost all databases and models have a common structure of employed items: production, consumption, imports, exports and change in stocks. As an example, the AGLINK model uses this structure. This general structure might be amended or extended by additional balance items when a specific commodity or a group of commodities is compiled. This can be shown with the AGLINK model: for crops, production is harvested production. Domestic use or consumption is split between food, feed and other use (with a forthcoming additional split into biofuel use). Oilseed crush is distinct in oilseed consumption. For meats, production is gross indigenous production. Net trade for meats includes meat trade and live trade in carcass weight meat equivalent and thus we also separate slaughter production. Milk production is fluid milk production from all sources. Milk consumption is split into fluid milk and industrial use. Dairy product production and consumption are food products only.

Table 3.2: Classifications used by agricultural market models and international databases

Model	Country	Product	Items	Year
AGLINK	ISO	HS/AGLINK	AGLINK	mixed year
AGMEMOD	Eurostat	Eurostat	Eurostat	mixed year
CAPRI/CAPSIM	Eurostat	Eurostat	Eurostat	mixed year
COSIMO	ISO 3166/FAO	HS/FAO	FAO	mixed year
ESIM	Eurostat	Eurostat	Eurostat	mixed year
FAPRI	FIPS	PS&D	PS&D	mixed year
GTAP	ISO 3166	CPC/GTAP	GTAP	calendar year
FARM	ISO	FARM	FARM	mixed year
IFPRI	ISO 3166/IFPRI	HS/FAO	IFPRI	calendar year
Eurostat	Eurostat	CN/Eurostat	Eurostat	mixed year
FAOSTAT	ISO 3166/FAO	HS/FAO	FAO	calendar year
USDA – PS&D	FIPS	SITC/PS&D	PS&D	mixed year

For the time-reference period, there are just three databases that use only one defined period: GTAP, IFPRI and FAOSTAT each use calendar years. The other models and international databases apply a variety of possible combinations of calendar year, market year, crop year or financial year, which is simply marked as mixed in the table.

4 Data sources and data handling

Agricultural market modelling groups tend to prefer certain international statistical databases, since the data necessary for their modelling purposes requires certain characteristics. These requirements might be induced through the nature of the model itself or by certain agricultural policy interests.

In most cases one data source might not be sufficient to fulfil the needs of a model database, as data from one source might include, e.g. data errors, missing values, and additional sector representation, or is not enough to answer a specific research question. Therefore, a database needs to draw additional information from other sources such as national statistical offices or ministries, agricultural research institutions or key experts.

In Table 4.1 the different data sources of the reviewed agricultural market models have been compiled. GTAP data needs are quite different from those from other models; it is a general equilibrium model and thus maps the production, demand, intra-sectoral input and trade of all sectors and all countries. It does not show quantities of production, demand or trade but instead displays everything in terms of value. As trade plays a major role in the model, trade data is of eminent importance. For the GTAP model, trade data is provided by COMTRADE consolidated by one key expert. In the absence of an Input-Output (I-O) table available at the GTAP level of aggregation, supplemental data are needed to provide guidance on how to disaggregate the agricultural and food commodities from the existing I-O table into the GTAP commodities. Because information on the value of production and trade is needed across many regions, the supplemental data should come from a database with consistent commodity definitions across regions, rather than data for individual countries – in the case of GTAP, the FAO supplies such data. The FAO commodity balance database provides information on the quantity of a commodity produced, imported and exported for a given

country. Prices for various products are obtained from a variety of other sources (FAO, USDA (various yearbooks)) (Dimaranan, 2006).

Table 4.1: Comparison of the various data sources of the reviewed databases

Database	Sources
AGLINK	National statistics, FAOSTAT, USDA, Eurostat, DG Agri
AGMEMOD	Eurostat, national statistics, research institutions, experts
CAPRI/CAPSIM	Eurostat (FSS, FADN), FAOSTAT, WB, DG Agri, experts
COSIMO	FAOSTAT, CBS, FO. Licht, ICB
ESIM	Eurostat, FAO, FAPRI, USDA, national statistics
FAPRI	USDA/PS&D, OECD, FAOSTAT, Global Insight
GTAP	National I-O tables, FAOSTAT, USDA, experts
FARM	Canada national statistics, AGLINK, FAOSTAT
IFPRI	FAOSTAT, World Development Indicators, experts
Eurostat	Member states (relevant ministries, national statistics)
FAOSTAT	Member states (relevant ministries), country representatives
USDA-PS&D	National statistics, U.S. attachés, FAOSTAT, int. org., int. traders, experts

The FAO sends out annual questionnaires to all of its member countries requesting data for its statistical database. In the case of Germany, the recipient of the questionnaire is the Federal Ministry of Food, Agriculture and Consumer Protection (BMELV). If inconsistencies (such as unreliable data or unclosing balances) are detected during the process of data compilation, FAOSTAT gets back to the relevant ministry to clarify such problems. COSIMO makes use of the FAOSTAT database, but also uses data from F.O. Licht, ICBs data series and an internal dataset.

In the case of AGLINK, for OECD member countries and four non-member countries (Brazil, Argentina, China and Russia), official national statistical data, when available, are used for the historical data in the database (since 1970). These are obtained both directly from national statistical databases and publications and through questionnaire responses provided by country representatives from agriculture departments/ministries. Statistics for China and Russia are obtained from the ERS of the USDA. Additional data are collected mainly from the web-based databases of FAOSTAT and PS&D. For the EU-27 aggregate (EU-15, Hungary, Poland and the remaining 10 EU countries) data are obtained both from Eurostat and directly from the Commission. In addition, the model uses data projections that are obtained primarily from the questionnaire responses of national representatives. In some cases, and for a limited number of series, national outlook publications are used as a source of these individual country projections. For the remaining countries that are under the responsibility of the FAO, FAOSTAT is used for historical commodity balances collected directly from internal FAOSTAT databases.

The FARM model is derived from the AGLINK model and is limited to the agricultural sector in Canada. Approximately 85 % of national data is supplied by Statistics Canada, while data for international markets are taken from the AGLINK/COSIMO model.

In the EU, the supply balances for various agricultural markets (such as cereals, rice, potatoes, sugar, wine, feed stuff, meat, etc.) are compiled by the EU MS and sent to Eurostat. MS, acceding countries and candidate countries are responsible for drawing up a number of SBS and for supplying national data to Eurostat so that EU-wide SBS can be compiled. Collecting and transmitting results for all agricultural commodities is not covered by legislation, but is carried out based on gentlemen's agreements. For example, in the meat

production sector, some commodities are not covered by legislation (equidae, poultry, other meats and offal), although binding health regulations do exist. The balances supplement the information on production sectors with data on external trade in products, stocks and domestic use (Eurostat, 2005b). Table 4.2 provides an overview of SBS that are compiled by MS and Eurostat.

Table 4.2: Different approaches for compiling SBS at Eurostat

Category 1 **National balances compiled by Member States**	**Category 2** **National balances compiled by Eurostat**	**Category 3** **EU-25 balances compiled by Eurostat**
Rice (paddy, husked, milled)	Eggs	Cereals:
Sugar beet, sugar	Eggs – total, Eggs for consumption	Total cereals, soft wheat, durum wheat, barley, grain maize, triticale, rye and maslin, oats and summer cereals, sorghum
Potatoes, starch	Meat – total	Honey
Fruit and vegetables:	Beef and veal	Oilseed and protein crops
Oranges, peaches, apples, pears (fresh and processed)	Pigmeat	Soya, sunflower for fruit, oil and oilcakes
Dried fruit	Sheepmeat and goatmeat	Maize for oilcakes
Cauliflowers, tomatoes (fresh and processed)	Horsemeat	Linseed for seed, oil and oilcakes
Citrus fruit	Poultry	Cotton for seed
Wine	Other meat	Maize germ for oil
Rape for fruit, oil and oilcakes	Offal	Dried pulses:
Olives for fruit, oil and residue	Milk products	Peas, horse and broad beans, sweet lupine
Prepared fat and oils:	Whole milk (raw material)	Fresh grapes
Margarine and white products	Fresh products (excl. cream)	Stone fruits
	Drinking milk	
	Cream	
	Condensed milk	
	Powdered whole milk	
	Powdered skim milk	
	Butter (product weight)	
	Cheese	
	Processed cheese	

Source: Eurostat (2005a)

Four models source the majority of their data needs from Eurostat: AGMEMOD, CAPRI, CAPSIM and ESIM. As these models mostly analyse policy scenarios of the EU and its CAP, Eurostat is the logical choice as a statistical database. According to specific data needs, these models draw additional information from various other sources. CAPRI and CAPSIM obtain their initial data from Eurostat (Economic Accounts for Agriculture (EAA), REGIO Database). Policy data is partially acquired from DG AGRI and other sources. Expert data from MS are collected during different research projects and other services. In addition, data are collected from FAOSTAT and the World Bank (WB), as well as other model projections. ESIM sources most of its product data from Eurostat and received additional data from FAOSTAT, FAPRI, US national statistics, US statistical yearbook, the USDA, ZMP, Eurostat (AgrIS database), and the Turkish State Institute of Statistics.

The USDA uses official country statistics, reports from agricultural attachés at US embassies, data from international organisations, publications from individual countries, information from traders both inside and outside a country, and other available information. In many cases, the eventual "final" figure used by the USDA is the official figure from the country's government. In some cases, alternative sources may be regularly used because they have proven over time to be more complete, more reliable, or more timely than the governmental source. For the current season, most data are forecasts. Some of these current year forecasts also are based on countries' official releases. For example, developed countries often release timely, survey-based updates as the producing season progresses. They also release monthly trade data and some consumption and stock data during the season. The USDA uses these data, except if a sharp change has altered the situation since the date of the countries' survey. In this case, the USDA will try to account for this change in its forecasts. For most other countries, the current season is a USDA forecast based on progress reports from other sources (PS&D, 2007).

The FAPRI model database is quite unique in that it almost exclusively draws data from the USDA/PS&D database. As this statistical database has complete and consistent product balances, there is hardly any need to use information from other sources. On the other hand, the FAPRI model does not cover all countries ("world") but rather the "most important" countries producing certain agricultural commodities. Price and policy information is collected from a combination of USDA, OECD, FAO, and country-specific sources. Biofuel data are collected from a variety of sources, including the Energy Information Administrations and some private firms. Most macroeconomic data are provided by Global Insight, a private forecasting group.

The IMPACT model at IFPRI uses FAOSTAT as the primary source for its data needs. However, these data are adjusted according to the required level of aggregation and the need for global net trade balances, along with other insights from the IFPRI team of experts and other published sources. The DREAM model at IFPRI also relies on the FAOSTAT as a data source, but additional data are drawn from the World Development Indicators (WDI) of the WB.

Technical data handling

The amount of data and information collected and stored by agricultural market modelling groups is enormous and tends to grow with each extension. Data management, data handling and analysis thereby become more demanding and it is thus crucial to implement database designs and file formats that support the easy handling of combined databases.

Most data suppliers have one or more database system for their data. These systems and the data are not available to the public, to which only a selection of the data is presented. Nowadays, most data is available through the Internet, but it can also sometimes be obtained on CD-ROM or DVD. Data suppliers offer a wide range of tables to be downloaded for further research. Many of these tables have different formats, e.g. ascii-files (tab-separated, comma-separated) or MS-Excel spreadsheets. Converting data from one format into another is called parsing. Often, in addition to the data tables, further documents and information are presented. Data suppliers like Eurostat and FAO make their data available in different formats and through different user interfaces. For instance, Eurostat makes tables available in a "Tab Separated Value" format (tsv-file). Further, one can find a great deal of information regarding the data on the Eurostat website, including classifications, legal rights, etc. (Meta-information) which is generally stored separately from the data.

At the moment, the important data suppliers are working on a new XML (Extensible Markup Language) format to present data and meta-information to the public. This Statistical Data

and Metadata eXchange (SDMX) format is an initiative that aims to foster standards for the exchange of statistical information, and is sponsored by BIS, European Central Bank (ECB), Eurostat, IMF,OECD, UN and the WB (see: www.sdmx.org).

The SDMX initiative clearly underlines that merely having data is not sufficient for most purposes. In addition to the data, there is also a need for extra information such as:

- what do the data columns mean, i.e. which classifications are used;
- what is the unit of the data;
- who is the copyright holder of the data.

The questionnaire utilised in this study included questions on the technical aspects of the model databases. In particular, the following questions were of interest:

- What is the format of your database? (e.g. GAMS, MS-Excel, csv)
- How do you import the data into your model?
- Which technical output format is used for your results?

Most models store their data and model information in Excel and CSV files (AGLINK, AGMEMOD, CAPRI, CAPSIM, COSIMO, DREAM, ESIM, FAPRI, FARM, IMPACT). This is advantageous for importing, exporting and exchanging, as these formats are widely used and thus reduce technical barriers if data access is needed. GTAP uses a rather unique format for importing, exporting and storing data, namely, the Header Array File format. Still, if the need arises this format can be converted into MS-Excel, MS-Word or txt-file formats. AGLINK and FARM also use the TROLL software, which creates bn1-files. The DREAM model stores additional data in FoxPro and ascii-files.

With respect to the programming code, a number of agricultural market modelling groups have implemented their models in GAMS (General Algebraic Modelling System). This is the case for AGMEMOD, CAPRI, CAPSIM, ESIM, IMPACT. While GTAP has its own software called GEMPACK, there is also a GAMS-version of GTAP available. The AGLINK and FARM model run in MS-Excel and TROLL. FAPRI is based on MS-Excel, while the DREAM model requires MS-Access.

AGLINK

The entire AGLINK database is released through an Outlook-dedicated website in the form of a query-based data application (called OECD.stat). Downloads of individual queries (as MS-Excel tables) are possible through the web application.

The data is first collected in MS-Excel databases, with individual databases for each country/region. These databases are maintained as the primary data source. Database files are generated from the MS-Excel databases in TROLL software, which are then used for individual country standalone models. Subsequently, the individual TROLL databases are merged into a complete, comprehensive TROLL database used for generating the commodity market baseline at the world level. The model baseline output forms the Outlook database, which is available as both a TROLL database and an MS-Excel database.

AGMEMOD

The data for AGMEMOD is stored in MS-Excel files, and is either downloaded from Eurostat or inserted into the spreadsheet manually. The data files are then imported into the GAMS programme. After the model calculations are carried out, the results will be stored in GAMS

or MS-Excel files. Input and output files are also available as gdx-files, which can be navigated and handled by a data viewer that includes a graphical tool.

CAPRI

Technically speaking, collecting data is conducted either via download from official web services, which provides the possibility of automatically incorporating the data into the data preparation step, or by using the conventional manual approach. The CAPRI system also includes a recently developed User Interface. This allows the modeller to not only begin the different preparation routines and run scenarios, but to exploit the results as well. The data viewer, as a part of the user interface, provides the possibility of navigating through different data dimensions and specifying the range for export. Formats such as MS-Excel, formatted HTML-tables, GAMS tables, csv-files, text format (fixed field width format) and a relational databases file structure are all supported. Also, if data is needed quickly, the clipboard feature (cut and paste) can be used to export the data along with column and row labels. This format is accepted by most word processors, text editors, spreadsheet programs and many DBMS systems. Furthermore, the CAPRI Exploitation Tool can be used to visualise the results by drawing maps and different sorts of diagrams.

FAPRI

Supply, use, and trade data are assembled by the USDA (PS&D). Other data are assembled by commodity analysts and placed in commodity and/or country-specific databases. The downloading of PS&D data is automated. Most other data are entered manually.

IMPACT

Data is collected via the Internet and processed in MS-Excel, Stata (or other statistical packages), and GAMS. Storage primarily takes place in MS-Excel workbooks stored on local hard drives.

USDA-PS&D

The USDA uses two stages of aggregation in their PS&D database. Output from the first stage goes to a csv-file with fixed-width fields, allowing the file to be read either as csv or as fixed-width. The fixed-width aspect of the format also allows quick visual scanning to find patterns and problems. The first stage also generates files that can be used to create TS-formatted files that can be read by the old TS-View software. The second stage of aggregation passes data to the linked system using FORTRAN direct access files. For historical reasons, the second stage passes data to models through text and Lotus-formatted files that are created automatically each time the program runs (Lotus-formatted files are read by MS-Excel, so there is no need for specific Lotus software.) The Lotus-formatted files will later be replaced by csv-files for data exchange when the second stage is updated.

Importing data into the database is done using an MS-Excel Visual Basic program to move the data. The results are presented in multiple formats: a fixed-format csv-file; prn-files containing tables that can be loaded into MS-Word; csv-files containing tables that can be loaded into MS-Excel; and TS-View-formatted files.

5 Harmonisation, completion and balancing within model databases

5.1 Applied consolidation methods

Data consolidation comprises various processes such as data completion, harmonisation, and balancing. In this study, data harmonisation will be defined so as to remove

discrepancies between different data sources if a (model) database draws information from various statistical databases (product balance sheets and trade data) or diverging time series within one statistical database (supply statistics and product balance sheets). Here, e.g. various definitions of variables have to be considered. This might not only apply to product classifications, but also to regional borders, the definition of years and items considered. In contrast, balancing deals with the closing of a given balance, which ensures that the sum of all supplies equals the sum of all uses. Here, differences might occur even when a balance is provided by one distinct source. But achieving this is very much influenced by the varying model requirements, since the demand of a net-trade model differs from models covering more balance items or even the Armington approach. In the following, completion includes problems related to missing values, typing errors, breaks or jumps in time series.

In recent years, database consolidation issues like harmonisation, completion and balancing have often been discussed with reference to trade, an effect induced by ongoing international trade negotiations and their conjoined impact analysis. To explain this special focus, one has to keep in mind that the possible simulated trade effects of further trade liberalisation might differ with the database used, and additionally, that various databases have to be combined to reveal optimal information. But the same or at least similar problems are related to baseline projections ("outlooks") and impact analysis of agricultural and environmental policies in general. Although the angle is a different one, methods applied to trade data may provide insights and general approaches that are worth examining with special reference to agricultural product balances.

In World Trade Organisation (WTO) negotiations, a procedure was adopted to harmonise the calculation of trade unit values based on different databases (Drogue and Bartova, 2006). Within this procedure, a number of guidelines were established which comprise, in addition to others, the following: if the chosen specific international database (WTO Integrated Data Base (IDB)) for a series (here: tariff line) in the base period 1999-2001 contains missing data, errors, systematic biases (i.e. when its data is always lower than in a second, different database), or the calculated values are considered to not reflect the true level, an alternative method will be used. In case of the first three categories of problems, the base period 1999-2001 can be extended by up to two years, or a comparable series (here: IDB import value of a closely related tariff line), a comparable series of a near country, or a different database at a higher aggregation level can be used. Concerning the last category, however, a weighted value combining two different databases will be applied.

To overcome some of the caveats in trade statistics[3], the Centre d'Etudes Prospectives et d'Informations Internationales (CEPII) has implemented a more advanced procedure (see Drogue and Bartova, 2006) in the *Base pour l'Analyse du Commerce International* (BACI); (in English, the Database for International Trade Analysis):

- The cif rates are used to adjust import and export declarations to solve deviating import values reported as cif, and exports reported as fob. Mirror trade flow ratios and gravity-type ordinary least square (OLS) estimates on pooled data are applied to generate cif. If a strong positive relation between the ratio of mirror flows in value and those in quantities occurs in the equation, observations will be weighted for implicit cif by the inverse of the

[3] Examples are: missing information; import quantities not matching import values; inconsistencies due to different sources; exports expressed in fob price (free on board) and imports in cif price (cost of insurance and freight included); goods in transit; recording of re-exportations; actual handling of taxes; mirror flows being inconsistent; national and international definitions of territory; and use of country of origin or country of shipment.

gap between reported mirror quantities (Min(QX_{ij},QM_{ji}) / Max(QX_{ij},QM_{ji}))[4], thus providing a higher weight. Estimated cif ratios are used to compute fob import values.

- To address greater discrepancies, the quality of country declarations is assessed, which allows harmonisation to be achieved in two steps: firstly, quality indicators of import and export declarations are calculated for each country. In order to evaluate the quality of the declarations, the absolute value of the ratios of mirror flows is decomposed using a variance analysis. The error variable (absolute value of the logarithm of the ratio of mirror flows) is regressed on four sets of fixed effects (for reporter, for partner, for the 6-digit product and for years). Secondly, these OLS estimators are used to weight each observation (trade flow) by the logarithm of the sum of the two reports. Fixed estimated effects provide the marginal impact on discrepancies between flows that can be attributed to all country or sector.

A further procedure is applied by the CEPII and the UNCTAD to reconcile tariffs within MAcMapHS6v1 (see Drogue and Bartova, 2006). Here, ad valorem equivalents are computed as trade weighted averages using the median unit value of a reference group the exporter belongs to as weight. Five groups are defined as clusters of countries grouped by real gross domestic product per capita and trade openness. The use of reference group unit values compared to world unit values or national unit values reduces problems such as variations in unit value data, or impacts of prohibitive tariffs on trade flows. The FAO deals with missing trade data in a different way: missing quantities or/and values (due to either a lack of reliable sources or unavailable data) are derived by trading partners' trade returns.

There are many problems that might occur in the time series:

- *Breaks* in the series: Because of changes in the definitions the time series might have breaks that occur in a specific year.
- *Incomplete series*: Data is not provided by the data provider for one or several years.
- *Unreliable data*: Data are provided by different institutions within one country, of which one might be less reliable than the other.
- *Typing errors*: Data handling in the international statistical offices may cause national data to end up in the wrong place in the database. Sometimes data are still manually input from paper versions provided by the reporting country.
- *Definition*: Data provided by national statistical offices may differ from internationally stored data because of different definitions used. Additionally, this data required by researchers and modellers cannot always be found in one source.

Completing time series not only involves filling the missing data points, but also means correcting typing errors and breaks in the time series. Different techniques are possible to employ when completing the time series in the database,each having its own advantages and disadvantages:

- *Manually*: This is perhaps the method that gives the best results because every entry needs to be judged separately. However, there are two serious drawbacks of this method – it is very time consuming and also error prone.
- *Inter- and extrapolation*: Interpolation is a very straightforward method for completing time series. When one or more years are missing between two known years, a simple interpolation can provide the data for the missing years. If missing data is at the beginning or end of a time series, extrapolation can be used. This method has the advantage of being simple. A disadvantage of this method is that interpolation is merely the drawing of a straight line between two points.

[4] Where QX_{ij} is the exported quantity from country i to country j, and QM_{ji} is the imported quantity in country j coming from country i.

- *Simple regression*: The idea of a simple regression is to find a complete series for an explanatory variable. The correlation between this complete variable and the incomplete variable is used to complete the time series. In most cases the years are used as an explanatory variable, but it can be any other complete series that is available in the database. This method is rather straightforward and easy to use. A disadvantage is that there is no attention given to annual cycles that may exist.
- *Multiple regression*: In the case of multiple regression, more than one explanatory variable is necessary. When there are a lot of time series to be completed, it can be a lot of work to indicate which explanatory variables have to be used for which incomplete time series. Rather lengthy administrative procedures are also necessary for this method to be successful.

Literature on the abovementioned methods can be widely found in textbooks on time series and basic statistics. Preliminary results reveal widely-scattered methods applied to data completion that range from expert knowledge and simple averages to regression and time series analysis methods weighted by complex filters and complemented by external sources of information (a priori information). But in the model databases used in this project, with the exception of the COCO database (CAPRI/CAPSIM), no advanced procedures for data completion seem to have been applied. Although methods like regression analysis, averages, smoothing averages or growth rates are used, they are employed ad-hoc or on a case-by-case basis. The decision of which method to use is often assigned to the respective desk officer and, hence, the method might vary according to the respective sector or country. Missing values and outliers are a nuisance to modellers and database users, and methods are only investigated to overcome problems, thus, if a certain procedure is made available, it may easily be accepted. When it comes to the actual methods, two principle procedures can be applied: time series analysis methods and econometric techniques such as Generalised Cross Entropy (GCE), which allows the incorporation of a priori information. The time series approach is easy to implement since good software packages are available. The second approach requires more of an investment since no software or generic tools are available.

The method used in the COCO database aims to carry out both completion and consistency in one step. The concept is accomplished by the minimisation of normalised least squares under different accountancy constraints (Britz et al. 2004). The objective function minimises the sum of two relative squared errors: (1) between corrected and given data, and (2) differences between trend forecasts and given data res. fitted data. Normalisation for the errors is based on the mean of the lower and upper boundary of the error term and the corrected data of the observed data value. Normalisation was necessary and helpful to illustrate the fact that the means of the time series entering the estimation deviate considerably. Hence, normalisation leads to the minimisation of relative errors instead of absolute ones. When focussing on the completion component in COCO, it becomes clear that it is based on trend estimations and the minimisation of differences between forecasts and given data.

Looking at methods that are used in several statistical and banking organisations, it appears that two software packages for time series are used widely:

1. TRAMO/SEATS (Time series Regression with Arima Noise, Missing observations and Outliers/Signal Extraction in ARIMA Time Series) performs standard time series analysis with an autoregressive moving average and has the advantage that it can easily be used on every separate time series. Further, a strong point of TRAMO/SEATS is that it takes some annual cycles into account that might be hidden in the time series. Also, TRAMO/SEATS smoothens the time series for breaks in the series, outliers and missing observations. TRAMO/SEATS, which are linked programs, were developed by Agustin Maravall and Victor Gomez at the Bank of Spain. TRAMO provides automatic ARIMA modelling (regARIMA models), while SEATS computes the

components for seasonal adjustment. A detailed description of TRAMO/SEATS[5], its documentation and software is given in: www.bde.es/servicio/software/econome.htm. The software has fully automatic options, including options for running many series at once.

2. X-12-ARIMA[6]. This is the seasonal adjustment software produced and maintained by the US Census Bureau. X-12-Arima is an improved version of X-11-ARIMA from Statistics Canada (US Census Bureau, 1999). Features include:

    - Extensive time series modelling and model selection capabilities for linear regression models with ARIMA errors (regARIMA models).
    - Wide variety of seasonal and trend filter options.
    - Diagnostics of the quality and stability of the adjustments achieved under the options selected.
    - The ability to process many series at once.

As the FAO undertakes considerable efforts to compile data on 140 food and agricultural items for more than 200 countries, regular consolidation of trade matrices and market balances pose an important overall task. Thus, a study was conducted to investigate whether mechanical procedures could play a greater role in the compilation of statistical databases offered by the FAO (Witzke and Britz, 2005).

In this study, a two-step approach was advocated:

First, trade flows were balanced and completed by means of import and export notifications from the two reporting countries if mirror flows deviated or were missing. The first option covered the use of a simple regression on time, whereas the second option dealt with regressions on trading partner notifications, but were limited due to limited degrees of freedom. For the regressions, R^2 were calculated to choose which should be subsequently used. Instead of using the regression results directly, they were smoothed using a Hodrick-Prescott-Filter (HP)[7]. The two completed bilateral notifications were then merged with weights

[5] The National Bank of Belgium (NBB) also uses TRAMO/SEATS but discovered some limitations of the software. For example, statistical algorithms must often be integrated in completely different tasks/environments; outlier detection can be used in batch processing of many series; seasonal adjustment must be embedded in some automated production chains, like business surveys; advanced graphical interfaces should also be available for detailed analysis, while, for some unskilled employees, black-box functions integrated in MS-Excel are the preferred solution. The current implementations of TRAMO-SEATS (DOS programs or TSW) do not satisfy all those needs. Because of the abovementioned problems, the NBB has established a freely-available library (www.bnb.be/app/dqrd/index.htm) that is fully Object Oriented (OO) and written in the .NET framework (based on dll's). Within this library not only are TRAMO/SEATS available, but also X-11-ARIMA and other tools that can be of interest.

[6] Extensive documentation on X-12-ARIMA and the software can be found on the Internet: www.census.gov/srd/www/x12a/. More detailed information about "time series" and the two packages is also on the website of Catherine Hood: www.catherinechhood.net.

[7] Hodrick-Prescott-Filter is defined as:

$$\min_{\bar{y}_t} obj = wgt \cdot \sum_{1 \le t \le T} \left[\left(y_{t+1}^{HP} - y_t^{HP} \right) - \left(y_t^{HP} - y_{t-1}^{HP} \right) \right]^2 + \sum_t \left(y_t^{HP} - y_t \right)^2 ,$$

where y_t^{HP} are the filtered estimates, y_t are the Hodrick-Prescott-Filter input values and wgt (chosen as unity) is a weighting factor attached to the first component of the objective obj. Where available, the HP input values are notifications from the given country, but before the first or after the last year with any notifications, one of the

depending on a 'revealed data quality indicator', thereby reflecting the fact that reliable and accurate results require weights to decline with the likelihood of measurement errors of the information (notified or estimated). The initial (a priori) information's lack of accuracy was expressed in terms of a standard deviation and specified as reciprocal weights of these. In the next step, the resulting import and export totals are treated as fixed, thereby allowing the estimation of each country to be independent from the others.

Step two lies more in the scope of this study: The extended versions of the FAOs' SUA were consolidated, i.e. gaps in time series were closed and eventual imbalances were removed. In the case of the SUA, hard raw data, production data, and the aggregated trade flows are examined, whereas information on human consumption, feed use, processing and other utilisation are missing or imprecise. Additionally, plausibility suggests that some series e.g. stocks, are quite unstable. As in the case of trade, missing values have to be completed but a priori information is of much more diverse quality than in trade. Furthermore, a number of constraints concerning the different SUA elements have to be regarded to ensure that the balance is closed. Additionally, stock behaviour, conversion rates, calorie intake, and processing rates have to be consistent.

Originally, an overall GCE approach was set-up, which proved to be computationally infeasible (Witzke and Britz, 2005). Instead, a Bayesian Highest Posterior Density (HPD) estimator was applied together with a number of constraints catching consistency and plausibility rules. The starting point was country-wise closing balances in all years and products. A number of equations were established as constraints based on hard data and reasonable coefficients, which might replace expert knowledge in the long run[8]. Most of the constraints acted on single SUAs, but two of them link all products. These are maximal yearly changes of calorie balances for humans and livestock (Witzke and Britz, 2005).

In the chosen Bayesian approach, the parameter vector β treats model parameters as stochastic variables with an associated prior density function, PD(β), pooling prior information before any data **y** was observed. The Likelihood function, L(β|**y**), displays information obtained in association with the assumed model. Posterior density, H(β**|y**), is the result of merging prior information and data information based on certain probability rules, as well as the Bayes theorem. In this context, "parameters" to be estimated are the SUA items, where ***sua*** is:

$$\mathrm{H}(\boldsymbol{sua}|\boldsymbol{y}) \propto \mathrm{PD}(\boldsymbol{sua})\,\mathrm{L}(\boldsymbol{sua} \mid \boldsymbol{y}).$$

Constraints are defined a feasible space $\Psi(\boldsymbol{y})$ conditional on data information **y**. The likelihood function assumed a form of an indicator function with a positive constant for feasible points and zero for infeasible points (Witzke and Britz, 2005).

When a normal distribution is applied as a prior density, the logarithmic formulation indicates that at the bottom line of the objective function there will appear a sum of normalised squared deviations:

---

two regression options will provide the input values. In case of missing values within a time series, no input values will be provided. In these cases, the HP Filter provides a convenient interpolation between the next known points (Witzke and Britz, 2005, p.6).

[8] E.g. such equations were: seed use is derive by the harvested area next year times seeding rate, looses by production and imports times loss rate, production of primary products by area harvested respectively living animals times yield, derived products by input times conversion rate, stock changes link stock levels of subsequent years, a three years moving average stock to use ratio. Concerning stock levels, bounds relative to expected values for production, imports, and exports were applied.

$$LPD = -\sum_{r,p,i,t}\left[ln\left(\sigma_{r,p,i,t}\right) + 0.5ln\left(2\pi\right) + 0.5\left(\frac{sua_{r,p,i,t} - \mu_{r,p,i,t}}{\sigma_{r,p,i,t}}\right)^2\right].$$

Limited data availability induced some complications, as some of the SUA items, such as food use, required estimation to be carried out either by expertise or by mechanical procedures. Thus, different terms were part of the objective function (Witzke and Britz, 2005):

$$obj = -Standard\ term - HP\ filter\ term - Share\ term - Stock\ to\ use\ term\ .$$

The standard term followed from the approach in the abovementioned equation, and was usually applicable to input, conversion factors, production and stock changes. The penalty for any deviation of the results from the mean was the standard deviation. The HP filter term[9] corresponds to zero a priori means for second differences when raw data was missing, especially with respect to the balancing items feed, processing, food and other use. The share term[10] in the objective function introduced the three-year average of the implied shares and was usually related to feed, processing, food and other use. Further, the share term expressed, depending on the standard deviation, an imprecise expectation about the mean importance. The stock to use term[11] is used to stabilise the three-year average stock to use ratio (Witzke and Britz, 2005).

The required a priori information was derived as follows: usually raw data for the expected means of input, harvested area, herd size, conversion rates, yields and production was supposed to be the given data **y** or HP filtered trend forecasts or interpolations. The associated standard deviations were established from the standard errors of simple trends. For the usual share items feed, processing, food and other use, the level was governed by a combination of the share term and the HP filter term. As a consequence, there was no need to specify an expected mean for the level of these items, but a priori expectations and standard deviations were required for the associated shares on total market appearances (Witzke and Britz, 2005):

$$\begin{aligned}\mu_{r,p,sink\ share,t} &= sua^{Trend}_{r,p,sink\ share,t}\\ \sigma_{r,p,sink\ share,t} &= \tau_{r,p,sink\ share,t}\end{aligned},$$

where $sink \in \{Feed, Processing, Human\ consumption, Other\ use, Residual\}$, $\mu_{r,p,sink\ share,t}$ is the expected mean of a share item, $\sigma_{r,p,sink\ share,t}$ is the corresponding standard deviation, $sua^{Trend}_{r,p,sink\ share,t}$ is the trend forecast, and $\tau_{r,p,sink\ share,t}$ is the estimated standard error of the trend line.

---

[9] $$HP\ \ filter\ term = \sum_{r,p,i,t}\left(\frac{\left(sua_{r,p,i,t+1} - sua_{r,p,i,t}\right) - \left(sua_{r,p,i,t} - sua_{r,p,i,t-1}\right)}{\sigma_{r,p,i,t}}\right)^2$$

[10] $$Share\ term = \sum_{r,p,sh,t}\left(\frac{\frac{1}{3}\sum_{s=t-1}^{t+1}\frac{sua_{r,p,sh,s}}{sua_{r,p,Production,s} + sua_{r,p,Imports,s}} - \mu_{r,p,sh,t}}{\sigma_{r,p,sh,t}}\right)^2$$

[11] $$Stock\ to\ use\ term = \sum_{r,p,t}\left(\frac{sua^{avg}_{r,p,StocktoUse,t} - \mu_{r,p,StocktoUse,t}}{\sigma_{r,p,StocktoUse,t}}\right)^2$$

Expected standard deviations for the sink items were needed in the HP filter term of the objective; they were specified as a certain percentage of the sum of the standard deviations of production and imports.

Hence, country-wise, the estimation process covered all products and all years simultaneously. The deviations between final estimates and a priori information were explicitly minimised, whereas the prior information included both parts considered precise (such as trade totals from step 1) as well as imprecise (such as 'reasonable' stock-to-use ratios). Thus, the estimation procedure was quite flexible.

For data consolidation purposes, the GCE is often used. The GCE permits the researcher to define ranges for missing data values and provides a means of differentiating the reliability of various sources in the exercise (e.g. Robinson and El-Said (2000); Wieck and Britz (2002), Robilliard and Robinson (2003)). In Robinson et al. (2000) methods for updating and estimating a social accounting matrix (SAM) using cross entropy are described; this is a problem comparable to the consolidation of SBS. The flexible cross entropy approach (CE) uses all available information and also allows the incorporation of errors in variables, inequality constraints, and prior knowledge about certain parts of the matrix. In the classical approach, the SAM is updated by the RAS method, in which an iterative procedure derives a unique set of positive (normalised) multipliers that satisfies the bi-proportional condition concerning rows and columns. But this estimation problem requires a balance to be updated, as well as new row and column totals to be added. To overcome the problem of incomplete information, Robinson et al. (2000) extended a maximum entropy econometric approach applied by Golan, Judge, and Robinson (1994); they used a numerical minimisation of the entropy distance between an existing coefficient matrix and the newly estimated coefficient matrix by applying Lagrange multipliers associated with the information on row and column sums:

$$a_{i,j} = \frac{\overline{a}_{i,j} \exp(\lambda_i y_j^*)}{\sum_{i,j} \overline{a}_{i,j} \exp(\lambda_i y_j^*)},$$

where λ_i are the Lagrange multipliers associated with the information on row and column sums, y_j^* are known new row and column sums, $a_{i,j}$ is a SAM coefficient matrix constructed by dividing each cell $t_{i,j}$ in each column *j* by the column sum y_j. This expression is analogous to Bayes' Theorem, thus, the use of additional information revises an initial set of estimates, whereas an efficient estimator satisfies the "Information Conservation Principle", namely, that the estimation procedure should neither ignore any of the input information nor inject any false information (Robinson et al. 2000, p.6f). In contrast to RAS, information to be included in the CE estimation can be quite diverse, varying from an earlier SAM via some not all row and column sums, economic aggregates and inequality constraints to zero cells. If a type of CE estimation were applied, the approach allows for the fact that the new information (sum of rows and columns) is subject to measurement errors and that the initial estimates are unbalanced. So Robinson et al. (2000) extended the CE criterion and tested it by employing a Monte Carlo method indicating that the approach as well as the incorporation of a wide range of information improves the estimation of SAM parameters. An implementation based on measuring the CE distance between two probability distributions to the CE-SAM estimation is to be found in Robinson and El-Said (2000). Here, a solution to the problem of finding a new set of *A* coefficients which minimise the cross entropy distance between the prior $\overline{A}$ and the new estimated coefficient matrix based on GAMS is presented:

$$\min I = \sum_i \sum_j A_{i,j} \ln \frac{A_{i,j}}{A_{i,j}}$$

Subject to: $\sum_j A_{i,j} y_j^* = y_i^*$ and $\sum_j A_{i,j} = 1$ and $0 \le A_{i,j} \le 1$.

Other techniques developed to overcome the abovementioned problems are described in Feenstra et al. (2005).

Another time series based approach has been set-up by Gomez and Maravall (1994) with TRAMO (Time series Regression with ARIMA (noise, missing observations and outliers)) to estimate and forecast regression models with possible ARIMA errors and any sequence of missing values. Following Gomez et al. (1999), several possible approaches can be used: firstly, some version of the Kalman Filter (KF) can be applied after "skipping" the missing observations. A maximum likelihood estimation of the ARIMA parameters is then possible, and the smoothing algorithm fixed point smoother (FPS), interpolates the missing values. Furthermore, missing values can be filled with arbitrary data, and then a maximum likelihood estimation of an ARIMA model with additive outliers (AO) can be applied. When the arbitrary value set by the user and its corresponding estimated parameter are compared, the conditional expectation of the missing value given the observed data is shown. If the unknown model parameters are to be estimated by maximum likelihood, the AO and the skipping approaches will lead to different results; this is due to the fact that in the case of the AO, the arbitrary value was included. Thus, the AO approach has to be seen as an approximate approach, especially studied for stationary ARIMA models. However, an AO approach in the general non-stationary case of missing observations estimation, called the "corrected AO" approach, was later derived. Gomez at al. (1999) also conducted different simulations to assess the approaches' performance (skipping, AO, and corrected AO). They concluded that there is a brief trade-off between the skipping and AO approaches. If the number of missing observations is limited, the additive outlier approach can be easier and faster to implement, but with an increasing number of missing observations, the skipping approach begins to be outperformed by the AO approach. A point in favour of the AO approach is that algorithms for automatic model identification and automatic outlier detection may be used in the case of no missing values. A procedure which interpolates these values, identifies and corrects for several types of outliers, and also estimates intervention variable type effects (in addition to a fully automatic model identification and outlier correction procedure) is available at (http://www.modeleasy.com/tramosea.htm). A description of the basic estimation, prediction and interpolation algorithms used in the TRAMO model can be found in Gomez and Maravall (1994). An HP filter can be also be integrated in such a model (Kaiser and Maravall, without year, p. 65). In contrast, the SEATS (Signal Extraction in ARIMA Time Series) program is used to estimate unobserved components in time series following the so-called ARIMA-model-based method. Here, trend, seasonal, irregular and cyclical components are estimated and forecasted, with signal extraction techniques applied to ARIMA models (Maravell, 1987).

5.2 Actual model database consolidation

This chapter deals with the methods employed by the various agricultural market modelling groups to create complete and consistent model databases. The various steps applied in order to achieve this goal are called harmonisation, completion and balancing, which were defined in Chapter 5.1.

Almost all agricultural market models reviewed in this study draw their information from more than one statistical database (see Chapter 2.2). Additionally, data from one source might be taken from diverging statistics (production statistics and supply balance sheets). Therefore, all modelling groups have to face the decision of how to deal with differing data from several sources.

For most statistical and model databases, the harmonisation method is, despite comprehensive documentation, not explained in detail. Instead, in most cases emphasis is placed on the sources used and the model's code. As the procedure of database consolidation requires a great deal of experience dealing with agricultural statistics and the various sources where these data can be obtained, (such as international databases, national ministries and agencies, agricultural research organisations, companies, key experts), it is therefore often termed "expert knowledge".

AGLINK

In this model, cross-country data consistency is achieved in the first instance using links to a single data source in the individual country's MS-Excel databases. For example, US beef imports from Canada are taken from the Canadian database of beef exports from Canada to the US. For each commodity, data series are selected from all available data in order to conform as much as possible to a homogeneous model-wide quality standard with a corresponding representative market price. All quantities are converted to standard measures, with no ad-hoc quality adjustments to the data. Missing series may be specifically requested from national representatives, but because of the current level of commodity classification, most series are available from at least one source. These harmonisation and unit conversion steps take place within the individual country's MS-Excel files.

For data assembly, links within individual country MS-Excel files are used to fill the final series from the different sources, such as questionnaire responses, but because there is no single source of data, no automation of the process has been undertaken. Once individual country databases are completed, an automated process creates wk1-files that are then used to create TROLL software databases (bn1-files). The final merged database is assembled within TROLL using a dedicated TROLL program that merges the datasets. Within these programs there are checks for completeness, calculations of world aggregates and additional variables (i.e. percentage changes, per capita consumption) and simulations which verify that the merged dataset and model yields the same results as the individual country modules.

For data problems such as breaks in time series, unclosing balances or incorporating data from deviating sources, no unique rule is applied in the AGLINK model. However, different solutions are envisaged, for example:

- search for a different source and apply the annual growth rate to the actual series;
- make use of a residual category if the complete balance is available;
- use a linear interpolation with the previous and following years;
- use a trend;
- use the last three-year average growth.

For certain problems, the AGLINK modelling team usually tries to apply a rule that allows them to revise the series backward from the breaking point (extrapolating backward using growth rates when there is a change in levels, use of reported shares to split an aggregated series, etc.).

When data are directly incorporated into the MS-Excel database from queries of national statistics and such data deviate from questionnaire responses, the AGLINK modelling team usually inquires directly to the national representative to determine the source of the

deviation, and whether a correction is possible. In the case of quality differences, several series in the database are maintained in order to keep a running adjustment of the deviating series.

AGMEMOD

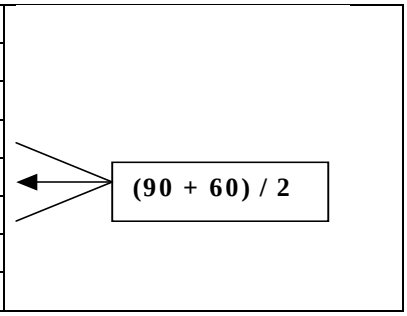
The AGMEMOD modelling group provides rather practical advice with respect to handling actual data inconsistencies or balancing problems. The overall guidelines are as follows:

- Where data is insufficient, other sources such as national statistics, data from research institutes or industry sources have to be sought.
- In the case of "unbalanced" market supply:
  - in so far as possible, data for production and domestic consumption have to be kept close to the value in the original dataset;
  - negative values in time series have to be avoided;
  - stocks will be the first to be adjusted, followed by trade variables (imports and exports), so as to maintain a coherence to data series; recent variables' true values in the original database will be adhered to as far as possible;
  - in some cases, the addition of a loss/statistical discrepancy variable can be used to ensure that the supply and use equilibrium is maintained.

In the following, some examples are given to illustrate the balancing methods applied in the AGMEMOD model (Levert and Chantreuil, 2006).

Negative values: # 1

Year	Production	Correction
1975	50	50
1976	80	80
1977	90	90
1978	- 50	75
1979	60	60
1980	110	110
1981	90	90



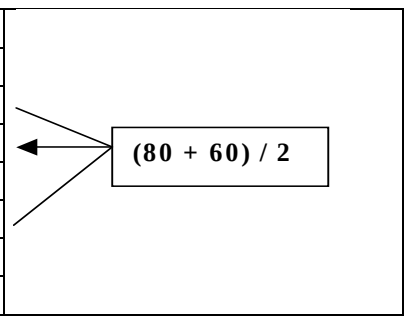
This first practical example simply shows that a negative value for a given variable should be recalculated as the average of previous and following year's data.

Negative values: # 2

By using a similar rule to that presented in the first case – though doubled – correcting a data series can be carried out.

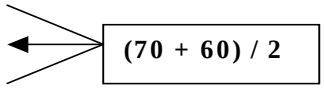
First step:

Year	Production	Correction
1975	50	50
1976	80	80
1977	- 90	70
1978	- 50	- 50
1979	60	60
1980	110	110
1981	90	90



Second step:

Year	Production	Correction
1975	50	50
1976	80	80
1977	- 90	70
1978	- 50	65
1979	60	60
1980	110	110
1981	90	90



Unbalanced data

The following example exhibits a problem with the supply and use identity in 1977.

Year	SPR	SMT	CCT(-1)	CCT	UXT	UDC	
1975	0	169	8	9	11	157	0
1976	0	263	9	14	21	237	0
1977	0	328	14	9	33	302	- 2
1978	0	359	9	4	24	333	0
1979	0	293	4	2	16	279	0

Since the data series ends in 2004, it is more practical to implement the supply and use identity, then change data for previous years.

The easiest way to proceed is to change the data of beginning stocks in 1977 so that supply and use identity holds.

The first step:

Year	SPR	SMT	CCT(-1)	CCT	UXT	UDC	
1975	0	169	8	9	11	157	0
1976	0	263	9	16	21	237	- 2
1977	0	328	16	9	33	302	0
1978	0	359	9	4	24	333	0
1979	0	293	4	2	16	279	0

Of course this induces some new unbalanced data series in 1976. Then, by implementing the same procedure in 1976 and in 1975, balanced data series can be obtain for each year.

The second step:

Year	SPR	SMT	CCT(-1)	CCT	UXT	UDC	
1975	0	169	5	7	11	157	0
1976	0	263	7	16	21	237	0
1977	0	328	16	9	33	302	0
1978	0	359	9	4	24	333	0
1979	0	293	4	2	16	279	0

CAPRI/CAPSIM

A well documented and rather complex database management approach has been developed by the CAPRI and CAPSIM modelling groups. The two groups pooled their

resources to build a new database in the context of the COCO (Complete and Consistent Database) project. COCO was primarily designed to fill gaps or to correct inconsistencies found in the statistical data and, additionally, to easily integrate data from non-Eurostat sources into the model. However, given the task of constructing consistent time series on yields, market balances, Economic Accounts for Agriculture (EAA) positions and prices for all EU MS, emphasis was placed on a transparent and uniform econometric solution to avoid manual corrections.

Three principle problems had to be solved when constructing the COCO database:

- Holes had to be filled in time series, either before the first available point, inside the range where observations are given, or beyond it.
- Some time series are missing altogether and had to be estimated, e.g. when there is data on animal production but none on meat output per head.
- Minimising corrections to given statistical data when not in line with the accounting identities.

The following principles were applied to the process of harmonisation:

- Accounting identities constrain the estimation outcome (items of the market balance sum up to zero, the difference between beginning and ending stocks equals the stock change etc.).
- Relations between aggregated time series (total cereal area) and single time series are used as restrictions in the estimation process.
- Bounds for the estimated values based on engineering knowledge or derived from first and second moments in the time series ensure plausible estimates and/or bind estimates to original data.
- As many time series as technically possible are estimated simultaneously to use the full extent of the informational content of the data constraints.

The first three points can be interpreted as a kind of "non-frequentist" approach, where additional information supplements the estimation. In the case of COCO, estimators are only of minor importance. Instead, consistent, plausible, and well-fitted values are more important for ensuring that room is left for expert knowledge.

Explicit data constraints were introduced in the estimation procedure; these involved the fitted values for each point, which were later taken as the content of the database.

The concept functions according to the following steps:

- Estimate independent trend lines for the time series;
- Estimate a Hodrick-Prescott filter using given data where available, and otherwise using the trend estimate as input;
- Define supports where there are (a) given data, (b) results from the Hodrick-Prescott filter times R^2 plus the last $(1-R^2)$ times the last known point.

The concept was carried out by minimising normalised least squares under the constraints:

$$(1.1)\quad \min_{a_i,b_i,c_i} \sum_{i,t\ if\ y_{i,t}} \left((y^*_{i,t} - y_{i,t})/y_{i,t}\right)^2 + w\sum_{i,t}\left(e_{i,t}/\left(\tfrac{1}{2}e^u_{it} - e^l_{it}\right)\right)^2$$

$$s.t.\quad (1.2)\quad a_i + b_i T + c_i T^2 + e_{i,t} = \begin{cases} y^*_{i,t}\ if\ not\ y_{i,t} \\ y_{i,t}\ if\ y_{i,t} \end{cases}$$

$$(1.3)\quad y^l_{i,t} \le y^*_{i,t} \le y^u_{i,t}$$

$$(1.4)\quad e^l_{i,t} \le e_{i,t} \le e^u_{i,t}$$

$$(1.5)\quad Accounting\ \ identities\ \ defined\ \ on\ \ y^*_{i,t}$$

where:

- Variables a,b, and c are the parameters for estimating and describing a polynomial trend fit, y is given, y^* is fitted values, and e is made up of the error terms from the estimation and T trend.
- The variable i represents the index of elements to be estimated (crop production activities or groups, herd sizes, etc.), t stands for the year, and superscripts l and u are the indices for upper and lower bounds of the estimates and errors.

The estimation process is carried out independently for each MS included in the model.

According to the different regional layers interlinked in the modelling system, it would be impossible to ensure consistency across all regional layers simultaneously. Hence, the process of building up the CAPRI database is divided into three main parts:

- A *Member State level* which integrates the EAA from Eurostat (valued output and input use) and a herd flow model for young animals. (COCO)
  (data at MS level - currently EU-27 plus Norway)
- A *regional level* which takes MS data as given and adds an input-allocation mechanism across activities and regions, as well as consistent acreages, herd sizes and yields at a regional level. Input allocation allows the calculation of region and activity-specific economic indicators as revenues, costs and gross margins per hectare or head. The regionalisation step also introduces supply-oriented CAP instruments such as premiums and quotas.
  (approximately 300 administrative regional units at NUTS 2 level for the EU-25)
- A *global level* which also takes EU MS data from the market balances as given and adds supply utilisation accounts for the other trade blocks in the market component of the model. Data on bilateral trade flows, trade policies (Most Favourite Nation Tariffs, Preferential Agreements, Tariff Rate Quotas, export subsidies) and data from domestic market support instruments (market interventions, consumption subsidies) are included.
  (16 non-EU regions broken down to 27 countries or country blocks)

For the CAPSIM model one further issue exists: the COCO database operates on a more differentiated product and activity list than CAPSIM, thus a data aggregation step is necessary to arrive at the aggregation level used in CAPSIM (Eurostat, 2003).

<u>COSIMO</u>

The FAO sends out annual questionnaires to all of its member countries requesting data for its statistical database. If inconsistencies are detected during the process of data compilation, FAOSTAT gets back to the relevant source of information to clarify such problems. Additionally, FAOSTAT uses a country-specific balance system for each commodity to ensure that total utilisation equals supply availability. In the COSIMO model, most data are

drawn from the FAOSTAT database. Therefore, data discrepancies and inconsistencies should already be reduced considerably. In the case of missing data in single years, mathematical techniques such as AR(p), log functions, and exponential growth functions are employed to deal with such problems. Historical balances are obtained outside the model, SUAs are provided by FAOSTAT. Aggregates and the report process are automated.

ESIM

Within the ESIM model, basic data adjustments are applied according to plausibility considerations and food balance sheets. If the problem of missing data occurs, two year-averages or one year data as base data will be used to calculate a three-year average. In the case of non-tradable products, residuals are calculated in the case of missing data for one item. These calculations are usually done in GAMS.

Unclosing balances do not matter much in the ESIM model since the residual is marked as foreign trade. Of course, trade data is reviewed and adjusted according to other sources and plausibility considerations if necessary.

Change in the regional coverage of states and change in the coverage of regional aggregates has not yet been an issue in ESIM. EU-intra trade is not considered in ESIM, but EU trade (as an aggregate) with third parties is (however, please note that ESIM is a net trade model).

World data includes figures for all countries of the world for which data is available.

FAPRI

Unclosing balances are rarely a problem in the FAPRI model, as it uses data from the PS&D database, which has consistent time series. In general, the PS&D product balances are complete and balanced. If necessary, the balance consumption item is adjusted by FAPRI. The particular adjusted consumption category depends on the commodity. In rare cases, an explicit "statistical discrepancy" term is added to the supply-demand balance. But, as previously stated, adjustments are needed only occasionally and are then left to the discretion of the commodity analyst.

In general, PS&D deals with most data problems in terms of supply and demand balances. Other data issues that occur in the FAPRI model are resolved by commodity analysts.

FARM

The FARM model has a fairly strong link to the AGLINK model. The Canadian component of AGLINK is a reduced form of FARM. AAFC used AGLINK to generate the world prices baseline which is passed into FARM to generate the larger, more detailed Canadian baseline. If major discrepancies exist between the initial Canadian value used in AGLINK and those generated by FARM, a further round of simulation is carried out until they are reduced to a minimum.

FARM uses TROLL software when calculations using different commodities variables are needed. Harmonisation is done at the outset when the data is stored in the xls-files; the process is automated, and consists of reading xls-files in TROLL.

Data problems are handled as follows: for breaks in time series, the model is adapted to the new series and the previous series' year-to-year percentage change is used to extrapolate the new series backwards. In the case of missing data, linear interpolation is applied. Market knowledge is applied when unreliable data appear in the database. If unclosing balances are discovered, then market experts are also required to fix such problems manually.

IMPACT/DREAM

Because the two models utilised by IFPRI are independently maintained, the applied consolidation methods vary accordingly. As described in Chapter 4, the IMPACT model uses FAOSTAT as its primary data resource. These data are adjusted according to the aggregations and the need for global net trade balances, along with other insights from IFPRI's team of experts and other published sources. The data for the base year of the model is based on a three-year average centred on the year of interest to avoid starting points that create difficulties for the long-term performance of the model. This is done by separate routines implemented outside the model. The process is an iterative, mostly manual calibration process that uses expert knowledge supplied by the modelling team.

Data problems such as missing data, breaks in time series or unclosed balances are handled through a combination of back-casting, interpolation and extrapolation, or with an "informed nearest neighbour" approach[12]. Regarding the incorporation of data from deviating sources, IMPACT uses additional methods such as expert judgement and proportion matching.

The DREAM model, being a spatial allocation model (SPAM), implements different methods for harmonisation. For example, data that is obtained from FAOSTAT is validated with other sources such as satellite imagery or agricultural census data. In general, harmonisation is done outside the model, but due to "unexpected" errors, some checking and harmonisation is done inside SPAM during data preparation, i.e. prior to the allocation calculations. After the data from FAOSTAT and WDI has reached the database (MPdb) the data assembly is done "automatically", based on parameters entered by the user (selection of product, aggregation patterns for regions), and can be repeated at any time.

Some data problems do not apply to the DREAM model (such as missing time series). Other problems are known, but no solution has thus far been found (e.g. handling intra-EU trade). Data problems such as breaks in time or unclosing balances are not addressed in a systematic way, but are rather fixed on an ad-hoc basis ("manually").

GTAP

The Center for Global Trade Analysis at Purdue University constructs the GTAP database with data that is supplied by the GTAP Network, which consists of individuals, agencies and institutions from around the world (Dimaranan, 2006).

The centrepiece of the GTAP database consists of bilateral trade, transport, and protection matrices that link all country/regional economic databases available in the GTAP model.

The trade data upon which the GTAP database is built originate from the United Nations (COMTRADE = COMmodity TRADE). The database is one of the most complete and exhaustive databases in terms of commodity and country coverage. Still, there are many problems concerning availability, quality, and consistency. The focus of the GTAP reconciliation effort is identifying and correcting inconsistencies in bilateral or counterpart trade statistics (Hertel, 1997).

In order to balance world trade, total exports of all goods and services must be equal to the total imports of all goods and services. Therefore, the difference between cif imports and fob exports must equal the value of shipping services. This is summarised as follows:

[12] A nearest neighbour is "informed", in that it can be a combination of geographic proximity, socio-economic similarities, or agro-ecological similarities depending on expert judgment.

Total merchandise imports (cif) + Nonshipping services		Total merchandise exports (fob) + Nonshipping services + Shipping services
Total imports	=	Total exports

Besides the bilateral international trade flows in GTAP, individual country Input-Output (I-O) tables are the main component of the GTAP database. Because the I-O tables that make up regional databases do not refer to a common base year (e.g. 2001), but rather to the latest available year, they must be updated to conform to the newest base year trade and macroeconomic data. GTAP accomplishes this task by using a special software package. The core data based on I-O tables comprise of the following details:

- An intermediate input matrix (firms' purchases) for domestic use of domestically produced commodities;
- An intermediate input matrix (firms' purchases) for domestic usage of imports;
- Firms' purchases of land, labour, and capital services (endowments);
- Final demands for domestically produced commodities by private households and the government, and for gross fixed capital formation;
- Final demands for imported commodities by private households and the government, and for gross fixed capital formation;
- Capital stock and endowments;
- Taxes and subsidies[13].

I-O tables provided by researchers around the world are checked at the GTAP data centre at Purdue University for structure, sectoral classification, sign, and balance. Depending on the outcome, they may be returned for further work, modified by the data centre or used unaltered.

Once this is done, the trade and regional databases may be merged. If everything has been done correctly, the database balances and the sum of all regions' savings must, by virtue of Walras' Law[14], equal global investment. This offers a final consistency check on the GTAP database (Hertel, 1997).

USDA/PS&D

With respect to harmonisation and completion, the USDA does not use a specific approach in the PS&D database such as that found e.g. in the COCO database. When discrepancies are known to exist (e.g. China's corn stocks are close to becoming negative, given the information available about other supply and use variables), the USDA "Interagency Commodity Estimates Committees" (ICECs) make current year and historical year changes to the time series based on judgements regarding the best data and information available; this results in a consistent set of supply and use data which balances. (Note: Chinese cotton does not balance because of inconsistencies in existing data but, in this case, the PS&D time series have not been adjusted.)

For grains, the USDA's World Agricultural Supply and Demand Estimates (WASDE) process adjusts World and Foreign (World-USA) consumption to balance supply and use. The data aggregation process used for modelling replicates those adjustments.

Before making projections, the Country-Commodity Linked System (CCLS) (one of the USDA's modelling systems) initialises a residual region that sets world imports equal to world

[13] Negative taxes.
[14] Walras' Law is a principle in general equilibrium theory which states that if markets for all but one good are in equilibrium, then all markets must be in equilibrium and the economy is in general equilibrium.

exports, and sets world consumption plus stock change equal to world production. However, the Residual region is not accepted as part of the official USDA data.

Generally, adjustments are made outside the models used to generate USDA projections. However, in a few limited cases (e.g. when a time series has been discontinued), we use the equations in a model to generate estimates of the most recent years of missing data. In the following, some examples are given:

Breaks in time series:

- In 2002, PS&D dropped other poultry, so broilers plus turkeys were used instead of total poultry for all years.
- For the EU, to the extent possible, the component countries are summed for all years. In recent years, intra-EU trade has been excluded from the raw PS&D data.

Missing data:

- The supply and use tables in the USDA's PS&D database nearly always provide estimates for all variables in the table, and they nearly always balance. Any exceptions would be in animal products. For some commodities and country pairs, stocks are always treated as zero, even if small stocks may be carried over from one year to another.

Unclosing balances:

- The ICEC's reconcile inconsistent (non-balancing) supply and use tables based on judgements on all available data and information, including S&U tables submitted from field offices of the USDA's Foreign Agriculture Service (FAS). Chinese cotton is a notable exception to this general practice.

Aggregation of world data:

- The CCLS modelling system has 40 country and rest-of-region models. The six rest-of-region models represent countries that are not handled by explicit country models within a given region. The commodity data in PS&D enable ERS to model total world grains, oilseeds, and cotton. For the meats, however, PS&D does not report data for a number of countries that are known to be minor producers and traders of meat products. Thus, for meats the PS&D database defines total producers and traders across all countries for which PS&D does report as "major exporters" and "major importers".

5.3 Consistency of items in commodity balance

Data consistency in a database is more difficult to establish than the completion of individual time series. The definition of *consistency* used in this report is that the interrelationship of the data holds over the four classifications (commodities, territories, items of the commodity balance, and years). For commodities, the sum of e.g. all cereals must be equal to data provided for total cereals. In the case of territories, this means for example, that for each variable in the database, the sum of all 27 individual MS of the EU must be equal to the data provided in the database for EU-27. The two examples mentioned above deal with rather simple aggregations. Establishing consistency over the items of the commodity balance is more difficult since the commodity balances should be balanced and closed.
It is not so easy to establish consistency over items in commodity balances. For most modelling databases, with the exception of the COCO database for the CAPSIM and CAPRI model, no advanced balancing procedures are employed. In general, the one decision to be made when closing the balances is whether to adjust consumption or stocks/stock changes if inconsistencies occur. Examples of currently applied advanced techniques are the HPD estimator and GCE approaches.

After completing individual time series, there is no guarantee that the commodity balances will still hold; for this reason, different consistency checks are necessary. Some of these checks are listed below:

- Total uses must be equal to total resources. To be more precise, one cannot be sure that production + import + stock change is still equal to domestic consumption + export.
- A further check would be that the sum of the individual items that are part of domestic consumption (such as losses, human consumption, animal feed, processing, industrial uses and seed) add up to the total provided in the commodity balance for domestic consumption.
- Another check is on stocks, i.e. the ending stock of year (t) must be equal to the beginning stock of year (t+1). Stock changes must be equal to the difference between beginning stock and ending stock.
- For commodity balances of processed products (such as sugar, oilseed, oilcake, etc.) the conversion factor between the amount of a basic commodity going to processing and the usable production in the balance of the processed product can be calculated and used for checking the consistency between the two linked balances. An example of this checking rule is that a conversion factor can be calculated between the amount of sugar beet going to processing and the usable production of sugar in the sugar balance.
- Finally, there are checks on indicators like per capita consumption that can be calculated by dividing the quantity used for human consumption by the population. The outcome of this calculation must be fairly stable on a year-to-year basis. Further, the result of this calculation can be used in comparison with data on per capita consumption in the commodity balances provided, but can also be used for error detection and reporting back or notifying the data providers.

5.4 Method for obtaining consistency

The previous section discussed some of the conditions that the information in the database is required to satisfy. If all those conditions are expressed as equations, a system of equations is obtained, and the objective of the consistency step is to pick a solution to that equation system. To facilitate this discussion, the desired set of database items are referred to as "variables" and represented by the vector z. The system of equations that is used to impose consistency is denoted by g(z), where consistency is obtained if, and only if, g(z) = 0.

If the number of equations in g is less than the number of elements in z, then the problem of finding z has (if any) an infinite number of solutions. Such a problem is called "underdetermined" or "ill-posed". There is not just one dataset z that is consistent, and the aim of this section is to propose a method that establishes an order of preference among all such datasets. As an aid, the researcher may have a set of observations that are assumed to be related to z by some error model, and possibly also additional types of information, like sign restrictions (e.g. prices are positive) or pure engineering information (e.g. the likely yields in agriculture). If the researcher wishes to select a solution z that is in some sense "close" to the observations and/or other prior information, then the problem is to choose the z that minimizes some criterion function v(z), which expresses this sense of "closeness", subject to the constraint that "z solves g(z) = 0". The methodological challenge is to postulate an appropriate error model and criterion function v. Below, two approaches for selecting the criterion function v are briefly discussed.

One approach to the solution of underdetermined systems of equations which deserves specific mentioning is that of Generalised Maximum Entropy (GME), extended by GCE, as introduced to a wider range of applied econometricians by Golan, Judge and Miller (1996). A typical application of GCE for solving underdetermined systems in economics is the problem of balancing a Social Accounting Matrix (SAM). Golan, Judge and Robinson (1994) propose

different entropy based methods for solving that problem, methods that have since been frequently used. With GCE as applied by Golan, Judge and Robinson (1994) (and many more), no explicit error model is formulated. Instead, each element in z is defined as the expectation of a discrete probability distribution, the supports of which are chosen by the researcher. The researcher then assigns probabilities to the individual supports of the discrete distribution in such a way that they maximise entropy (minimise cross-entropy), meaning that a distribution is preferred which is "close" to the uniform (reference) distribution. If the expectation of the assumed discrete probability distribution under the reference probabilities is y, then the combined effect of the supports, the reference distribution and the entropy function is to penalise deviations of z from y. See Preckel (2001) for a discussion of entropy estimators and penalty functions.

Another way of viewing the problem of selecting the criterion function v is described in Jansson (2007). That author interprets the problem in Bayesian terms, and proposes using the posterior density function to choose the solution z that has the highest probability density value, the HPD estimator. The same estimator is also known as "Generalised Maximum Likelihood" (DeGroot, 1970), and is also referred to by various authors as "Posterior Mode" or "Maximum A-Posteriori". An HPD estimator can be designed for the consistency problem by using the observations in conjunction with the error model to postulate a prior distribution for z. For example, if some observation is the sum of a large number of records, the central limit theorem suggests that the error in that variable is normally distributed, and in cases where only upper and lower bounds are known, a uniform distribution is appropriate. The prior distributions are confronted with the actual data and the model through the likelihood function. In this case, the likelihood function is conveniently defined as the indicator function that assigns a likelihood of "1" to all z vectors, plus error vectors that satisfy g(z) = 0 and the chosen error model, and "0" otherwise. As an example, under normally distributed priors, the HPD estimator is the vector that minimises the sum of weighted squared deviations from prior means, subject to consistency constraints g(z) = 0.

A recent study, commissioned by the FAO and carried out by EuroCARE and the University of Bonn, describes the consolidation of trade flows and market balances (Witzke and Britz, 2005). In this study, they propose ensuring the consistency of the SUAs of the FAO by using the Bayesian HPD estimator for the objective function and a number of constraints to impose consistency. In their paper, Witzke and Britz have listed all the constraints that are used to define the solution area and have described the objective function as a sum of normalised squared deviations using expected means and standard deviations for each item of the commodity balance. They go on to argue that the Bayesian HPD approach has less computational disadvantages than the GCE, which is commonly used for this kind of problem. They do, however, remark that the whole estimation for the SUAs of the FAO took 48 hours and still required some tricks.

Jansson (2007) shows that by appropriately interpreting the elements of the GCE estimator, the supports, reference distribution and entropy criterion together constitute the prior distribution (compare also with the penalty-function interpretation by Preckel (2001)). Golan et al. (1994) also recommend selecting support points based on the standard deviation of the elements to estimate. Interpreted this way, the GCE estimator is a special case of HPD estimator, where the prior distribution is restricted to a special class. Depending on the choices of prior distributions, the HPD estimator can be computationally simpler than the GCE estimator. If k is the number of support points per element of z used in GCE, and n the number of elements in z, then GCE requires (k + 1)*n variables, of which at least k*n will have non-linear derivatives (by the logarithmic entropy function); and it also introduces additional equations for adding up probabilities and definitions of expected values of the elements of z. The variable k must be at least 2, but is commonly set to higher numbers (e.g. Oude Lansink (1999) and Paris and Howitt (1998) use k = 5). If, in contrast, the priors are

assumed to be normally distributed, only n variables are required, and the first derivatives of the criterion will be linear functions. For large-scale applications, this manifold difference in number of variables can be crucial. An additional potential problem with GCE is created by the fact that the support points demark an open interval within which the estimates must lie. In large-scale applications it may be a numerical challenge to find any feasible point at all. Putting the outer supports very wide apart circumvents that problem, but since the GCE estimator then approaches a least square objective (Preckel, 2001) the added value of using the entropy criterion function in such cases can be questioned.

5.5 Relation between balance sheet and other statistics

The balance sheets play a central role in setting up consistent databases. In this report, the focus is only on these balance sheets. When expanding the database to other statistical data, it is important to have the balance sheets first be consistent. There are several relations between other statistical data sets and the possible commodity balances. A short overview is provided below:

- The “usable production” balance sheet item is the link between activities (areas, heads) and yield calculation in the production statistics. This calculation of yields can be used for error detection and reporting. This solution was recently presented by Eurostat in their Working Party Meeting in Luxemburg, which dealt with agricultural data collection and data validation.
- The import and export items on the balance sheet can be linked to trade statistics (compare Witzke and Britz, 2005). One has to be careful here because although the commodity balances in many cases already use trade statistics, they also sometimes take into account imports and exports of products that contain a certain commodity as an ingredient. For this purpose, processed products are converted into a basic commodity by using a conversion factor between product and commodity.
- A link with industrial statistics (for example, the feed industry and the processing industry) can be established through those items in the balance sheet that deal with domestic use.
- Finally, there is also a link to production values through price statistics and usable production. Prices can also be the result of a division of values through their corresponding volumes. This calculation of prices can be used in comparison with prices found in price statistics, but can also be used for error detection and reporting back or notifying data providers.

6 A fully operational MetaBase system

A database system consists of a database for storing data, and a user interface for extracting it and performing various operations on it. When a database system not only contains the data itself but also information about the data it contains, this database system becomes a so-called MetaBase system.

To build a fully operational MetaBase system, a lot of effort is required to obtain the data from the data suppliers, to get the corresponding meta-information (e.g. classifications), and finally to store all this information in a database system. Since data suppliers continuously update their data (e.g. Eurostat changes 1/3 of their 4000 tables every month) a MetaBase system should be automated as much as possible. Building software (also called Graphical User Interface or GUI) is a very complex task because a MetaBase system requires a great deal of functionality. Moreover, this functionality changes over time, e.g. because the used scientific techniques change over time.

During the course of analysing the possibilities, a prototype has been constructed to directly approach most of the foreseen challenges and to develop a potential MetaBase.

6.1 The five spheres approach

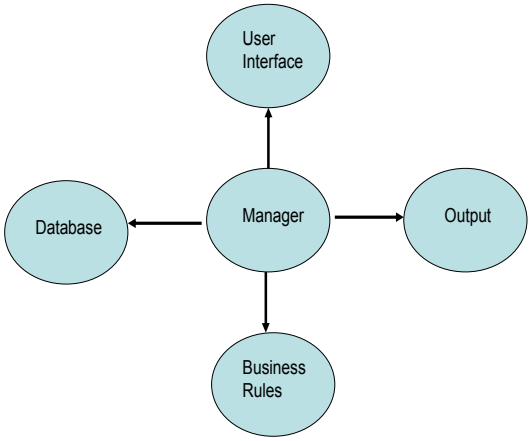
To handle complex software and database problems, the contractor has developed a standard concept which is called the five spheres approach (Figure 6.1). Central steering of a system is done by the manager, who will start a user interface, receive data from a database and apply the business rules to generate output.

In modern programming languages, the usage of objects is an accepted standard, but how one organises and structures the code is left to the programmer. Most software code is very screen- or “single problem solving”-orientated and doesn’t have a good structure. As with building research models, it turns out that a well-structured code will:

- reduce the amount of errors in the code;
- make maintenance easier;
- make extending the functionality easier;
- make sharing knowledge easier;
- make reuse of the code easier;
- save time and money.

Building applications and databases should be done in such a way that they are generic, flexible, extendable and modular. Before programming is begun, research should always be carried out to determine if some software (preferably open source) is already available that can perform certain tasks and can be incorporated into the code. Re-using software will not only speed up development time and reduce the amount of time spent on bug fixing, but the project will also benefit from the fact that other experts are building, improving and maintaining the software, and hence a much better product will be available. Re-using software becomes easier when a MetaBase is programmed with the 5 spheres approach.

Figure 6.1: Software building with the five spheres approach



The five spheres approach splits the software code into separate functional parts. Because the code is split, it becomes easier to fix bugs or to extent the functionality, but above all it becomes possible to easily adapt to large changes and to re-use code from other applications. For instance, when the five spheres approach is followed, this means that changing a Windows application into a Web-application can easily be done by only changing

the code in the User Interface portion of the five spheres. Indeed, since the whole functionality of the Business Rules and Database sphere does not change, perhaps the only requirement would be some additional Output routines. When starting the software, the Manager could even then offer, based on the specific needs at the time, either the MS-Windows interface or the Web-application, making it possible for local users to have different possibilities and a different Graphical User Interface (GUI). Another good example is the Database itself. When systems evolve over time, it is often discovered that the wrong database system was used. Switching the Database system is not a difficult task in the five spheres context. Even better, one can have several databases and database systems and the Manager will decide which one to use.

6.2 Functionality

A working MetaBase system should have the following functionality:

- The ability to convert data from different suppliers into one format.
- The possibility of presenting the suppliers' data and the harmonised data in an organised structure (a tree is the best way to present structured data).
- A multidimensional data viewer for viewing data in tables and graphs.
- A display showing available meta-information for data and data suppliers (e.g. data ownership, copyright statements, contacts, legal documents, additional documentation, etc.).
- A display of classifications used by the data tables and also the meta-information about the classification.
- A means of illustrating the relationships between the different classifications (concordances).
- The possibility of adding, changing and updating the contents of the database, i.e.
    - data sources;
    - meta-information;
    - classifications (e.g. adding new elements);
    - how the contents of the data suppliers (the tree) are presented;
    - procedures for harmonisation, completeness and consistency.
- Whenever possible, the processes should be fully automated, e.g. when a data supplier delivers new data, this data should automatically update classifications and the presentation tree, automatically convert and run the harmonisation, completeness and consistency procedures that contain the new data, and generate reports for all the data suppliers on differences in data and problems with classifications.
- For software, it should be determined whether software (freeware, shareware or commercial) that can be used already exists. It is important to make a "make or buy" decision based on the costs involved and the time needed to write one's own code. When writing tools, these tools should become freely available for others (GNU licence scheme).

6.3 Graphical User Interface (GUI)

A good user interface for a MetaBase depends on making things visible on the screen, with the database constructed in such a way that it includes many tables for screen presentation purposes only. It was also decided that no original data should be stored in the database, but that only the reference to a file would be stored. This makes it possible for data to be stored somewhere else (e.g. the data suppliers website) and above all it will keep the MetaBase

database small. Because the database is relatively small and the usage is simple, it is not essential to choose between the formats of possible database suppliers.

A good reason to use MS-Windows is that the majority of MetaBase users will be running a MS-Windows platform. But one essential part should be the possibility of showing multidimensional data. This can be data from the different data suppliers, but can also be the harmonised data. Because of the size and structure of the data, there is currently no good web-based viewer available. The LEI has developed a multidimensional DataViewer[15] that runs under MS-Windows. This viewer could be the ideal tool for allowing customers to look at multidimensional data and generate tables and graphs from it. A MetaBase system could make use of this.

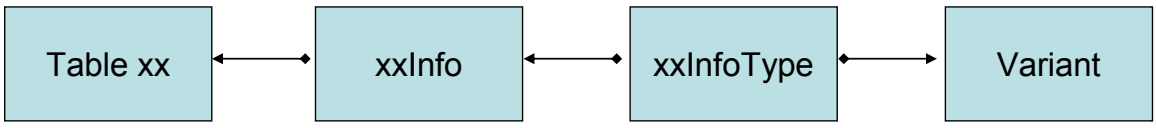
Most data suppliers have one or more database systems for their data. These systems and the data are generally not directly available to the public. Rather, only a selection of the data is presented. Most data is currently obtainable through the Internet, but also sometimes on CD-ROM or DVD. The data supplier offers a wide range of tables to be downloaded for further research. Many of these tables have different formats, e.g. ascii-files (tab-separated, comma-separated) or Excel spreadsheets. Often, not only the data tables but also additional documents and information are presented.

At the moment, the important data suppliers are working on a new XML format to present data and meta-information to the public. This Statistical Data and Metadata eXchange (SDMX) format is an initiative to foster standards for the exchange of statistical information, and is sponsored by BIS, ECB, EUROSTAT, IMF, OECD, UN and the World Bank (www.sdmx.org). For the DataViewer, it is best to store data in the GDX (Gams Data eXchange) format, in combination with a Gtree reference file (gref-file). A gdx-file is a compressed binary file and hence very dense, while the gref-file contains all meta-information needed for displaying the data (e.g. classification element labels). For converting the suppliers' data into GDX, so-called parsers are needed. The process of transforming the suppliers' data into the MetaBase format should not be complicated, and the tools used for this should be freely available. If and when SDMX becomes a de facto standard, it will be worthwhile to consider abandoning gdx/gref and switching to SDMX.

6.4 Database structure

A MetaBase is not only about showing and combining data, but also about presenting available knowledge about the data. This means that the MetaBase is created in such a way that whatever knowledge is available can be added to the system. When we have, for example, Table xx, meta-information is added to the database in three additional tables (Figure 6.2).

Figure 6.2: Adding meta-information related to a Table xx.



[15] Details on the DataViewer can be found in the documentation of Gtree: http://www.lei.dlo.nl/nacquit//downloads/gtree.doc.

When adding meta-information, it is first necessary to have the table xxInfoType, which contains a list of themes for which the meta-information is available, e.g. xxInfoType could contain a list such as: Distribution Conditions, Web-addresses, Documentation, Legal statements, Update Date, etc.

The Variant table will advise what to do with the information, e.g. it will know that a web-address is a URL, that Documentation is a Memo, the Update Date is a date, etc. Knowing this information makes it possible for a GUI to collect and present the information contents stored in table xxInfo, which contains the meta-information about entries in Table xx.

Often it is not desirable to present the contents of a table in a list, because a list can become very long, or because the content of the list is structured. A tree structure is the best way to organise lists and present them in a GUI.

When building a database for agricultural commodities, it is important to first have a clear structure of what the database should look like. Classifications play an essential role in this structure. What has been done so far is similar to a first draft. Indeed, establishing a complete and fully operational set of harmonised classifications is a huge and complex task. Further, it is important to define a clear procedure for loading the data and establishing its completeness and consistency. Additionally, it is important for database users to know what data is original (stemming from official statistical bodies) and what data is calculated, and how this is done (transparency).

To link different statistical and modelling databases into one harmonised, complete and consistent database, a unique classification system is necessary:

- For commodities, the Harmonised Commodity Description and Coding System (HS) was found to be best suited as a starting point for commodity classification.
- Most databases of the different international organisations follow the official United Nations classification for countries, which is based on the ISO-3166. Other classifications can easily be linked to this classification.
- Commodity balances are the subject of this research. The commodities themselves are already captured by the commodity classification. However, when dealing with balances, one also needs a classification for the commodity items. Since the Eurostat item classification turned out to be the most complete, it is proposed to use this classification as a foundation.
- The classification for years is uncommon because all data, and sometimes even the unit in which the data is expressed, are time-dependent. The choice was made to transform all appearances of a year into a calendar year.

6.5 Feedback to data suppliers (owners)

A MetaBase can use its information and tools to show the contents of suppliers' data, but can also show irregularities in the classifications they use. Often, it appears that a data supplier has used a well-defined classification (e.g. NACE or SITC), but while converting/parsing the data (into gdx-files) it turns out that the supplier has in fact used their own extension/redefinition of some elements in the classification. Informing the data supplier that they are not using the official classification could start an investigation over whether it is possible to supply the data according to the official classification. If that is not possible, the used/redefined elements should be added to the classification used in MetaBase and all concordances of the classification should also be updated

Whenever a data supplier provides new data, all procedures that include the new data table should run automatically, and it is possible to start a procedure that will compare the different outcomes for the various data suppliers and provide feedback to the data suppliers when there is a difference between the suppliers, or when the difference exceeds a certain threshold. How often they will receive a feedback report and also how detailed this report should be would depend on the wishes of the data supplier. Also, the format of the report should be discussed. Looking at the SDMX initiative, the next logical step would be to construct the feedback routines in such a way that they can generate SDMX.

6.6 Outlook for further developments

What follows is a list of concepts that are worthwhile to implement, and that will greatly improve the usefulness of the MetaBase system.

- Database aims:
    - possibility of adding database content (e.g. meta-information, classifications, and concordances) by making use of the GUI;
    - more classification and classification concordances;
    - extending the amount of data sources;
    - extending and improving units, dimension and their conversions;
    - more prepared selections for harmonisations;
    - adding more data from suppliers;
    - automating the updating processes for important data suppliers (such as EUROSTAT data in MetaBase).
- Software aims:
    - implementing TRAMO/SEATS and X-12 for completeness;
    - implementing an HP filter for completeness (in GAMS);
    - implementing the HPD procedure (in GAMS);
    - showing tree structures in a data viewer;
    - generating reports for data suppliers illustrating the differences between supplied and harmonised data.

A MetaBase would not only greatly benefit data suppliers in that they receive feedback on the quality of and errors in their data, but would also benefit the research community. In this project, a first prototype was built that already contains a database with metadata (links to data sources and data suppliers, classifications and linking of classification), a user-interface, parsers, procedures for harmonisation and a data viewer.

The prototype shows the potential and possible usefulness of a MetaBase system. However, the prototype does not currently fully implement a working system according to the five spheres approach in .NET, and adding some functionality in the near future is still worth considering. The system should build as much as possible on available open source software. GAMS is only used in the proposed procedures (Harmonising, Completing and Consistency of data), and when these are completed, no GAMS licences are needed for those who want to use the MetaBase system to query and show the data.

For the completion of data, TRAMO/SEATS (Time series Regression with Arima noise, Missing observations and Outliers/Signal Extraction in ARIMA Time Series) was chosen. This method comes close to a simple regression, with years being the explanatory variable. TRAMO/SEATS carries out standard time series analysis with an autoregressive moving average and has the advantage that it can easily be used on every separate time series.

Another strong point of TRAMO/SEATS is that it takes into account some annual cycles that might be hidden in the time series. Besides this, TRAMO/SEATS smoothens the time series for breaks, outliers and missing observations.

Consistency between the data in a database is more difficult to establish than the completion of individual time series. The definition of consistency in this project is that the interrelationship of data over the four classifications (commodities, territories, items of the commodity balance, and years) holds. The most suitable method applied in this study for reaching consistency seems to be the Bayesian HPD approach. In describing the consolidation of trade flows and market balances, Witzke and Britz (2005) argue that the Bayesian HPD approach has fewer computational disadvantages than the GCE which is commonly used for this kind of problem.

7 Conclusions

This study aims to provide a structure for a consolidated database for policy modelling which does not alter existing databases. Within this report, existing databases are analysed to derive key insights for setting-up a harmonised metabase. As available databases comprise statistical databases as well as scientific model databases, both groups are studied. For the purpose of this study, statistical databases are defined as providers of the information that international institutes receive from their reporters, while the reporters are required to provide harmonised, complete, consistent, and where possible, timely data series for establishing models or other quantitative methods. Nevertheless, a statistical database can also serve as a model database, such as e.g. PS&D. Statistical databases from international institutions (FAO, USDA, Eurostat), as well as model databases (AGLINK/COSIMO, AGMEMOD, CAPRI/CAPSIM, ESIM, FAPRI, GTAP, FARM, IMPACT), were studied to find ways of consolidating data and providing insights that allow for a better comparison of model results. For this reason, various classification schemes used in agricultural statistics were reviewed (country, product, balance item, year, unit), as was the manner in which the different modelling groups have dealt with these classifications in their databases.

In general, agricultural market model databases are related to a certain statistical database from an international statistical organisation. This is primarily because data from one single source considerably alleviates data management. As the main focus of modellers is to work on issues of certain agricultural, environmental, or general policy interest, the necessary data for these modelling tasks require certain databases driven by the specific model needs to vary with the model used and the policy issues analysed. Modellers thus tend to prefer unique international statistical databases, as they are also thought to be consolidated.

Most additional information is only retrieved if required data is missing or of poor quality. As this is frequently the case, modellers may apply different practices, which range from expert knowledge and alternative sources via estimation techniques, to complex mathematical procedures for overcoming these problems. Thus, the original data might be changed. However, certain routines and rules are also applied if the commodity balances do not close. These balancing procedures regularly differ from the completion procedures employed.

Due to the relation between a model's database and the respective statistical database, the underlying classification concepts of the statistical database are also of great importance for the model databases. However, model documentation rarely mentions the classifications which are explicitly or implicitly used. Instead, in most cases emphasis is placed on the sources used and the model's code. Additionally, data are mostly available only as

preformatted sets of tables and not provided as a database, or might be not easily accessible for bulk data downloads.

The issue of classification is relevant for all attributes describing statistical and projected variables, including countries/regions, products, balance items, years and units. For the first four attributes, classification is important. Units are a special attribute for which there will be no classification. To link the units, a conversion matrix with coefficients (e.g. from kg to tonnes) is needed. Furthermore, the unit can be time dependent (e.g. exchange rates) which means that a time series with conversion rates is required. Problems in comparing results appear limited when is comes to countries. In this context, deviations are mostly observed with regional aggregates such as the EU, or the world or the rest of the world. With regard to balance items, a classification based on the most detailed Eurostat list of balance items is employed. While the main concept appears to be quite consistent over sources, minor deviations frequently occur when special items are regarded. Major problems exist with respect to classifications. When we tried to set up mappings between the different classifications and databases, some concordances could only be linked by arbitrary choices. Further, in the area of SBS/SUA, no well-defined product classifications are displayed, which seems to reflect the merging of different sources into SBS/SUA (e.g. production statistics and trade statistics). Concordance between trade classifications (HS/CN code) will be used to generate links to the SBS products.

To allow for a regular comparison of results from statistical and model databases, the diverting statistical concepts, as well as the data consolidation procedures applied to the databases themselves have to be taken into account. As the classification codes are not generally documented, it is difficult to set up mappings or concordance tables for different model results. A further factor not to be neglected when updating model databases is the presence of different time scales.

Proposing a harmonised database structure for product balances which allows for the comparison of statistical databases from international statistical organisations and modelling groups requires additional steps. If no classification (explicit or implicit) of a model database is available, a classification will have to be created artifically. This classification then has to be linked and mapped to its respective international statistical database for which concordance tables should be available. To achieve this, a combined classification for each of the five variable attributes will be needed. Although a combination of all classifications may be impossible, such an approach offers the option of displaying discrepancies when classifications do not match. In principle, differences between model data and the underlying statistical databases can also be depicted.

Consolidating model databases will be difficult to achieve as long as the data in the underlying statistical databases vary due to: a) classification and consolidation issues, and b) methods of data collection. A higher degree of compliance between the statistical databases will induce a similar effect on model databases, although this might lead to adjustments in the current simulation models since the existent model parameters are estimated and calibrated to the present model databases, respectively.

Besides a common classification, a harmonised database for market modelling purposes will require further efforts to be applied to a consolidation effort for the original data. Such a procedure must be supplemented by methods dealing with completion and balancing. Preliminary results from the questionnaires and literature review reveal that widely scattered methods have thus far been applied. These efforts range from expert knowledge and simple averages to regression and time series analysis methods weighted by complex filters and a priori information. However, it appears appropriate to distinguish between a completion and a balancing process. For missing observations and outliers, the TRAMO model provided by the Bank of Spain seems to be well-accepted, but other, simpler approaches could also be

made available to the user. In the long run, a useful feature would be to also provide the procedures applied by the model databases. Introducing balancing procedures directly in the MetaBase appears to lie in the medium-term scope. In principle, the HPD estimator, as well as GCE approaches, might provide an adequate solution. But here, further research will be helpful. With regard to consolidation, data generated in this process should not deviate much from original data supplied by the respective countries, as these are well-known to national experts; larger differences could undermine the credibility of model results and the MetaBase itself.

The main conclusions of this report can be summed up as follows:

- A harmonised MetaBase must track the original data from model and statistical databases and needs to provide a comparison tool.
- This MetaBase should be made available free of charge.
- The model databases reflect the underlying statistical databases in all classification attributes, but are consolidated.
- Model results will be more comparable if the classifications are harmonised. This will require linking supply balance sheets' respective supply utilisation accounts through multidimensional mapping tables for the various classification schemes of the attributes, items and involved technical coefficients.
- Properly setting-up these mappings requires additional input, as not all questions have been satisfyingly answered.
- Comparability might be improved by completing and balancing statistical databases. Such a completion procedure (e.g. TRAMO) should also be supplied by the MetaBase.
- Concerning an advanced balancing procedure, at least two possibilities exist, both of which require further research.

References

AAFC (2003). FARM – a documentation of the food and agriculture regional model. Ottawa, Canada, 2003.

AGMEMOD (2007). URL: http://www.tnet.teagasc.ie/agmemod/

Angelini, A. (2003). Development of nomenclatures in agricultural statistics. PowerPoint presentation, 2003.

Banse M., H. Grethe and S. Nolte (2004). European simulation model (ESIM) in GAMS: model documentation.

Britz W. (ed.) (2005). CAPRI modelling system documentation.

Britz W., C. Wieck and T. Jansson (2004). National framework of the CAPRI database – the COCO module. CAPRI Working paper 02-04.

COM EU (Commission of the European Union) (2007) Prospects for Agricultural Markets and income in the European Union 2006 – 2013, Brussels.

DeGroot, M.H. (1970). Optimal statistical decisions McGraw-Hill, New York.

Dimaranan, B.V. (ed.) (2006). Global Trade, Assistance, and Production: The GTAP 6 Data Base, Centre for Global Trade Analysis, Purdue University.

Drogué, S., L. Bartova et al. (2006). A critical survey of databases on tariffs and trade available for the analysis of EU agricultural agreements. TRADEAG working paper 2006/16.

Eurostat (2002). Untitled document.

Eurostat (2003). The CAPSIM model – reference document.

Eurostat (2005a). Supply balance sheets. Meeting of the standing committee for agricultural statistics. Luxembourg, 2005.

Eurostat (2005b). Compiling supply balance sheets.

Eurostat (2006). Statistics on the trading of goods – user guide. Luxembourg 2006.

Eurostat (2007a). Database; URL: http://epp.eurostat.ec.europa.eu/portal/page?_pageid=0,1136206,0_45570464&_dad=portal&_schema=PORTAL

Eurostat (2007b). RAMON EUROSTAT metadata server; URL: http://ec.europa.eu/eurostat/ramon/index.cfm?TargetUrl=DSP_PUB_WELC

Eurostat (2007c). Information received via email, 27.02.2007.

FAO (2001). FAO – food balance sheets. A handbook. Rome, 2001.

FAOSTAT (2005). The Cosimo work programme at FAO. Committee on commodity problems, sixty-fifth session. Rome, 2005.

FAOSTAT (2007). URL: http://faostat.fao.org/

FAPRI (2006). U.S. and World Agricultural Outlook, FAPRI Staff Report 06-FSR 1, Ames January 2006.

FAPRI (2007). URL: http://www.fapri.iastate.edu/tools/outlook.aspx

Feenstra R.C., R.E. Lipsey, H. Deng, A.C. Ma and H. Mo (2005). World Trade Flows: 1962-2000.

Golan, A., G. Judge, and D. Miller (1996). Maximum Entropy Econometrics, Chichester UK: Wiley.

Golan, A., G. Judge, and S. Robinson (1994). "Recovering Information from Incomplete of Partial Multisectoral Economic Data." The Review of Economics and Statistics 76(3):541-549.

Gomez, V. and A. Maravall (1994). Estimation, Prediction, and Interpolation of Nonstationary Series with the Kalman Filter, Journal of American Statistical Association 89, pp.611-24.

Gomez, V., A. Maravall and D. Pena (1999). Missing observations in ARIMA models: Skipping approach versus additive outlier approach, Journal of Econometrics 88 (1999), p.341-63.

Hertel, T. (ed.) (1997). Global Trade Analysis: Modeling and Applications, Cambridge University Press.

IFPRI (2007a). URL: http://www.ifpri.org/themes/impact.htm

IFPRI (2007b). URL: http://www.ifpri.org/dream.htm

ISO (2007). URL: http://www.iso.org/iso/country_codes

Jansson, T. (2007). Econometric specification of parameters of constrained optimization models. Dissertation, Bonn University (forthcoming).

Kaiser, R. and A. Maravall (without year). Notes on Time Serie Analysis, ARIMA Models and Signal Extraction, Madrid.

Levert F. and F. Chantreuil (2006). AGMEMOD partnership – notes and guidelines no. 4. Building the AGMEMOD database: use of Eurostat data and common rules for a coherent database (revision 6).

Maravall, A. (1987). Minimum Mean Squared Error Estimation of the Noise in Unobserved Component Models, Journal of Business and Economic Statistics, 5, pp. 115-120.

OECD (2006). Documentation of the AGLINK-COSIMO model. Paris, 2006.

OECD database (2007). URL: http://stats.oecd.org/wbos/default.aspx

OECD/FAO (2006). OECD/FAO Agricultural Outlook 2006-2015, Paris 2006.

Oude Lansink, A. (1999). Generalised maximum entropy estimation and heterogeneous technologies. European Review of Agricultural Economics 26: 101-115.

Paris, Q., and R.E. Howitt. (1998). An analysis of ill-posed production problems using maximum entropy. American Journal of Agricultural Economics 80: 124-138.

Preckel, P.V. (2001). Least squares and entropy: A penalty function perspective. American Journal of Agricultural Economics 83: 366-377.

PS&D (2007). URL: http://www.fas.usda.gov/psdonline/psdhome.aspx

Robilliard, A. and S. Robinson (2003). Reconciling household surveys and national accounts data using a cross entropy estimation method. Review of Income and Wealth 49 (3), 395-406.

Robinson, S., A. Cattaneo and M. El-Said (2000) Updating and Estimating a Social Accounting Matrix Using Cross Entropy Methods, TMD Discussion Paper No. 58, Washington, published in: Economic Systems Research, Vol. 13, No.1, pp. 47-64, 2001.

Robinson, S., and M. El-Said (2000). GAMS code for estimating a social accounting matrix (SAM) using cross entropy methods (CE). TMD Discussion Paper No. 64. Washington, D.C.: International Food Policy Research Institute.

UN (2007). URL: http://unstats.un.org/unsd/methods/m49/m49.htm#ftn1

U.S. Census Bureau (1999), X-12-ARIMA Reference Manual, Final Version 0.2, Washington, DC: U.S. Census Bureau.

Wieck C. and W. Britz (2002). Completeness and Consistency in a Multidimensional Data Base using Constrained Simultaneous Estimation Techniques. Bonn, 2002.

Witzke, H.P. and W. Britz (2005): Consolidating trade flows and market balances globally using a Highest Posteriori Density estimator, Paper on the 8th Annual Conference on Global Economic Analysis , June 9 - 11, 2005, Luebeck, Germany.

European Commission

EUR 23417 EN – Joint Research Centre – Institute for Prospective Technological Studies
Title: Potentials of a Harmonised Database for Agricultural Market Modelling
Authors: David Verhoog, Michael Heiden, Petra Salamon, Wietse Dol and Frans Godeschalk
Editors: Stephan Hubertus Gay, Marc Müller and Federica Santuccio.
Luxembourg: Office for Official Publications of the European Communities
2008
EUR – Scientific and Technical Research series – ISSN 1018-5593
ISBN: 978-92-79-09459-0
DOI 10.2791/33791

Abstract

In a study carried out by the Joint Research Centre, Institute for Prospective Technological Studies (JRC-IPTS), the potentials of harmonising and improving databases that are currently used for agricultural market modelling have been analysed. The study supports DG AGRI in improving quality and timely availability of data for market modelling and ensuring that data from different sources are consistent. This study aims to provide a structure for a consolidated database for policy modelling which does not alter existing databases. Within this report, existing databases are analysed to derive key insights for setting-up a harmonised metabase. As available databases comprise statistical databases as well as scientific model databases, both groups are studied. Statistical databases provide information which statistical organisations received from their reporters, while the aim of scientific databases is to provide harmonised, complete and consistent data series that can be used for research and modelling. Nevertheless, a statistical database can also serve as a model database, such as e.g. PS&D. Statistical databases from international institutions (FAO, USDA, Eurostat), as well as model databases (AGLINK/COSIMO, AGMEMOD, CAPRI/CAPSIM, ESIM, FAPRI, GTAP, FARM, IMPACT), were studied to find ways of consolidating data and providing insights that allow for a better comparison of model results. For this reason, various classification schemes used in agricultural statistics were reviewed (country, product, balance item, year, unit). Apart from this, the study investigated the manner in which the different modelling groups have dealt with these classifications in their respective databases. Besides a common classification, a harmonised database for market modelling purposes will require further efforts to be applied to a consolidation effort, supplemented by methods dealing with completion and balancing.

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

JRC
EUROPEAN COMMISSION

LF-NA-23417-EN-C

ISBN 978-92-79-09459-0

Publications Office
Publications.eu.int