EMIEL CARON

Explanation of Exceptional Values in Multi-dimensional Business Databases

Explanation of Exceptional Values in Multi-dimensional Business Databases

Explanation of Exceptional Values in Multi-dimensional Business Databases

Verklaren van exceptionele waarden in multi-dimensionele bedrijfsdatabanken

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. H.A.P. Pols

and in accordance with the decision of the Doctorate Board.

The public defence shall be held on

Thursday, 14 November 2013 at 15:30 hrs

by

Emiel Antonius Maria Caron
place of birth, Dongen.

Doctoral Committee

Promotors:	Prof.dr.ir. H.A.M. Daniels
	Prof.dr. G.W.J. Hendrikse
Other members:	Prof.dr.ir. B.M. Balk
	Prof.dr. U. Kaymak
	Prof.dr. J. van Hillegersberg

Erasmus Research Institute of Management - ERIM
The joint research institute of the Rotterdam School of Management (RSM)
and the Erasmus School of Economics (ESE) at the Erasmus University Rotterdam
Internet: http://www.erim.eur.nl

ERIM Electronic Series Portal: http://hdl.handle.net/1765/1

ERIM PhD Series Research in Management, 296
ERIM reference number: EPS-2013-296-LIS
ISBN 978-90-5892-345-5

Design: B&T Ontwerp en advies www.b-en-t.nl

This publication (cover and interior) is printed by haveka.nl on recycled paper, Revive®.
The ink used is produced from renewable resources and alcohol free fountain solution.
Certifications for the paper and the printing production process: Recycle, EU Flower, FSC, ISO14001.
More info: http://www.haveka.nl/greening

Acknowledgements

Writing a dissertation is a project you do not accomplish on your own. This is a very good moment to thank some people who directly or indirectly contributed to this dissertation.

First and foremost, I would like to thank my promoters Hennie Daniels and George Hendrikse. Hennie thank you for your supervision, many suggestions, and patience. I have enjoyed our discussions and walks over the years and hope that finishing this dissertation will not be the end of those. George thank you for keeping me on track.

Secondly, I want to thank the other members of the inner committee Bert Balk, Uzay Kaymak, and Jos van Hillegersberg, for evaluating this thesis. In addition, I thank Leo Kroon and Cornelis van Bochove for being a member of the PhD committee.

Of course, a happy working environment is also important when writing a thesis. I have enjoyed being a member of the department of Decision & Information Sciences at the Rotterdam School of Management. In particular, I have enjoyed working with Jo van Nunen, Jelle van Willigen, and Ksenia Iastrebova. A special thanks goes to Albert Veenstra with whom I collaborated on the PROTECT project. During two research visits I had a very pleasant and stimulating working place at the campus of University College Dublin in Blackrock. There I collaborated with Tony Brabazon. I hope to drink some Irish beers with him again in the future. I also have enjoyed working for the Econometrics department in Rotterdam, especially for the former capacity group Economics & Informatics. I discussed on occasion my research and other topics with my colleagues there. I therefore would like to thank Uzay Kaymak, Wim Pijls, Rob Potharst, Michiel van Wezel, Eelco van Asperen, Robert Dekker, Flavius Frasincar, and Gert van der Pijl.

For my research I could use you several data sets free of charge. I thank Statistics Netherlands for letting me use the production statistics for companies, IBM Cognos

for the sales data set, and the Foundation for Tackling Vehicle Crime for the figures on stolen vehicles in The Netherlands.

Several parts of this dissertation are the result of collaboration with students, in particular students of the Economics & Informatics program. I want to acknowledge Jeroen den Heijer, Arjen Gideonse, Wim Zuiderwijk, Arno van den Berg, and Arjen Huberts. Your contributions really made a difference.

A special thanks goes to Arno van de Camp, Ad Feelders, and my paranymphs. I am very grateful to Arno van de Camp for letting me know that an interesting PhD position was available. Without informing me, my life would have been very different. I want to thank Ad Feelders for being a 'founding father' of the explanation formalism. This model is the basis for this dissertation. Furthermore, I thank my paranymphs Bart and Tjeerd for being with me on the day of defense.

Finally, I would like to thank my family and friends. Oma Jo, thank you for offering me a quiet place to study. Gonny & Frans, Rik & Diana, thank you for your support. Papa and Marieke, thank you for always being there for me. Mama, thank you for the love I still feel. It is such a sorrow that you cannot be here at my defense. Last but certainly not least, I want to thank my wife Annemarie for her unconditional love, support, and patience during my long PhD trajectory. You stood by me the whole way and I am very grateful for that. I am so proud of our sons Marijn and Maarten!

Emiel Caron
Oosterhout, September 2013

Table of Contents

Chapter 1

Introduction

1.1 Problem definition

"How can the functionality of multi-dimensional business databases be extended with diagnostic capabilities to support managerial decision-making?" This question states the main research problem addressed in this thesis. Before giving an answer, the question first requires clarification and delineation. In this chapter, the research question is placed briefly into context, both regarding academic and business relevance. This leads to the formulation of three specific research questions. Subsequently, a section is dedicated to each specific research question. An outline of this thesis concludes the chapter.

1.1.1 Business Intelligence

In management theory, the *managerial decision-making process* is often viewed as a phase model composed out of the phases: *intelligence*, *design*, and *choice* (Simon 1960). Similar phase models can be found in Emory and Niland (1968), Bonge (1972), and Mintzberg et al. (1976). In the intelligence phase, the business environment is scanned for conditions calling for decisions. During the design phase, possible courses of actions are developed and analyzed. And the choice phase concerns the selection of a specific course of action from those available. All phases of this process can be supported by the use of *business intelligence* (BI) (Turban et al. 2007). BI is an umbrella term that combines methodologies, processes, technologies, and

applications needed to transform company data into information and knowledge, that drive business decisions and actions (Raisinghani 2004). The main objective of BI is to enable access to historical and current company data, to enable manipulation of data, and to give business decision-makers the ability to analyse data that enables them to make more informed and hopefully better decisions. In this sense the term BI can also be used as the product of the transformation process in the form of the generated information and knowledge useful for decision-making.

In Figure 1.1, the conceptual architecture of the *BI framework* and its main components are depicted, based on the idea of the *enterprise information factory* (Inmon 1996). The framework describes how companies conduct and organize BI. In the framework, BI is arranged in components for (1) data production, (2) data assembly, logistics, and storage, and (3) data processing, analysis, and consumption. Starting from the left of Figure 1.1, company data flows from various operational production databases in back-end OnLine Transaction Processing (OLTP) systems, Enterprise Resource Planning (ERP) systems, and external data providers to the data warehouse. The *data warehouse* is the cornerstone of the BI framework. It is a large repository, that integrates data from several data production sources in a company, and is designed specifically to support managerial decision-making. Moreover, it is characterized as a set of *subject-oriented*, *integrated*, *time-variant*, and *nonvolatile* decision-support databases (Inmon 1996). Whereas a data warehouse combines databases across the organisation, a *data mart* is a subset of the data warehouse and focuses on a particular subject or department. Before the data can be stored into the data warehouse or data marts, the data usually needs to be assembled into a form ready for data analysis via the Extraction, Transformation, and Loading (ETL) staging area. The data in the data warehouse is finally processed by various BI front-end applications and consumed by business decision-makers, such as, financial analysts, accountants, and managers. The front-end applications allow decision-makers to access and analyze data from the data warehouse via a broad category of applications and techniques for gathering, analyzing, and providing access to data to support managers in decision-making. These BI applications include query and reporting tools, multi-dimensional or OnLine Analytical Processing (OLAP) databases[1], data

[1]The terms multi-dimensional database and OLAP database have the same meaning in this

mining and statistics, data visualization, and knowledge and business performance management systems.

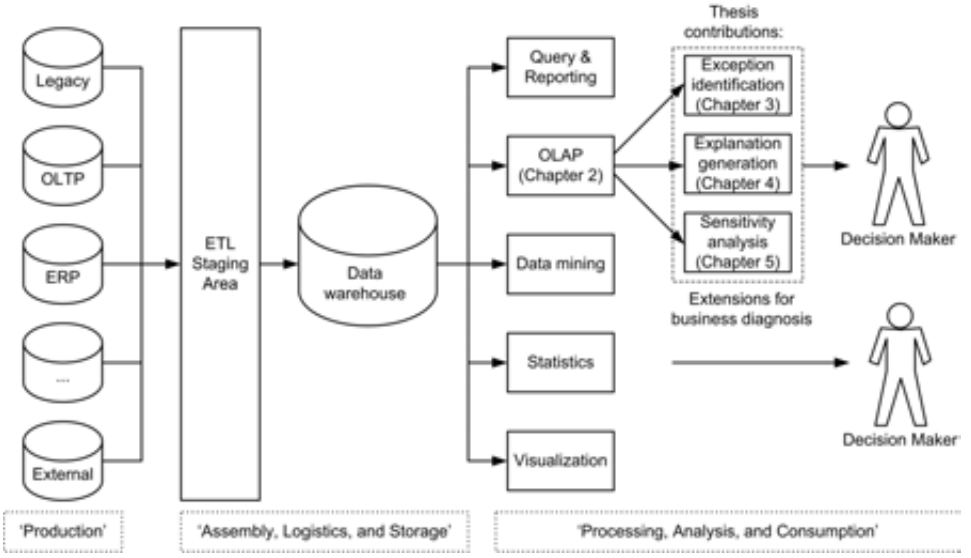


Figure 1.1: Business intelligence framework (based on Inmon (1996)).

1.1.2 Multi-dimensional database

The multi-dimensional database is an important component of the BI framework. It is used to provide business decision-makers with the ability to perform dynamic data analysis. With OLAP software technology[2], users gain access to the data warehouse. This type of access provides decision-makers with the potential to improve their understanding of business changes and their ability to identify or generate possible solutions for a variety of decision problems. Decision-makers tend to have questions that are often multi-dimensional in nature and demand fast access to large amounts of aggregated data (Kimball 1996). A typical business question might be: "What was

dissertation. In Section 2, these terms are described in detail.

[2]With the terms OLAP software (technology) or OLAP (information) system we mean the OLAP database plus the software to connect with a back-end source database and the software to analyse the OLAP database.

the profit of product A this year, in region X, per sales office, compared with the previous version of the product, compared to the targeted profit?". For decision-making purposes it might be necessary that the answer to this question is explored further, for example, on the quarter, month, and week level. This functionality is provided by OLAP technology (Thomsen 1997).

OLAP databases are a popular BI technique in the field of enterprise information systems for business analysis and decision support. The functionality of decision support systems (DSS), management information systems (MIS), and executive information systems (EIS) is combined in OLAP and extended with multi-dimensional views or data cubes, dynamic data analysis with intuitive navigational operators, and graphical data representation (Koutsoukis et al. 1999). OLAP systems support a variety of activities in business departments. Finance departments use OLAP for applications such as budgeting, activity-based costing, financial performance analysis, and financial modeling (Thomsen 2002). Sales analysis and forecasting are two of the OLAP applications found in marketing departments. Among other applications, marketing departments use OLAP for market research analysis, promotions analysis, sales forecasting, customer analysis, and market/customer segmentation. In addition, OLAP is considered more and more as an integral part of an Enterprise Resource Planning (ERP) system as can be seen in SAP's Business Information Warehouse (SAP 2006). To stress the importance of BI and OLAP products we just mention that the size of the world BI market is estimated about $10.5bn in 2010, and is still growing (Gartner 2011).

Business analysts are relying more and more on OLAP data[3] for business decision-making. However, today's OLAP databases have limited explanation or diagnostic capabilities. The diagnostic process is now carried out mainly manually by business analysts, where the analyst explores the multi-dimensional data to spot exceptions visually, and navigates the data with operators like drill-down, roll-up, slice, and dice to find the reasons for these exceptions (Han and Kamber 2005). It is obvious that human analysis can become problematic and error-prone for large data sets that commonly appear in practise. For example, a typical multi-dimensional data set has five to seven dimensions and an average of three levels hierarchy on each dimension

[3]With the term OLAP data we mean the actual business data that is stored in an OLAP database.

and aggregates more than a million records (Pendse 2006). Thus in practise, OLAP databases are often too large and have too many dimension hierarchies for analysts to browse effectively by hand. Therefore, computerized diagnosis in OLAP data, to help analysts discover the interesting parts of the OLAP database, is an important topic.

The goal of this thesis is to largely automate the current manual diagnostic discovery process in OLAP systems and to extend these systems with more powerful analysis and reporting functions. This functionality can be provided by extending the conventional OLAP system with an explanation formalism, which supports the work of human decision makers in diagnostic processes, as part of the intelligence phase in the decision-making process. Here diagnosis[4] is defined as *finding the best explanation of unexpected behaviour (i.e., symptoms or exceptions) of a system under study* (Verkooijen 1993). This definition captures two tasks that are central in problem diagnosis, namely *problem identification* and *explanation generation*. It assumes that we know which behaviour we may expect from a correctly working system, otherwise we would not be able to determine whether the actual behaviour is what we expect or not. Mintzberg et al. (1976) describe problem identification as an activity "in which opportunities, problems and crises are recognized and evoke decisional activity" and explanation generation as an activity "in which management seeks to comprehend the evoking stimuli and determine cause-effect relationships".

1.1.3 Diagnostic problem solving

The ability to generate explanations is generally considered to be an important aspect of knowledge-based systems in various application domains. Therefore, the formalization of diagnostic problem-solving or diagnostic reasoning is a subject that has been studied extensively in the field of Operations Research (OR) and Artificial Intelligence (AI) since the 1970's, and has applications in diverse domains as the medical, physical, and business and management domain. A short summary of diagnosis in the medical and physical domain is given in Appendix A. Diagnosis in the business and management domain is an important research area, where diagnostic support is often

[4]Obtained from the Greek words dia = by and gnosis = knowledge.

integrated in business information systems, like MIS and DSS, designed to support decision-making in various forms. Moreover, diagnosis occurs in a number of different business disciplines, like finance, accounting, marketing, and so forth. Typical diagnostic reasoning tasks include (Hamscher 1990): financial assessment, interfirm comparison, auditing, tax planning, and cost control. A special application of diagnosis in this domain is diagnosis integrated in a multi-dimensional database. In this thesis we focus on this relatively new application domain, because this is a critical, but rather unattended, aspect of the decision-making process of business analysts using these information systems.

The objective of the diagnostic process is to find an explanation for significant discrepancies between actual and expected system behaviour. In general, the diagnostic process is seen as a complex problem solving task with different kinds of interacting knowledge, as depicted in Figure 1.2, based on Davis and Hamscher (1988). Symptom identification is obviously necessary before the diagnostic process can be initialized to generate explanations for symptoms. This task basically requires three kinds of knowledge. Two are related to the input, and one to the interpretation of possible discrepancies (Benjamins 1993):

- the *actual model* with observations of the actual behaviour or definitions of the structure of the system to be diagnosed;
- the *normative model* with a description or a prediction of the expected behaviour of the system;
- and *domain knowledge* concerning the quality and preciseness of the observations and the expected behaviour as well as *comparison knowledge* (e.g., the type of statistical model applied, threshold values, etc.) to decide whether a discrepancy is significant.

A major activity in symptom identification is the specification of the degree of deviation from the norm. When a discrepancy between actual system behaviour and expected behaviour is discovered, and has been qualified as unacceptable with respect to some specified norm, the next step is to explain this using our "understanding" of the system. A positive decision usually results in an "explanatory path" that

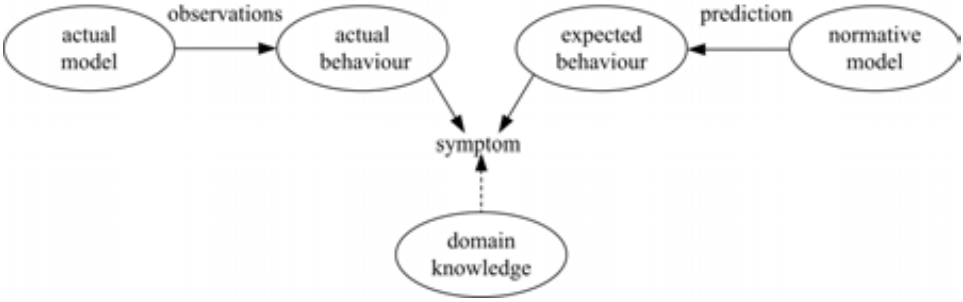


Figure 1.2: The general diagnostic task, adapted from Davis and Hamscher (1988).

leads from the observed symptoms, through the various abnormalities, to the causes. The objective of the *diagnostic process* is to find an explanation for a symptom. An explanation is a hypothesis that one or more abnormal states of the system have caused the observed discrepancies. In this work, the diagnostic process is conceived as a hypothetico-deductive process consisting of the following three consecutive tasks: problem identification, explanation generation, and explanation discrimination. This decomposition of the diagnostic process is motivated by the work of Davis and Hamscher (1988). In this thesis *sensitivity analysis* is also considered to be a part of the diagnostic process. The objective of sensitivity analysis is to determine how changes in one or more causes affect the identified symptom in some system. In our view, sensitivity analysis is considered to be the reverse of explanation generation in the diagnostic process, in the sense that in sensitivity analysis the reasoning proceeds from cause to effect.

1.1.4 Research question

Recall the main research question of this thesis:

> *How can the functionality of multi-dimensional business databases be extended with diagnostic capabilities to support managerial decision-making?*

As mentioned, business diagnosis is considered to be composed of three successive managerial decision routines or functionalities: (1) *exception identification*, (2) *explanation generation*, and (3) *sensitivity analysis*. Therefore, the main research question

breaks down into the following specific questions:

1. *How can we identify exceptional values in a multi-dimensional database?*
2. *How can we generate explanations for these exceptional values?*
3. *How can we extend the multi-dimensional database functionality with sensitivity analysis?*

Basically, the objective of our research is to propose three extensions to the OLAP database to support the business analyst in exploration of OLAP data. The OLAP database is extended with novel functionality for the detection of exceptional values, explanation generation, and sensitivity analysis. At the right side of Figure 1.1, the contributions of the thesis are depicted in the BI framework as new analytical components for OLAP databases. These components are described in Chapters 3, 4, and 5. Important in the treatment of the research questions is the development of a formal notation for multi-dimensional databases, suiting our research objective. This notation is presented in Chapter 2. In Chapter 6, we present typical business applications of the three analytical components for OLAP databases in a number of case studies. These case studies show the business relevance of answering the research questions. The analyses in the case studies are realized with prototype software applications which are developed for this purpose.

1.2 Identification of exceptional values

Business analysts who are browsing OLAP data cubes[5] are often looking for exceptions at any level in the data, because exceptions often lead to identification of problem areas or new business opportunities. This is the idea behind management by exception reporting (Judd et al. 1981). For example, chain store managers often pay special attention to areas with unusually high or low sales. Analysts from credit card companies would like to find anomalous transactions for either fraud detection or marketing reasons (Knorr and Ng 1998). Intuitively, an exceptional value in a data cube is a cell with a value significantly different from the value that is expected

[5]See Definition 2.6 for a description.

for this cell. This is a rather vague criterion that is formalized mathematically in Chapter 3. The expected behaviour in an OLAP data cube is derived from some normative model, that has been defined by business goals that have been expressed by the management. In this thesis we investigate the applicability of various types of normative models in OLAP databases. In research by Pounds (1969), it is shown that normative models are based on trends, comparable situations inside or outside the company, expectations of other people, or on theoretical models such as statistical models. A statistical model for multi-dimensional data, should estimate a cell value in the context of its position in the data cube and consider the value variation patterns over all dimensions and aggregates them relative to the cell it belongs to. Appropriate statistical models for OLAP data cubes are all kinds of ANalysis Of VAriance (ANOVA) models (Scheffé 1959; Hoaglin et al. 1988) for continuous data and models of independency for discrete category data (Bishop et al. 1975). In Chapter 3, we propose algorithms to detect exceptions automatically so that analysts could easily identify them even when the data cube is very large.

1.3 Explanation of exceptional values

Two independent surveys illustrate the need for information systems enhanced with diagnostic explanatory capabilities in the domain of business and management. Wierenga and van Bruggen (2001) evaluated 12 brand managers' experiences with existing marketing information systems and ERP systems. These brand managers were not very satisfied with the existing systems. In particular, they evaluated ERP systems negatively. One remark they made, on the business intelligence functionality of these systems, was "You only see the symptoms; you do not see the causes". In the development of a better information system the brand managers wanted a system that did not just record events but also explained them. Moreover, the authors conclude that these managers rely mainly on the problem-solving modes of reasoning and analogizing. Under these reasoning modes, they claim, one cannot determine an absolute best solution for the problem at hand. Therefore, the object of decision support is not to produce a precise recommendation on what to do, but rather to support the brand manager's decision process. For this purpose computerized diagnosis should

provide information about what is going on in the business environment and actively draw a manager's attention to specific events. In a survey of DSS users by Meador et al. (1984), it was found that needs assessment and problem diagnosis were rated as the most important factors in DSS development. This has led to the conclusion that existing techniques provide usually adequate support for problem finding but very limited support for problem diagnosis. Because this research problem only has received marginal attention in later research (Section 4.8), the problem still largely exists. For this reason, we research the possibilities here to extend OLAP systems with functionality for problem diagnosis.

Most OLAP software products rely heavily on the business analyst's intuition to manually drive the diagnostic process. Typical questions like "Why has sales increased in 2008 compared to 2009" or "Why is performance of our branch office ABC low compared to the average" can be answered by manual inspection of multi-dimensional data cubes. Such ad hoc user-driven exploration becomes complicated as data dimensionality and size increases. Moreover when it comes to an efficient in depth examination of the underlying causes of a symptom, there is still a shortage of tools to intelligently prune a large tree of causes to its essential branches. The goal of this thesis is to support the manual diagnostic discovery processes by adding explanation functionality. This can be provided by extending the conventional OLAP system with an *explanation formalism* for diagnosis of atypical values (Section 4.2.1). For this purpose, a methodology for diagnosis in the OLAP context is proposed here. The method first supports the analyst in the problem identification phase by detecting abnormal patterns in multi-dimensional data. In the subsequent explanation generation phase, the analyst is supported by returning reasons for significant drops, or increases, by generating the most important causes at lower level data. In doing so, a full explanation tree of causes at successive levels can be generated. If the tree is too large, the analyst can use appropriate filtering measures to prune the tree to a manageable size, to reduce information overload. The methodology has a wide range of applications such as interfirm comparison, analysis of sales data and the analysis of any other data that possess a multi-dimensional hierarchical structure (Chapter 4).

Hamscher (1990) and Verkooijen (1993) investigate the appropriateness of the business domain for diagnostic reasoning techniques by searching for similarities with

the physical and medical application domain. In Appendix A, we examine whether these similarities also hold for diagnosis in multi-dimensional databases. Two similarities are the presence of decomposable structures (i.e. the actual model) and behaviour prediction (i.e. the normative model). Both measures and dimensions hierarchies in the data cube have decomposable structures. An OLAP cube in the financial domain could consists of the following decomposable structures: financial statements, accounts, flows of goods and materials, market segments, etc. For example, the measures could represent a sales model M of a firm by means of quantitative equations derived from its sales database (see Table 1.1). In such quantitative financial models the dependent variables can be decomposed into its constituent independent variables in the explanation generation process. In addition, the dimensions in the data cube usually have hierarchies that specify aggregation levels. These dimension hierarchies are by definition decomposable structures. For example, $month \prec quarter \prec year$ is a hierarchy on the time dimension and $productcode \prec producttype \prec productline$ is a hierarchy on the product dimension. The measures are aggregated to various levels of detail of the combination of dimension hierarchy attributes using functions like sum and average. For example, the gross profit of some year can be decomposed into the gross profits of its constituent quarters, and the gross profit of a quarter can again be decomposed into the gross profits of its constituent months. When using the common additive aggregation function this decomposable structure is expressed as the mathematical equation: $\text{gross profit}(year) = \sum_{i=1}^{4} \text{gross profit}(quarter_i)$. When dimension hierarchies are expressed as mathematical equations, the diagnosis task resembles other more traditional diagnostic tasks that are represented by a structural model (Appendix A). In conclusion, we state that OLAP business databases have indeed decomposable structures, and the business entities described in the database structure have normative behaviours. These features suggest that the multi-dimensional database is an appropriate domain for automated reasoning and explanation techniques.

Table 1.1: Measures in a sales model M.

1. Gross Profit = Revenues − Cost of Goods
2. Revenues = Volume × Unit Price
3. Cost of Goods = Variable Cost + Indirect Cost
4. Variable Cost = Volume × Unit Cost
5. Indirect Cost = 0.3 × Variable Cost

1.4 Sensitivity analysis

Currently, OLAP business databases offer little support for sensitivity analysis or what-if analysis. Sensitivity analysis is the analysis of how changes in the output of a quantitative model can be apportioned to different sources of variation in the input of the model. In an OLAP context this naturally leads to "What if...?"-questions and scenario analysis. For example, questions of the form: "What happens to an aggregated cell value in the dimension hierarchy if I change the value of this cell value by amount X?" These types of questions are important for business analysts wanting to analyze the effect of changes in sales and costs figures on a product's profitability in a sales cube. Nowadays, multi-dimensional databases and software are rather static and have limited support to make such analyses. The OLAP analyst that wants to answer what-if questions, now has to do separate calculations in some special analysis environment (e.g. in MS Excel) or has to build SQL-queries to alter the database. In Chapter 5, we propose methods to transform the current static multi-dimensional database into a more dynamic environment, where we partly automate sensitivity analysis. The idea is to treat the OLAP database as a system of equations with respect to dimension hierarchies and relations between measures. For this purpose, we elaborate on two important mathematical conditions for sensitivity analysis in the OLAP context, namely consistency and solvability of the system of OLAP equations. We distinguish between linear systems of OLAP equations, associated with dimension hierarchies and business models, and nonlinear systems of OLAP equations, sometimes associated with business models.

1.5 Outline of the thesis

This thesis is organized as follows. In Chapter 2, we provide a general introduction to multi-dimensional business databases and present their background and context. Next we formalize the notion of the multi-dimensional database and formulate a new mathematical representation of it. This notation serves as a basis for the extensions to the OLAP framework.

In Chapter 3, we develop a framework for the identification of exceptional values. This provides an OLAP analyst the possibility to identify regions of exceptions in an OLAP data cube during navigation, representing new business opportunities or specific business problems. In addition, we elaborate on the exception identification process in the OLAP context. Here we discuss suitable classes of normative models for problem identification. A distinction is made between managerial and statistical normative models. In particular, we focus on two classes of statistical models: multi-way ANOVA models for continuous OLAP data and contingency table models for discrete OLAP data. Finally, a general algorithm for exception identification is proposed for a general OLAP cube.

Chapter 4 is the main chapter of this thesis. Here we extend the multi-dimensional model with the functionality to generate explanations for exceptional values in an OLAP data cube. We present a method that gives the OLAP analyst explanations for significant decreases or increases in business measures, identified at an aggregated level. Our method for automated diagnosis is based on a generic explanation formalism, as described in Feelders (1993) and Feelders and Daniels (2001). Explanation generation is supported by the two internal structures of the OLAP data cube: the business model and the dimension hierarchies. Therefore, we develop a multi-level explanation method for finding significant causes in these structures, based on an influence-measure which embodies a form of ceteris paribus reasoning. This method is further enhanced with a look-ahead functionality to detect hidden causes. Explanation generation is continued until a contributing cause cannot be explained further. The result of the process is an explanation tree, where the main causes for a symptom are presented to the analyst. We also propose a top-down approach for explanation in systems with both OLAP drill-down and business model equations, and a greedy

approach for explanation in systems that consist purely of drill-down equations. Furthermore, to prevent information overload, several techniques are created to prune the explanation tree. Finally, the construction of consistent chains of reference objects is discussed for various types of normative models applicable in the OLAP context.

In Chapter 5, the multi-dimensional model is extended with the functionality for sensitivity analysis. We discuss sensitivity analysis in systems that consist of purely drill-down equations and also in systems that consist of business model equations.

In Chapter 6, we show the applicability of the extended OLAP framework in a number of practical case studies. The following case studies are presented. In Case 1, computerized interfirm comparison with financial data about Dutch retail companies is discussed. In Case 2a and 2b, the top-down and greedy explanation are illustrated respectively in a case study on the analysis of multi-dimensional sales and financial data. In Case 3, the explanation method is used in a case study on the analysis of multi-dimensional vehicle crime data. Finally in Case 4, sensitivity analysis is discussed in a case study on the analysis of multi-dimensional supermarket sales data. The analyses in the various case studies are carried out with prototype software, that is described in the same chapter. Finally, in Chapter 7 we summarize the main results of this thesis.

In Appendix A, a brief overview of computer-based diagnosis is given. In Appendix B, we present the variables, data, and software, for the case study on interfirm comparison (Section 6.2). In Appendix C, we present background statistical information and data for the case study on explanation in financial OLAP data (Section 6.3). The mathematics in matrix notation to prove solvability and uniqueness of solutions of the OLAP equations are given in Appendix D. As far as we know, this has never been pointed out in the existing literature.

Chapter 2

Multi-dimensional business databases

2.1 Introduction

An important and popular front-end application for business analysis and decision support is the *OLAP* or *multi-dimensional database*. OLAP databases are capable of capturing the structure of business data in the form of multi-dimensional tables which are known as *data cubes*. Manipulation and presentation of information through interactive multi-dimensional tables and graphical displays provide important support for the business decision-maker.

Analytical data processing in OLAP databases is different from transaction data processing in OLTP databases. In the past, business data was mainly stored in the OLTP databases of transaction systems. The OLTP databases are normalized and designed using Entity-Relationship (ER) modeling. This design makes the OLTP database efficient for transaction processing but rather inefficient for managerial decision-support and complex query handling. Only recently researchers have realised the need to analyse the data and store it in a different format, the star model or snowflake model, that is utilised specifically for decision-making purposes. This research has led to a distinction between OLAP and OLTP databases. Codd (1993) and Han and Kamber (2005) provide a detailed comparison.

Moreover, OLAP databases have a strong similarity with *statistical databases*.

They both utilize the star data model and gain insight into data through fast, consistent and interactive access. However, an important difference lies in the origin of application areas. Whereas the statistical database area is mainly motivated by socio-economic databases derived from census bureaus, as for example Statistics Netherlands, which are usually the domain of statisticians, the OLAP area is driven by business applications, and their analysis for the purpose of decision-making. This is the main reason that an OLAP system is considered a component of the data warehouse. Decision-makers are not necessarily statisticians, but more typically business managers and analysts. We refer to Shoshani (1997) for a detailed overview of the similarities and differences between OLAP and statistical databases.

The remainder of this chapter is structured as follows. In the remainder of this Section, we give a short introduction to the basic concepts of the data warehouse and the multi-dimensional business database, its data model, and its implementation. In Section 2.2, we formalize the notion of the multi-dimensional database. In particular, we present a new concise mathematical notation, particular suited for combining the basic structures in the multi-dimensional database: dimensions, dimension hierarchies, cubes, cells, and measures. In Section 2.3, we elaborate on two types of equations that are present in the structure of OLAP databases: drill-down equations and business model equations. In Section 2.4, we discuss related work. Finally, we draw conclusions in Section 2.5.

2.1.1 Multi-dimensional model

The highly normalized form of the relational data model for OLTP databases is inappropriate in an OLAP database for performance reasons. Therefore, OLAP database implementations typically employ a *star model* or *star scheme* (Kimball 1996), which stores data *de-normalized* in a central fact table and associated dimension tables. This type of data model allows for fast query access because the number of table joins is heavily reduced compared to the relational model. The fact table contains linkages to the dimension tables and the actual measured data. In a star scheme, data is organized into *measures* and *dimensions*. Measures are the basic numerical units of interest for analysis and textual dimensions correspond to different perspectives for viewing measures. Dimensions are usually organized as *dimension hierarchies*, which

offer the possibility to inspect measures on different dimension hierarchy levels.

Example 2.1.1. A star model representing a multi-dimensional financial database is shown in Figure 2.1. It is taken from the case study in Section 6, and used as an illustrative example in this thesis. This database, called GoSales, contains the financial figures from a generic fictitious company that sells sports equipment, obtained from the Cognos OLAP product PowerPlay (IBM Cognos Software 2012). Figure 2.1

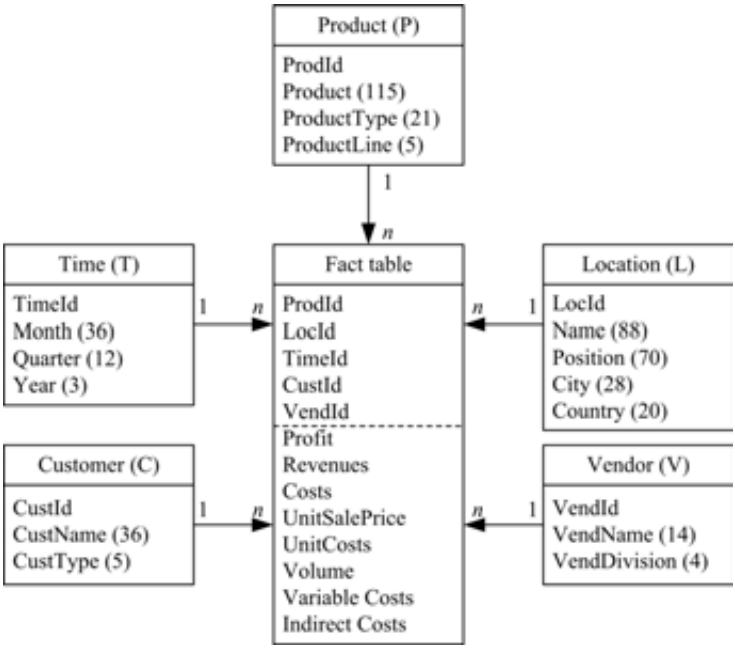


Figure 2.1: Star model with five dimension tables and a central fact table representing the financial data set.

depicts a central *fact table* and five *dimensions tables*. The central fact table represents the financial data set. It lists the measures of the data, like profit, revenues, costs, etc. The financial data set has five *dimensions tables*: Time (T), Product (P), Location (L), Customer (C), and Vendor (V), and all dimensions have a 2-4 level hierarchy.

Aggregating measures up to a certain dimension level, creates a multi-dimensional view of the data, also known as the data or OLAP cube. A data cube is not necessarily a three-dimensional geometric structure, but is essentially n-dimensional. In the upper left of Figure 2.3 on page 29, a financial data cube is shown, derived from the case study described in Section 6.3.

The star scheme's fact table has one row for each fact in the data cube. It has a column for each measure, containing the measure value for the particular fact[1]. In Table 2.1 an example fact is given for the financial fact table in Figure 2.1. As Figure 2.1 shows, a star schema has one table for each dimension and a 1-to-many (n) relationship with each dimension table. The dimension tables have corresponding key columns and one column for each dimension level, for example, Year, Quarter, Month, and TimeId. No column is needed for the top dimension level All, which will always hold the same value. The dimension table's primary key column is normally an integer identifier. In Table 2.2, a data example is presented of the dimension table Time. Moreover, the number in brackets in Figure 2.1 indicates the cardinality of that level of the dimension hierarchy. Obviously, data redundancy occurs in dimension tables. For example, because the year 2010 has 12 month values the year value 2010 is repeated 12 times in a table for the Time dimension.

Table 2.1: Example fact from the financial fact table.

ProdId	LocId	TimeId	CustId	VendId	Profit	Revenues	Costs	...
1	1	1	1	1	1295.00	3885.00	2590.00	...
...	...	...	...	...	...	...	...	...

Table 2.2: Example instance from the dimension table Time.

TimeId	Year	Quarter	Month
1	2009	Q1	January
...	...	...	...

[1] As well as a column for each dimension that contains a foreign key referencing a dimension table for the particular dimension.

2.1.2 Implementation of multi-dimensional databases

The implementation of multi-dimensional databases in OLAP software products has two basic forms (Pedersen et al. 2001; Han and Kamber 2005):

- *Relational OLAP* (ROLAP) systems use relational database structures for storing data. Such systems employ indexing methods, such as bit-mapped indexing and join indexing, to achieve good query performance.
- *Multi-dimensional OLAP* (MOLAP) systems store data in multi-dimensional database structures. Such systems contain methods for dealing with sparsity and often use indexing and hashing techniques to improve query permance.

MOLAP databases use multi-dimensional arrays as the basic data structure and implement the OLAP operators as defined in Section 2.2.3 over the arrays. MOLAP systems usually offer "more space-efficient storage as well as faster query response times (Pedersen et al. 2001)". ROLAP systems typically scale better in the number of facts they can store, are more flexible with cube redefinitions, and provide better support for frequent updates. The virtues of the two approaches are combined in the hybrid OLAP approach, which uses MOLAP technology to store higher-level summary data and ROLAP systems to store detailed data (Thomsen 1997). In this chapter we abstract from the type of implementation, in the sense that our notation can be incorporated in a ROLAP as well as a MOLAP system.

2.2 OLAP notation, concepts, and operators

2.2.1 Dimensions and dimension hierarchies

The basic unit of interest in the multi-dimensional database are numerical measures, representing countable information (Lenz and Shoshani 1997) concerning a business process. A measure can be analysed from different categorical perspectives, which are the dimensions of the multi-dimensional data. Dimensions are represented by $D_1^{i_1}, D_2^{i_2}, \ldots, D_k^{i_k}, \ldots, D_n^{i_n}$, where each domain $D_k^{i_k}$ represents a dimension k, e.g. Time, Location, Product and so on, from the associated business process. Each

dimension has a set of *dimension levels* $i_k \in \{0, 1, \ldots, \max_k\}$, e.g. the Time dimension might have the following levels: Day, Week, Month, Quarter, Season, and Year. Each domain corresponds with a dimension table in the star scheme. Furthermore, the dimension levels are organised in multiple *dimension hierarchies* or dimension paths (Vassiliadis 1998).

Definition 2.1. The domain D_k is a hierarchy $D_k^{i_k}$ partially ordered by

$$D_k^0 \prec D_k^1 \prec \ldots \prec D_k^{\max_k},$$

where D_k^0 is the lowest level and $D_k^{\max_k}$ is the highest level in D_k.

Moreover, each level in the hierarchy $D_k^{i_k}$ has an unique categoric label $A_k^{i_k}$ corresponding with a column name from the dimension table. For example, the column names in Table 2.2 on page 18 correspond to the categoric labels for the Time dimension.

The presentation of a dimension hierarchy has a *schema* component and an *instance* component (Shoshani 1997). The dimension levels and their structure as in Definition 2.1 constitute the schema, and the *dimension level instances* constitute the instances (i.e., values) for this schema. A single instance of a dimension level $D_k^{i_k}$ is denoted by $d_k^{i_k}$, where $d_k^{i_k} \in D_k^{i_k}$. The total number of instances in $D_k^{i_k}$ is denoted by $|D_k^{i_k}|$.

Example 2.2.1. For the Time dimension $D_k = T$ we have the following labelled hierarchy schema: T[Month]$\prec$ T[Quarter] $\prec$ T[Year] $\prec$ T[All-Times] or in short $\mathrm{T}^0 \prec \mathrm{T}^1 \prec \mathrm{T}^2 \prec \mathrm{T}^3$, where the level instances at level 0 are $\mathrm{T}^0 = \{$2009.Q1.Jan, 2009.Q1.Feb, 2009.Q1.Mar, $\ldots\}$, at level 1 are $\mathrm{T}^1 = \{$2009.Q1, 2009.Q2, 2009.Q3, 2009.Q4, $\ldots\}$, at level 2 are $\mathrm{T}^2 = \{2009, 2010, 2011\}$, and $\mathrm{T}^3 = \{\textit{All-Times}\}$. Here we use the dot-notation as formulated in Definition 2.4, to indicate instances of the dimension hierarchy. An example of the instantiated dimension hierarchy is 2009.Q1.Jan $\prec$ 2009.Q1 $\prec$ 2009 $\prec$ *All-Times*, where 2009.Q1.Jan $\in T^0$, 2009.Q1 $\in T^1$, 2009 $\in T^2$, and *All-Times* $\in T^3$.

In addition, the top level of a dimension always has a single level instance $D_k^{\max_k} = \{\text{All-}D_k\}$, thus $|D_k^{\max_k}| = 1$, since analysis requires that measure instances that are bound to different level instances, must be aggregated up to a single value. The schema representation belonging to the hierarchy of the Time dimension is depicted

at the left hand side of Figure 2.2. Underlying the schema representation the OLAP system stores and maintains the instances and their relationships, called the representation of the instances. This representation is depicted as a tree at the right hand side of Figure 2.2.

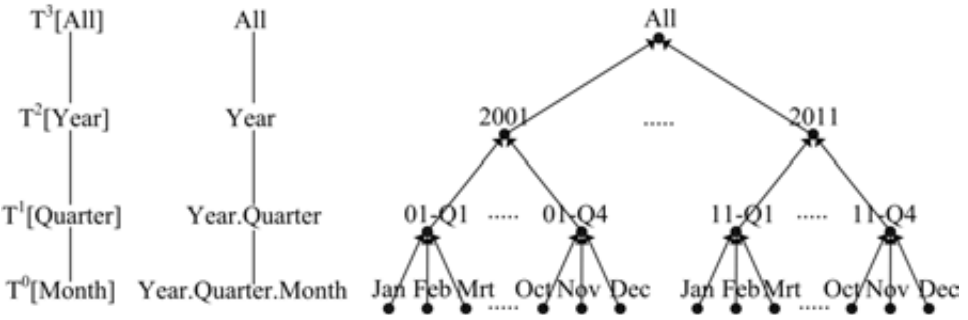


Figure 2.2: The left hand side represents the hierarchy schema of the Time dimension; the middle figure represent its schema dot-notation; the right hand side represents the rooted tree of its dimension hierarchy instances.

With each dimension hierarchy in domain D_k a *rooted tree* $T(D_k) = (V, E)$ is associated, called the *dimension hierarchy tree* of D_k, as follows. The vertex set $V(D_k)$ contains as elements all *dimension level instances* appearing in Definition 2.1. Suppose that $D_k^{i_k} \prec D_k^{i_k+1}$ is part of the dimension hierarchy, and furthermore suppose that $d_k^{i_k+1} \in D_k^{i_k+1}$ and $d_k^{i_k} \in D_k^{i_k}$, such that $d_k^{i_k} \prec d_k^{i_k+1}$. The edge set $E(D_k)$ contains a directed edge from vertex $d_k^{i_k}$ to vertex $d_k^{i_k+1}$. The instance element $d_k^{i_k+1}$ is called a *parent* and $d_k^{i_k}$ is called its *child*. The instance *All* at level $D_k^{\max_k}$ is the *root* of the tree and the instances at level D_k^0 are the *leaves* of the tree on the lowest level. An example tree of the hierarchy T^0[Month] $\prec$ T^1[Quarter] $\prec$ T^2[Year] $\prec$ T^3[All-Times] of the Time dimension is depicted at the right side of Figure 2.2. For example, in the tree the year 2009 is the parent of the children {2009.Q1, 2009.Q2, 2009.Q3, 2009.Q4}. We define an operator to determine the parent of some child element in the hierarchy of a single domain $D_k^{i_k}$.

Definition 2.2. A *1-dimensional roll-up operator* is defined as

$$r^{+1}(D_k^{i_k}) = D_k^{i_k+1}.$$

Reversely, we define an operator to determine the children of some parent element in the dimension hierarchy.

Definition 2.3. A *1-dimensional drill-down operator* is defined as

$$r^{-1}(D_k^{i_k}) = D_k^{i_k-1}.$$

These operators r^{+1} and r^{-1} can also be applied on any subset $X_k^{i_k}$ of $D_k^{i_k}$, and the operators can be applied both on the schema as on the instance level. For example, on the schema level as $r^{-1}(T^2[\text{Year}]) = T^1[\text{Quarter}]$, or on the instance level as $r^{-1}(2009)$, to determine the quarters of some specific year.

A hierarchy schema structure in dot-notation is associated with each domain $D_k^{i_k}$ on some level i_k in the hierarchy.

Definition 2.4. A *hierarchy schema structure in dot-notation* is defined as

$$D_k^{i_k} = A_k^{\max_k}.\cdots.A_k^{i_k+1}.A_k^{i_k},$$

where $A_k^{i_k}$, $A_k^{i_k+1}$, ..., $A_k^{\max_k}$ are column names from the associated dimension table. This structure represents the *ancestry* $A_k^{\max_k}.\cdots.A_k^{i_k+1}$ of descendant $A_k^{i_k}$. For example, the domain T^0 has the dot hierarchy structure Year.Quarter.Month, and the domain T^1 has the dot hierarchy structure Year.Quarter. In the middle of Figure 2.2, the dot-notation for the Time dimension is represented.

Similarly, a hierarchy instance structure in dot-notation is associated with each instance $d_k^{i_k} \in D_k^{i_k}$.

Definition 2.5. A *hierarchy instance structure in dot-notation* is defined as

$$d_k^{i_k} = a_k^{\max_k}.\cdots.a_k^{i_k+1}.a_k^{i_k},$$

where $a_k^{i_k} \in A_k^{i_k}$, $a_k^{i_k+1} \in A_k^{i_k+1}$, ..., $a_k^{\max_k} \in A_k^{\max_k}$ are column entries from the associated dimension table.

An alternative way of representing Definition 2.5 is $d_k^{i_k+1}.a_k^{i_k}$, where $d_k^{i_k+1}$ are the ancestors of $a_k^{i_k}$. For example, $d_k^0 = 2009.\text{Q1.January}$ is a dot instance representation of Year.Quarter.Month from T^0, where $d_k^2 = 2009$ and $d_k^1 = 2009.\text{Q1}$ are ancestors of 2009.Q1.January.

2.2.2 Cubes and cells

The key structure in the multi-dimensional database is the data cube.

Definition 2.6. A *cube* C is defined as the Cartesian product of subsets of available domains

$$C = X_1^{i_1} \times X_2^{i_2} \times \ldots \times X_n^{i_n}, \text{ where } X_k^{i_k} \subseteq D_k^{i_k}.$$

For example, $C = \{2010,\ 2011\}^2 \times \{\text{Germany}\}^3 \times \text{Product}^2$ is an example of a cube. Additionally, an alternative database representation of a cube is given by $(X_1^{i_1}, X_2^{i_2}, \ldots, X_n^{i_n})$. For example, the alternative notation for the cube in the upper left of Figure 2.3 on page 29 is given by (Year, Country, ProductLine).

A *full cube* C_F is defined as a special cube, where the cube contains all elements of its associated domains on some level, i.e. $X_k^{i_k} = D_k^{i_k}$. The full cube is specified on the schema level and is given by $C_F = D_1^{i_1} \times D_2^{i_2} \times \ldots \times D_n^{i_n}$, or alternatively by $(D_1^{i_1}, D_2^{i_2}, \ldots, D_n^{i_n})$, or $[\mathbf{i}] = [i_1, i_2, \ldots, i_n]$ in shorthand notation. For example, $\text{Time}^2 \times \text{Location}^3 \times \text{Product}^2$, $T^2 \times L^3 \times P^2$, or $[2, 3, 2]$ in shorthand, and so on, are full cubes in the example of Figure 2.3. Notice that according to this definition also a single dimension hierarchy is composed out of full cubes, e.g. the left hand side of Figure 2.2 shows the full cubes that make up the Time dimension.

A cube C is composed out of one or more cells.

Definition 2.7. A *cell* c is defined as an instance element of a cube C

$$c = (d_1^{i_1}, d_2^{i_2}, \ldots, d_n^{i_n}),$$

where $d_1^{i_1} \in X_1^{i_1}$, $d_2^{i_2} \in X_2^{i_2}$, ..., $d_n^{i_n} \in X_n^{i_n}$.

For example, (2006, United States, Golf Equipment) is a cell in the upper left cube $\text{Time}^2 \times \text{Location}^3 \times \text{Product}^2$ of Figure 2.3. The total number of cells in a cube C is $|C| = |X_1^{i_1}| \times |X_2^{i_2}| \times \ldots \times |X_n^{i_n}|$.

The instances at the lowest dimension levels of each of its domains are cells of a special cube, called the *base cube* $C_B = X_1^0 \times X_2^0 \times \ldots \times X_n^0 = [0, 0, \ldots, 0]$. For example, in the financial database described in Example 2.1.1 the full base cube is represented by $\text{Time}^0 \times \text{Location}^0 \times \text{Product}^0 \times \text{Customer}^0 \times \text{Vendor}^0$ or alternatively as (Month, Product, Name, CustName, VendName). The base cube

can be aggregated to higher hierarchical levels. When all dimension hierarchies are aggregated to the highest level, we derive the 0-dimensional *apex* or *top cube* $C_T = X_1^{\max_1} \times X_2^{\max_2} \times \ldots \times X_n^{\max_n} = [\max_1, \max_2, \ldots, \max_n]$. The top cube consists of only one cell $(All, All, \ldots, All)$. Notice that aggregating a dimension hierarchy to $D_n^{\max_n}$ is similar to removing it from the cube.

2.2.3 Navigational operators

With *navigational operations* the business analyst can manual explore OLAP cubes, allowing interactive querying and analysis of the data. Usually, a large number of records is stored in the fact table. Therefore, operations exist to materialize different views on the data and summarize measures in meaningful ways. By applying suitable operators, the level of detail is altered and lower level cubes are mapped to higher level cubes and vice versa. Often multiple operators are combined in one OLAP analysis. The results of an OLAP operation are usually stored in presentation tools, like reports and graphs, for the decision-maker. The navigational operators or queries for cubes are drill-down, roll-up, slice, unslice, matrix slice, and matrix unslice, they are defined in Definitions 2.8 to 2.13.

Definition 2.8. The *drill-down* operator in dimension q, given by R_q^{-1}, is defined as

$$R_q^{-1}(X_1^{i_1} \times \ldots \times X_q^{i_q} \times \ldots \times X_n^{i_n}) = X_1^{i_1} \times \ldots \times r^{-1}(X_q^{i_q}) \times \ldots \times X_n^{i_n}.$$

Drill-down de-aggregates a cube to a lower dimension level. For example, a drill-down operation R_{Time}^{-1} on the Time dimension from the level Year to the level Quarter, applied to the full cube $\text{Time}^2 \times \text{Location}^3 \times \text{Product}^2$ results in the full cube $\text{Time}^1 \times \text{Location}^3 \times \text{Product}^2$.

Definition 2.9. The *roll-up* operator in dimension q, given by R_q^{+1}, is defined as

$$R_q^{+1}(X_1^{i_1} \times \ldots \times X_q^{i_q} \times \ldots \times X_n^{i_n}) = X_1^{i_1} \times \ldots \times r^{+1}(X_q^{i_q}) \times \ldots \times X_n^{i_n}.$$

Roll-up aggregates a cube along one or more dimension hierarchies to a higher dimension level. For example, a roll-up operation R_{Time}^{+1} on the full cube $\text{Time}^2 \times \text{Location}^3 \times \text{Product}^2$ results in the full cube $\text{Time}^3 \times \text{Location}^3 \times \text{Product}^2$. Obviously, drill-down and roll-up are the inverse of each other: $R_q^{+1}(R_q^{-1}(C)) = R_q^{-1}(R_q^{+1}(C)) = C$.

With these operators, we can determine the parents and children of a cube C. A *parent cube* C' is determined as the result of the roll-up operation $R_q^{+1}(C) = C'$ and reversely a *child cube* C is determined as the result of the drill-down operation $R_q^{-1}(C') = C$. A cube C might have multiple parent cubes, i.e. each applicable roll-up operation on C gives a parent cube. For example, the cube $C = [i_1, \ldots, i_q, \ldots, i_n]$ has $[i_1 + 1, i_2, \ldots, i_n]$, $[i_1, i_2 + 1, \ldots, i_n]$, ..., $[i_1, i_2, \ldots, i_n + 1]$ as its parent cubes, corresponding to all the different roll-up operations. Oppositely, a cube C might have multiple child cubes, i.e. each applicable drill-down operation on C gives a child cube. For example, the cube C has $[i_1 - 1, i_2, \ldots, i_n]$, $[i_1, i_2 - 1, \ldots, i_n]$, ..., $[i_1, i_2, \ldots, i_n - 1]$ as its child cubes, corresponding to all the different drill-down operations. Cubes with the same parent are siblings of each other.

Drill-down and roll-up operations are commutative.

Lemma 2.2.1. (*Commutativity of drill-down operators*). $R_p^{-1} \circ R_q^{-1}(C) = R_q^{-1} \circ R_p^{-1}(C)$ for any pair of drill-down operations.

Proof: This follows immediately from Definition 2.8. □

Commutativity between drill-down operators in the general situation of more than two dimensions, where $C = D_1^{i_1} \times D_2^{i_2} \times \ldots \times D_n^{i_n}$ is a straightforward generalization of Lemma 2.2.1, denoted by

$$\begin{aligned} R_1^{-i_1} \circ R_2^{-i_2} \circ \ldots \circ R_n^{-i_n}(C) &= R_n^{-i_n} \circ \ldots \circ R_2^{-i_2} \circ R_1^{-i_1}(C) \\ &= R_1^{-i_1} \circ (R_n^{-i_n} \circ \ldots \circ R_2^{-i_2}(C)) \\ &= R_1^{-i_1} \circ R_2^{-i_2} \circ (R_n^{-i_n} \circ \ldots \circ R_3^{-i_3}(C)) \\ &= \ldots \\ &= R_1^{-i_1} \circ R_2^{-i_2} \circ \ldots \circ (R_n^{-i_n}(C)) \\ &= R_1^{-i_1} \circ R_2^{-i_2} \circ \ldots \circ R_n^{-i_n}(C), \end{aligned}$$

where $R_q^{-n} = R_q^{-1} \circ R_q^{-1} \circ \ldots \circ R_q^{-1}$ and $q = 1, 2, \ldots, n$.

Lemma 2.2.2. (*Commutativity of roll-up operators*). $R_p^{+1} \circ R_q^{+1}(C) = R_q^{+1} \circ R_p^{+1}(C)$ for any pair of roll-up operations $R_p^{+1} \circ R_q^{+1}$.

Proof: This follows immediately from Definition 2.9. □

Definition 2.10. *Slice* is defined as

$$S^{X_q = Y_q}(X_1^{i_1} \times \ldots \times X_q^{i_q} \times \ldots \times X_n^{i_n}) = X_1^{i_1} \times \ldots \times Y_q^{i_q} \times \ldots \times X_n^{i_n},$$

where $Y_q^{i_q} \subset X_q^{i_q}$.

Slice performs a selection on the dimension level instances within a single domain of the cube, resulting in a *subcube*. For example, a slice operator with criterion (Year="2011") on the full cube $S^{\text{Year = 2011}}$ (Year × Country × Product) results in the subcube 2011 × Country × Product, or represented similarly as (2011, Country, Product). By definition combinations of slice operators are commutative $S^{X_q=Y_q}(S^{X_p=Y_p}(C)) = S^{X_p=Y_p}(S^{X_q=Y_q}(C))$.

In addition, the *dice* operator - which performs a selection on the dimension level instances within multiple domains of the cube - is defined as a composition of slice operators. For example, the following dice operator $S^{\text{Year = 2009}}$ ($S^{\text{Country = USA}}$ ($S^{\text{Product = Golf equipment}}$ (Year × Country × Product))) results in the cell (2009, United States, Golf Equipment), where all elements are instances.

Definition 2.11. *Unslice* is defined as

$$U^{X_q=D_q}(X_1^{i_1} \times \ldots \times X_q^{i_q} \times \ldots \times X_n^{i_n}) = X_1^{i_1} \times \ldots \times D_q^{i_q} \times \ldots \times X_n^{i_n}.$$

Unslice transforms one domain of the cube from the instance level to the schema level. It is the reverse of a slice. For example, an unslice operator with criterion Year on the cube U^{Year}(2011 × Country × Product) results in the full cube Year × Country × Product.

Definition 2.12. *Matrix slice* is defined as

$$S^{A_q^{i_q}=a_q^{i_q}}(X_1^{i_1} \times \ldots \times A_q^{\max_q}.\cdots.A_q^{i_q+1}.A_q^{i_q} \times \ldots \times X_n^{i_n}) = X_1^{i_1} \times \ldots \times A_q^{\max_q}.\cdots.A_q^{i_q+1}.a_q^{i_q} \times \ldots \times X_n^{i_n},$$

where $a_q^{i_q} \in A_q^{i_q}$.

Matrix slice performs a specific selection on a dimension level instance within a hierarchy, as described in Definition 2.4. Additionally, $A_q^{\max_q}.\cdots.A_q^{i_q+1}.A_q^{i_q}$ might be sliced on any other element of its ancestry. For example, the matrix slice operation $S^{\text{Month = January}}$ (Year.Quarter.Month × Country × Product) results in the cube Year.Quarter.January × Country × Product.

Definition 2.13. *Matrix unslice* is defined as

$$U^{a_q^{i_q}=A_q^{i_q}}(X_1^{i_1} \times \ldots \times a_q^{\max_q}.\cdots.a_q^{i_q+1}.a_q^{i_q} \times \ldots \times X_n^{i_n}) = X_1^{i_1} \times \ldots \times a_q^{\max_q}.\cdots.a_q^{i_q+1}.A_q^{i_q} \times \ldots \times X_n^{i_n}.$$

Matrix unslice is the reverse of a matrix slice. For example, the matrix unslice operation U^{Month} (2009.Q1.Jan $\times$ Country $\times$ Product) results in the cube 2009.Q1.Month $\times$ Country $\times$ Product.

In addition, we verify for other combinations of navigational operators whether they commute with each other or not. The drill-down (roll-up) operator and the (matrix) slice operator between two dimensions D_p and D_q are commutative, i.e.

$$S^{X_q=Y_q}(R_p^{-1}(C)) = R_p^{-1}(S^{X_q=Y_q}(C)), \tag{2.1}$$

and the roll-up (drill-down) operator and the (matrix) unslice operator are commutative as well, i.e.

$$U^{X_q=D_q}(R_q^{+1}(C)) = R_q^{+1}(U^{X_q=D_q}(C)). \tag{2.2}$$

In a single dimension D_p commutativity holds for the drill-down operator and the (matrix) slice operator but not for the roll-up operator and the (matrix) unslice operator.

Example 2.2.2. The following two successive operations on some cube C given by Year $\times$ Country result in the cube C' given by 2009.Quarter $\times$ Country:

- a drill-down followed by a matrix slice, i.e.

$$S^{\text{Year}=2009}(R_{\text{Time}}^{-1}(C)) = C',$$

- and, a slice followed by a drill-down, i.e.

$$R_{\text{Time}}^{-1}(S^{\text{Year}=2009}(C)) = C'.$$

Furthermore, other OLAP operations are *rank*, i.e. order the data points in the cube's cells in a specified order, and *pivot*, i.e. rotate the data axes of the cube. We refer to Han and Kamber (2005) for an elaborate overview on these navigational operators. We illustrate the working of the operators on the running-example of the multi-dimensional financial database.

Example 2.2.3. Figure 2.3 shows the three-dimensional financial cube $\text{T}^2 \times \text{L}^3 \times \text{P}^2$ derived from Example 2.1.1 and the effect of a number of roll-ups on the cube. The

cube contains the sports equipment sales data of a global chain store, the GoSales-company, collected for different countries over the past years. The cube in the north-west visualizes the result of the slice operator $S^{\text{ProductLine = Golf Equipment}}(T^2 \times L^3 \times P^2)$ with a dark grey color. Notice that this selection 'slices of' a part of the cube. Moreover, the result of a specific dice operator, the cube 2006 × United States × ProductLine is visualized with a light grey color. Notice that due to the dice a row in the cube is selected.

Definition 2.14. The *context of a cell c* is defined as the cube C after the application of one or more (matrix) unslice operations on the cell c of the form $U^{d_q=D_q}(c) = C$.

Obviously, a cell has many *context cubes*, dependent on the number of domains and hierarchies of the cube. If a cell is unsliced over all its associated domains, we obtain the full cube as the cell's context. For example, the cell (2006, United States, Golf-Equipment) in Figure 2.3 might be unsliced, with the operations: U^{Year}, U^{Country}, $U^{\text{ProductLine}}$, or any combination of these operators, to its various context cubes. Moreover, the maximum number of context cubes an arbitrary cell can be (matrix) unsliced to is:

$$\mathrm{T} = (\prod_{i=1}^{n} 2^{l_i}) - 1, \tag{2.3}$$

where l_i is the number of levels of dimension i (excluding the top-level All).

Example 2.2.4. Suppose that (2009.Q2, Germany) is a cell in the full cube Year.Quarter × Country then Year.Q2 × Country is an example context cube and the total number of context cubes is $(2^2 \cdot 2^1) - 1 = 7$.

2.2.4 Aggregation lattice

Given a cube C and a set S of roll-up operators we can generate an *aggregation lattice* L of cubes by applying all possible subsets of S to the cube C. The minimal element of L is C and the maximal element of L is the cube where all operators in S are applied to C. The minimal element is also called the *base cube* of the lattice and the maximal element is the *top cube*. This is stated more formally in Definition 2.15.

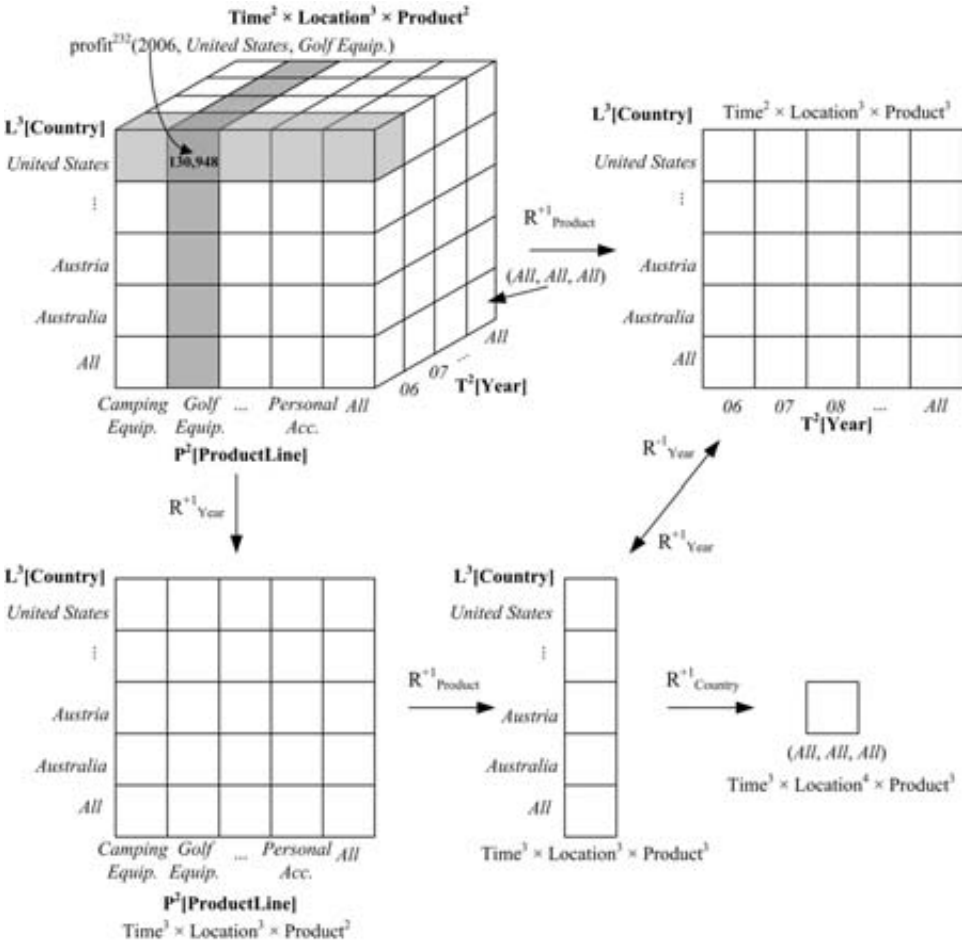


Figure 2.3: The cube $T^2 \times L^3 \times P^2$ represents the example financial database with the measure profit and the dimensions Time, Location, and Product in the north-west. The effects of roll-up and drill-down operations on the financial cube are depicted in the other figures. In the figures it can be seen that the cube $T^2 \times L^3 \times P^2$ can be rolled up via R^{+1}_{Product}, R^{+1}_{Year}, and R^{+1}_{Country} or via R^{+1}_{Year}, R^{+1}_{Product}, and R^{+1}_{Country} to the cube $T^3 \times L^4 \times P^3$, which is equivalent with the cell (All, All, All). Moreover, it can be seen that the cube $T^3 \times L^3 \times P^3$ is drilled down to the cube $T^2 \times L^3 \times P^3$ with the operation R^{-1}_{Year}. In the north-west cube the result of a slice and dice operation is visualized, see Example 2.2.3 for a description. The figure is adapted from (Han and Kamber, 2005).

Definition 2.15. Given a cube $C = X_1^{i_1} \times X_2^{i_2} \times \ldots \times X_n^{i_n}$ and integers $I_k \geq 0$ for $k = 1, 2, \ldots, n$ we define the lattice of cubes

$$L = \{X_1^{i_1+j_1} \times X_2^{i_2+j_2} \times \ldots \times X_n^{i_n+j_n} \mid 0 \leq j_i \leq I_i, \text{ for } i = 1, 2, \ldots, n\}.$$

The lattice L can alternatively be denoted by

$$L = \{R_1^{+j_1} \circ R_2^{+j_2} \circ \ldots \circ R_n^{+j_n}(C) \mid 0 \leq j_i \leq I_i, \text{ for } i = 1, 2, \ldots, n\}. \tag{2.4}$$

Notice that the lattice structure of $(L, \leq)$ is isomorphic to the lattice of indices defined by

$$\{[j_1, j_2, \ldots, j_n] \mid 0 \leq j_i \leq I_i, \text{ for } i = 1, 2, \ldots, n\},$$

where the partial ordering is defined by

$$[l_1, l_2, \ldots, l_n] \leq [k_1, k_2, \ldots, k_n] \text{ iff } l_i \leq k_i \text{ for } i = 1, 2, \ldots, n.$$

Figure 2.4 depicts a simple lattice with $n = 3$, $I_1 = 1$, $I_2 = 1$, and $I_3 = 1$. If C_B is the base cube in the hierarchy of the OLAP structure and if we apply all possible roll-ups to C_B we get the complete lattice of cubes $L_{\max}$, where $C_B = [0, 0, \ldots 0]$ is the base cube and $C_T = [\max_1, \max_2, \ldots, \max_n]$ is the top cube. The total number of cubes in the lattice $L_{\max}$ is (Han and Kamber 2005)

$$|L_{\max}| = \prod_{k=1}^{n} (I_{\max_k} + 1). \tag{2.5}$$

The *downset* $\{\downarrow C\}$ of a cube C in a lattice L is the set of all cubes that can be obtained by applying drill-down operators on C. Or alternatively,

$$\{\downarrow C\} = \{C' \in L | C' \leq C\}, \tag{2.6}$$

and the *upset* is defined analogously

$$\{\uparrow C\} = \{C' \in L | C' \geq C\}. \tag{2.7}$$

Given two cubes C and C' we define their *join* as follows

$$\bigvee(C, C') = \min\{E | E \geq C \text{ and } E \geq C'\}, \tag{2.8}$$

i.e. the smallest cube in the intersection of $\{\uparrow C\}$ and $\{\uparrow C'\}$. Similarly, we define the *meet* of two cubes C and C'

$$\bigwedge(C, C') = \max\{E | E \leq C \text{ and } E \leq C'\}, \tag{2.9}$$

i.e. the largest cube in the intersection of $\{\downarrow C\}$ and $\{\downarrow C'\}$.

Example 2.2.5. In Figure 2.4 an example lattice L with the 3-dimensional cube $C_B = [0, 0, 0]$ at level 0 as its base and $C_T = [1, 1, 1]$ at level 3 as its top. In this lattice it can easily be observed that all cubes can be derived from C_B, by the application of one or more roll-up operations in a specific order.

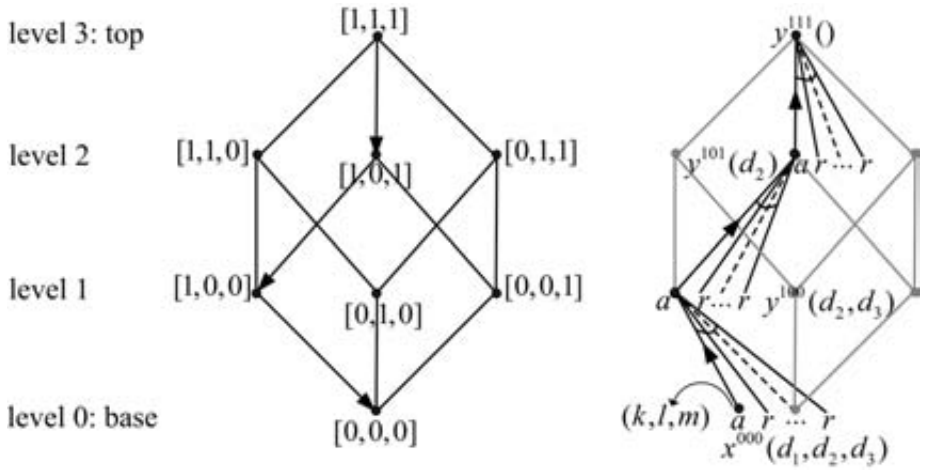


Figure 2.4: The lattice of cubes L is formed by rolling-up the base cube $[0, 0, 0]$ over all its domains and dimension hierarchies, in any order, to the top $[1, 1, 1]$ (left). An analysis path is projected in the lattice (right). This path is formed by rolling up the cell $x^{000}(k, l, m)$ over the path $[0, 0, 0] \rightarrow [1, 0, 0] \rightarrow [1, 0, 1] \rightarrow [1, 1, 1]$ (right) in L. In this way a path from $x^{000}(k, l, m)$ to y^{111}(All, All, All) is created.

In general, we can determine a specific *ancestor* (*descendant*) of a cube C by the application of a number of roll-up (drill-down) operations in L. If C_p and C_q are two cubes in L, where $C_p \leq C_q$, then C_q is an ancestor of C_p, and C_p is a descendant

of C_q. Obviously, in L the base cube C_B has no children and the top cube C_T has no parents, by definition. The aggregation lattice L can now alternatively be defined as a partially ordered set in which any two cubes have an unique join (i.e. smallest common ancestor) and meet (i.e. largest common descendant). The downset of a cube C in L is the *set of all its descendants* and the upset of the cube C in L is the *set of all its ancestors*. Moreover, L is a bounded lattice because is has a least element, the base cube C_B, and a maximum element, the top cube C_T. A property of the base cube C_B is that all cubes in L can be obtained from it by applying roll-up operations in a specific order. Conversely, a property of the top cube C_T is that all cubes in L can be obtained from it by applying drill-down operations in a specific order.

Furthermore, we define the level of a cube C as the number of roll-ups that must be applied to C_B to get C (see Figure 2.4).

Example 2.2.6. In the lattice depicted in Figure 2.4, the upset of cube $[0,1,0]$ is given by $\{\uparrow [0,1,0]\} = \{[0,1,0],[1,1,0],[0,1,1],[1,1,1]\}$ and is obtained by the following roll-up operations $R^{+1}_{D_1}([0,1,0])$,$R^{+1}_{D_3}([0,1,0])$, and $R^{+1}_{D_1}(R^{+1}_{D_3}([0,1,0]))$. In the same lattice, the downset of cube $[1,1,0]$ is given by $\{\downarrow [1,1,0]\} = \{[1,1,0],[1,0,0],[0,1,0],[0,0,0]\}$ and is obtained by the following drill-down operations $R^{-1}_{D_2}([1,1,0])$, $R^{-1}_{D_1}([1,1,0])$, and $R^{-1}_{D_1}(R^{-1}_{D_2}([1,1,0]))$. It can easily be seen that $\{\downarrow [1,1,0]\}$ is a sublattice L' in L with base cube $[0,0,0]$ and top cube $[1,1,0]$.

2.2.5 Analysis paths

The business analyst working with the multi-dimensional database can create an analysis path in L by the application of navigational operators.

Definition 2.16. An *analysis path* p is defined as a sequence of cubes in L, such that each of its cubes is a drill-down (roll-up) of its parent cube in the sequence.

The length of a path is the number of drill-down operations that is used in the path. In the path $p(C, C')$, the cube C is the start cube and C' is its end cube. If $C = [i_1, i_2, \ldots, i_n]$ and $C' = [j_1, j_2, \ldots, j_n]$ are cubes in L where $C \leq C'$, then the length of the path is $|p(C, C')| = (i_1 + i_2 + \ldots + i_n) - (j_1 + j_2 + \ldots + j_n)$. Notice that all paths from C to C' have the same length.

Moreover, an analysis path p can be represented as a binary *analysis matrix* where the columns represent the dimensions of the cube from $D_1, D_2, \ldots, D_n$ and the rows represent the levels of the lattice L from level $(i_1+i_2+\ldots+i_n)$ to level $(j_1+j_2+\ldots+j_n)+1$. Each row in the matrix has one cell with the value -1, that represents a single drill-down in dimension D_q from one level to the next in L, the other cells in the row have the value 0. The first row in the matrix corresponds with the first drill-down operation in some dimension from level $(i_1+i_2+\ldots+i_n)$ to level $(i_1+i_2+\ldots+i_n)-1$, the second row in the matrix corresponds with the second drill-down operation in some dimension from level $(i_1+i_2+\ldots+i_n)-1$ to level $(i_1+i_2+\ldots+i_n)-2$, and so on, until the last row in the matrix.

Obviously, there are usually multiple paths in L from its top to its base, or vice versa, corresponding with different analyses that might be created by the analyst. For example, in the lattice of cubes depicted in Figure 2.4, the following sequence of drill-down operations from C_T to C_B, $R^{-1}_{D_2}(C_T)$, $R^{-1}_{D_3}(R^{-1}_{D_2}(C_T))$, and $R^{-1}_{D_1}(R^{-1}_{D_3}(R^{-1}_{D_2}(C_T)))$, creates the drill-down path $[1,1,1] \to [1,0,1] \to [1,0,0] \to [0,0,0]$. This path represented in matrix notation is given by

$$\begin{array}{c} \\ \text{level 3} \\ \text{level 2} \\ \text{level 1} \end{array} \begin{array}{c} \begin{array}{ccc} D_1 & D_2 & D_3 \end{array} \\ \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ -1 & 0 & 0 \end{pmatrix} \end{array}.$$

Another drill-down path from C_T to C_B in this figure is given by $[1,1,1] \to [0,1,1] \to [0,0,1] \to [0,0,0]$. Reversely, an example of a roll-up path from C_B to C_T is given by $[0,0,0] \to [0,0,1] \to [0,1,1] \to [1,1,1]$.

The total number of analysis paths P in L can be very large. Suppose that n_k is the number of possible levels in dimension k. Then the length of a drill-down from C_T to C_B is given by $n_1 + n_2 + \ldots + n_k$.

Theorem 2.2.3. The total number of drill-down paths from C_T to C_B is

$$P = \frac{(n_1 + n_2 + \ldots + n_k)!}{n_1! n_2! \ldots n_k!}. \tag{2.10}$$

Proof: If in some drill-down path $p(C_T, C_B)$ the order of the drill-down operators is changed, we get a different path. Accordingly, the number of paths would be equal

to the number of permutations of the sequence of drill-down operators: $(n_1 + n_2 + \ldots + n_k)!$. However, there is no change if two operators are interchanged that act on the same dimension, therefore we have to divide by $n_1!n_2!\ldots n_k!$.

Example 2.2.7. Figure 2.5 provides an illustration. The figure depicts two example lattices. The number of drill-down analysis can be computed using formula (2.10).

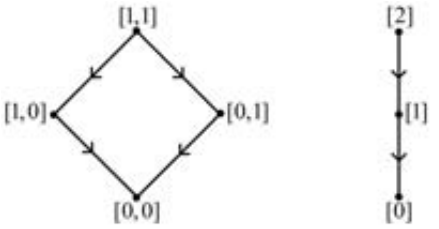


Figure 2.5: In the example lattice on the left hand side there are $P = 2!/1!1! = 2$ drill-down paths, where $n_1 = 1$ and $n_2 = 1$, and in the example lattice on the right hand side there are $P = 2!/2! = 1$, where $n_1 = 2$.

Example 2.2.8. Suppose we have 4 dimensions where each dimension has three levels. The base cube of the lattice of cubes is $C_B = [0, 0, 0, 0]$ and the top is $C_T = [3, 3, 3, 3]$. The length of a path p from the top to the base is $3+3+3+3 = 12$ and the number of paths from C_T to C_B is $12!/3!3!3!3! = 479,001,600/1,296 = 396,600$.

2.2.6 Measures

The measures are derived from the column names of the star scheme's fact table, and the *measure values* are entries of the fact table.

Definition 2.17. A *measure* y is defined as a function on a cube C

$$y^{i_1 i_2 \ldots i_n} : D_1^{i_1} \times D_2^{i_2} \times \ldots \times D_n^{i_n} \to \mathbb{X},$$

where measure values are $\mathbb{X} = \mathbb{N}$, $\mathbb{Z}$, or $\mathbb{R}$.

We sometimes use the term *variable* instead of measure.

Data are the measure values of a measure y in a particular cell like, for example, profit^{232} (2006, United States, Golf Equipment)= 130,948. The combination of a cell and a measure is called a *data point*. Each cube C can be viewed as a specific collection of cells, where we can store measure values. The measure's upper indices indicate the level of its cube or cell, i.e. the measure $y^{i_1 i_2 \ldots i_n}(C)$ is a function on the cube $C = [i_1, i_2, \ldots, i_n]$. For example, profit^{232} is a measure on the cube $T^2 \times L^3 \times P^2$ in Figure 2.3. If necessary, we use the shorthand notation $y^{\mathbf{i}}(C)$ for $y^{i_1 i_2 \ldots i_n}(C)$ where $\mathbf{i} = i_1 i_2 \ldots i_n$ or the shorthand notation $y^{i_q}(C)$ for $y^{i_1 \ldots i_q \ldots i_n}(C)$, where D_q is some arbitrary dimension D_q. Besides, if no confusion can arise we will leave out the upper indices, and write profit(2006, United States, Golf Equipment).

Furthermore, if a measure is not defined for a particular cell then $y^{\mathbf{i}}(c) = \emptyset$. We call such a cell an *empty cell* or missing value. Empty cells in an OLAP cube can have various causes (Thomsen 1997). For example, data for a cell can be missing but also forthcoming, like a late sales report. In some situations an empty cell means that data can never apply to the cell, such as the name of a bachelor employee's spouse. In other situations an empty cell means that zeros are being "suppressed" like the zero associated with individual product sales in a store that carries many products but that only sells 5% of its items on any one day.

In summary, we presented an original, generic notation in Section 2.2 to capture the structures of the dimension table and the fact table in the star model. In the first place, the concept of a domain $D_k^{i_k}$ represents the dimension table, the dimension hierarchy schema represents the table's column names, and the dimension hierarchy instances represent the entries of this table. From these notions we compose the concept of a cube $C = [i_1, i_2, \ldots, i_n]$ that lives in an aggregation lattice L. Each cube in L can be manipulated by set of navigational operators. In the second place, the concept of a measure $y^{\mathbf{i}}(C)$ represents the fact table's column names and the measure values represent the entries of this table. In particular, a measure is defined as a function on a cube C. In this way, a measure is connected with a cube, consistent with a fact table that is connected with a set of the dimension tables. Correspondingly, a multi-dimensional database can be interpreted as a lattice of cubes specified by the star scheme.

2.3 OLAP equations

The cell values of the base *data cube* are denoted by $y^{00\ldots0}(C_B)$. From $y^{00\ldots0}(C_B)$ the measures can be aggregated by typical *aggregation functions* to higher levels. These functions are SUM($y(C)$), COUNT($y(C)$), MAX($y(C)$), MIN($y(C)$), and AVG($y(C)$), and are implemented in most OLAP software packages. For example, the measure profit may be aggregated over the Time dimension of Figure 2.2, with the SUM($y(C)$) function from the monthly profit on T^0 to the quarterly profit on T^1 or the yearly profit on T^2. In general, aggregating measure values in $y^{i_1 i_2 \ldots i_n}(C)$ with some function along the hierarchies of different domains in its upset creates multi-dimensional views on the data.

The application of a specific aggregation function f on the measure values of each cube $y(C)$ in L creates a *system of drill-down equations*, given by

$$y^{i_1 \ldots i_q \ldots i_n}(C) = f(y^{i_1 \ldots (i_q - 1) \ldots i_n}(R_q^{-1}(C))). \tag{2.11}$$

In the above system of equations we distinguish between base and non-base variables.

Definition 2.18. The measure values $y^{00\ldots0}(c)$ in the base cube C_B are called the *base variables*.

A base variable is sometimes denoted by $x(c)$ to distinguish them clearly from dependent variables. Obviously, base variables are non-aggregated and are directly derived from the star model's fact table. The total number of base variables $y^{00\ldots0}(c) \neq \emptyset$ in C_B, corresponds with the number of rows in the fact table. A *non-base* or dependent variable $y^{i_1 i_2 \ldots i_n}(c)$ where $i_1 + i_2 + \ldots + i_n > 0$ can be computed by using Equation (2.11) repeatedly.

Definition 2.19. $y^{max_1 max_2 \ldots max_n}(C_T)$ is defined as the *root variable*.

The root variable is a non-base variable that *only* appears on the LHS of an equation in (2.11).

If we consider a sublattice L' of L we derive a *subsystem of drill-down equations*. In this subsystem we call variables $y^{00\ldots0}(c)$ in the base cube C'_B, i.e. the base of the

L' re-indexed to $[0, 0, \ldots, 0]$, the subsystem's base variables. All other variables are called the subsystem's non-base variables.

In Figure 2.6, a graphical representation of a system of drill-down equations is shown. In this figure, the lattice of cubes in Figure 2.4, is instantiated for base variables $x(c)$ in $C_B = [0, 0, 0]$. This system is composed out of 27 equations with 8 base variables and 19 non-base variables. Each separate equation in the system is denoted in the figure by a small arc '$\smile$' between the edges.

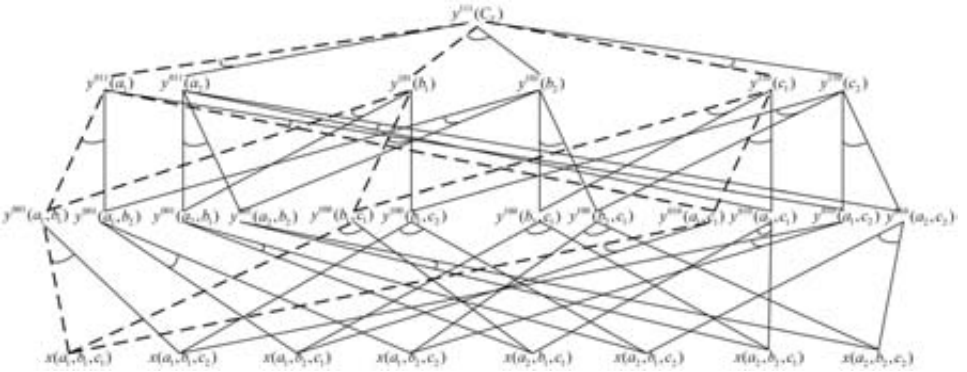


Figure 2.6: Graphical representation of a system of drill-down equations. The result of the instantiation is a semilattice with top $y(C_T)$ and base variables $x(c)$. The upset of each base variable $x(c)$ forms a lattice with $y^{111}(C_T)$ as its top.

Suppose we have a lattice L where a measure y, aggregated by some function f, is instantiated from the schema level to the instance level. The result of the instantiation is a *semilattice* SL with multiple base elements and a single top. Each base variable $x(c)$ in the base cube C_B represents a base element of SL. The root variable $y^{\max_1 \max_2 \ldots \max_n}(C_T)$ represents the top element in SL. The instantiated system of equations constitutes a semilattice because any two measure values $y(c_p)$ from cube C_p and $y(c_q)$ from cube C_q in it have an unique smallest common ancestor, i.e. it is an *ancestor-semilattice*. However, it is not a full lattice, see Definition 2.15, because *not* any two measures values $y(c_p)$ from cube C_p and $y(c_q)$ from cube C_q have an unique largest common successor.

In SL, the upset $\{\uparrow c\}$ of a cell c is defined as the set of all its ancestor cells and the downset $\{\downarrow c\}$ of a cell c is the set of all its descendant cells. Basically, SL is composed out of a set of related lattices, because the upset of each cell c forms a lattice L (see Definition 2.15), where the cell c is the base cell and C_T is the root cell. In Figure 2.6, the upset $\{\uparrow (a_1, b_1, c_1)\}$ is represented graphically with dashed lines to provide an illustration. This illustration shows that $\{\uparrow (a_1, b_1, c_1)\}$ is a lattice with base cell (a_1, b_1, c_1) and root cell C_T. Furthermore, in SL the downset of each cell c forms a semilattice, where the cell c is the root cell and a subset of descendant cells from C_B represent its base elements.

2.3.1 Drill-down equations

This section considers the most common types of aggregations of (2.11); the additive SUM($y(C)$) and COUNT($y(C)$) function, and the non-additive AVG($y(C)$), MAX($y(C)$) and MIN($y(C)$) function.

Additive drill-down equations

Definition 2.20. The measure y is an *additive measure* if for every cell $c \in C$, we have

$$y^{i_1 \ldots i_q \ldots i_n}(c) = \sum_{c' \in R_q^{-1}(c)} y^{i_1 \ldots (i_q - 1) \ldots i_n}(c'). \tag{2.12}$$

In words, the value of the measure of cell c is the sum of the values of the children of c in any dimension (see also Lenz and Shoshani (1997)). Equation (2.12) can also be written on the functional level as

$$y^{i_1 \ldots i_q \ldots i_n}(C) = \sum_{C' \in (S_q(R_q^{-1}(C)))} y^{i_1 \ldots (i_q - 1) \ldots i_n}(C'), \tag{2.13}$$

where the sum is over the slices of the cube $R_q^{-1}(C)$ in the dimension q.

Example 2.3.1. From our example database, we could inspect the measure revenues as a function on the subcube C, given by 2011 × All-Locations × Productline. This cube is part of the lattice L with cubes revenues(C), formed by rolling-up with the

SUM(revenues(C)) aggregation function. By applying Equation (2.13) two times we get

$$\begin{aligned}\text{revenues}^{242}(C) &= \sum_k \text{revenues}^{142}(S^k_{\text{Year}}(R^{-1}_{\text{Year}}(C))) \\ &= \sum_l \sum_k \text{revenues}^{132}(S^l_{\text{Location}}(R^{-1}_{\text{Location}}(S^k_{\text{Year}}(R^{-1}_{\text{Year}}(C))))).\end{aligned}$$

The same equation on the cell level reads

$$\begin{gathered}\text{revenues}^{242}(2011, \text{All-Locations}, \text{Golf-Equipment}) = \\ \sum_{j=1}^{4} \sum_{k=1}^{20} \text{revenues}^{132}(2011.\text{Quarter}_j, \text{Country}_k, \text{Golf-Equipment}),\end{gathered}$$

where $S^j(2011.\text{Quarter}) = 2011.\text{Quarter}_j$.

Moreover, if there is no confusion about the level in the lattice or dimension hierarchy we occasionally use the short-hand notation $y(\ldots, +, \ldots)$ for an additive measure, for the LHS of Equation (2.13), where the plus sign signifies summarization over that dimension.

The additive COUNT($y(C)$) function is defined similarly, and treated as a special case of the SUM($y(C)$) function, only this operator summarizes dimension hierarchy instances instead of measure instances. Basically, this function counts the number of instance elements in the base cube $[0, 0, \ldots, 0]$ for the downset of variable $y^{i_1 i_2 \ldots i_n}(d_1, d_2, \ldots, d_n)$. For example, the function is used to compute the number of products per P[ProductLine] or P[ProductType] or the number of employees per Country or City. The COUNT($y(C)$) function can also be interpreted as an additive measure where all values on the base level are 0 or 1.

For any two cubes C and $C' \in \{\downarrow C\}$ we may consider the set of all drill-down paths from C to C' $P(C, C')$ and the lattice L of all cubes in $P(C, C')$ (see Definition 2.15). This lattice has a top cube $C_T = C$ and base cube $C_B = C'$. Along any path from C_T to C_B we can apply Equation (2.12) repeatedly to the cells of the cubes in the drill-down sequence of the path. In doing so, the value of an additive measure in the cube C in the sequence, is expressed as a sum over its values on the cells of a child cube down in the hierarchy.

This typical OLAP feature is expressed in the following theorem.

Theorem 2.3.1. The measure values of the top cube's cell can be expressed as the sum of the measure values of the base cube's cells

$$y^{\max_1 \max_2 \ldots \max_n}(c) = \sum_{c_n \in R_n^{-\max_n} \circ \ldots \circ R_2^{-\max_2} \circ R_1^{-\max_1}(c)} x(c_n), \tag{2.14}$$

where $R_n^{-\max_n} \circ \ldots \circ R_2^{-\max_2} \circ R_1^{-\max_1}(C_T) = C_B$ is a drill-down path from C_T to C_B, $(n = \max_1 + \max_2 + \ldots + \max_n)$, c is a cell in C_T and $c_n \in C_B$. Furthermore, expression 2.14 is independent of the drill-down path chosen from C_T to C_B.

Proof. By applying Equation 2.12 repeatedly on a cell $c \in C_T$ we get

$$\begin{aligned}
y^n(c) &= \sum_{c_1 \in R_1^{-1}(c)} y^{n-1}(c_1) \\
&= \sum_{c_1 \in R_1^{-1}(c)} \sum_{c_2 \in R_2^{-1}(c_1)} y^{n-2}(c_2) \\
&= \ldots \\
&= \sum_{c_1 \in R_1^{-\max_1}(c)} \sum_{c_2 \in R_2^{-\max_2}(c_1)} \cdots \sum_{c_n \in R_n^{-\max_n}(c_{n-1})} y^0(c_n) \\
&= \sum_{c_n \in R_n^{-\max_n} \circ \ldots \circ R_2^{-\max_2} \circ R_1^{-\max_1}(c)} x(c_n).
\end{aligned}$$

Here c_i are cells on level $n - i$, $i = 1, 2, \ldots, n$. If we choose another path from C_T to C_B only the order of the drill-down operators changes in expression 2.12. But by commutativity of the drill-down operator (Lemma 2.2.1) this results in the same sum. It is only the order of summation that changes along different drill-down paths. □

Remark 2.3.1. From Theorem 2.3.1 it follows that the system of additive drill-down equations is *uniquely solvable*.

Remark 2.3.2. If C_T is the top cube of the whole lattice and consists of a single cell, then the sum in expression 2.14 extends over all cells in the base cube and is the grand total.

Example 2.3.2. We illustrate Theorem 2.3.1 with an example. In Figure 2.6, we show the composition of the value of the additive measure y^{101} in cell (b_1) into two drill-down equations related to the path $p([1,0,1],[0,0,0])$, specified in detail as $[101] \to [100] \to [000]$

$$\begin{aligned}
y^{101}(b_1) &= y^{100}(b_1, c_1) + y^{100}(b_1, c_2) \\
&= x(a_1, b_1, c_1) + x(a_2, b_1, c_1) + x(a_1, b_1, c_2) + x(a_2, b_1, c_2).
\end{aligned}$$

Here the final equation is the expression of $y^{101}(b_1)$ into the sum of an unique set of base variables. If we would select a different path $p([1,0,1],[0,0,0])$, i.e. the path $[101] \rightarrow [001] \rightarrow [000]$, we would get the same expression, because this result is independent of the selected path.

Additivity criteria

Additivity is an important criterion for the quality of multi-dimensional database design because it ensures the correctness of aggregations. The violation of this condition can lead to erroneous conclusions and decisions. Both Horner et al. (2004) and Lenz and Shoshani (1997) have studied additivity in OLAP databases and statistical databases respectively.

Horner et al. (2004) make a distinction between *additive*, *semi-additive* and *non-additive measures*. In Kimball (1996) it is argued that the "the most useful facts are numeric and additive". Additive measures have the property that they can be meaningfully aggregated along any dimension. For example, it makes sense to add total sales for the Product, Location, and Time dimension because this causes no overlap among the real-world phenomena that generated the individual sales. A measure is semi-additive if it is only additive across certain dimensions, but not all. For example, the measure stock-at-hand cannot be aggregated along a Time dimension because they represent a "snapshot" of a level or balance at one point in time. But this measure can be aggregated along a Product dimension and return a valid total. In practice, multi-dimensional databases are littered with non-additive measures. Percentages and ratios are examples of non-additive measures.

The structure of the dimension hierarchy is of central concern with respect to additivity, because the primary method of rolling-up and drilling-down data is along these pre-defined hierarchies. Therefore, two standard requirements are that dimension hierarchies have to be *strict* and *complete* (Lenz and Shoshani 1997). Most multi-dimensional data models as well as the one used in this thesis demand that the hierarchies of a dimension are strict. This means that there exists a many-to-one relationship between the level instances of two dimension levels D_i^{q+1} and D_i^q, with $D_i^q \prec D_i^{q+1}$, to ensure correct aggregation of measure values (Lenz and Shoshani 1997). The term completeness in dimension hierarchies means that all children of a

parent in the hierarchy tree are accounted for, i.e., that there is no missing or inaccurate data. In addition, other additional requirements for additivity placed on dimension hierarchies in multi-dimensional data models are that they have to be *onto* and *covering*, see Pedersen et al. (1999) for more detail.

Non-additive drill-down equations

Other examples of non-additive measures include measures that are derived by using an aggregation function like AVG($y(C)$). For instance, in the health care domain, it is often important to analyse the number of patients admitted to a hospital, like the average number of hourly admissions. The average number of hourly admissions cannot be combined along any dimension, because the aggregation function prevents combining lower-level averages to higher level averages. A formal definition runs as follows:

Definition 2.21. The measure y is an *average measure* if for every cell $c \in C$, where C is a cube in the lattice L, the following holds

$$\bar{y}^{i_1 \ldots i_q \ldots i_n}(c) = \frac{1}{|R_q^{-1}(c)|} \sum_{e \in R_q^{-1}(c)} y^{i_1 \ldots (i_q - 1) \ldots i_n}(e). \tag{2.15}$$

If Definition 2.21 is instantiated for a single cell $(\ldots, A, \ldots)$ in cube C then we obtain its instance representation:

$$\bar{y}^{i_1 \ldots i_q \ldots i_n}(\ldots, A, \ldots) = \frac{1}{J} \sum_{j=1}^{J} y^{i_1 \ldots (i_q - 1) \ldots i_n}(\ldots, A.a_j, \ldots). \tag{2.16}$$

where $A \in D_i^{i_q}$ is a parent, $A.a_j \in D_i^{i_q - 1}$ is a child, q is some level in the dimension hierarchy, and J represents the number of level instances in $D_i^{i_q - 1}$.

A typical OLAP feature is expressed in the following theorem.

Theorem 2.3.2. The measure values of the cell of the top cube can expressed as the average of the measure values of cells of the base cube

$$y^{\max_1 \max_2 \ldots \max_n}(c) = \frac{1}{|C_B|} \sum_{c_n \in R_n^{-\max_n} \circ \ldots \circ R_2^{-\max_2} \circ R_1^{-\max_1}(c)} x(c_n), \tag{2.17}$$

where $R_n^{-\max_n} \circ \ldots \circ R_2^{-\max_2} \circ R_1^{-\max_1}(C_T) = C_B$ is a drill-down path from C_T to C_B, $(n = \max_1 + \max_2 + \ldots + \max_n)$, c is a cell in C_T and $c_n \in C_B$. Furthermore, expression 2.17 is independent of the drill-down path chosen from C_T to C_B.

The proof of Theorem 2.3.2 is similar with the proof of Theorem 2.3.1 with the difference that the RHS of each drill-down equation is divided by the number of cells of the cube under consideration.

For completeness we mention two other non-additive measures. The maximum measure is

$$y^{i_1 \ldots i_q \ldots i_n}(c) = \max(y^{i_1 \ldots (i_q - 1) \ldots i_n}(R_q^{-1}(c))), \tag{2.18}$$

and the minimum measure is

$$y^{i_1 \ldots i_q \ldots i_n}(c) = \min(y^{i_1 \ldots (i_q - 1) \ldots i_n}(R_q^{-1}(c))). \tag{2.19}$$

The measure in the top cube is the maximum respectively minimum of all the values in the base cube, denoted by

$$y^{\max_1 \max_2 \ldots \max_n}(c) = \max_{c \in C_B} x(c), \tag{2.20}$$

and

$$y^{\max_1 \max_2 \ldots \max_n}(c) = \min_{c \in C_B} x(c). \tag{2.21}$$

2.3.2 Relations between measures

The measures that can be analysed by the same set of domains $D_1^{i_1} \times D_2^{i_2} \times \ldots \times D_n^{i_n}$ are described by the fact table in the OLAP database. A business model M is a system of relations between measures in this table. This model represents relevant financial and operating variables and relations between them. These relations can be derived from many business domains, like finance, accounting, logistics, and so forth.

Definition 2.22. Relations between measures are denoted by

$$y^{\mathbf{i}}(C) = f(\mathbf{x}^{\mathbf{i}}(C)), \tag{2.22}$$

where y and $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ are measures on the same cube $C = [i_1, i_2, \ldots, i_n]$, as specified in Definition 2.17.

The function f can have various functional forms here, representing a business model with (mixed) relations that are additive, multiplicative, ratio, and so on. A *system of business model equations*, where each equation is of type (2.22) is denoted by M. Business model equations hold for individual cubes. Therefore we may leave out the upper indices in Equation (2.22). If (2.22) holds for all cubes we may write $y = f(\mathbf{x})$.

Table 1.1 on page 12 depicts relations from the financial database example. For example, a typical instance of a business model equation

$$\text{profit}(c) = \text{revenues}(c) - \text{costs}(c)$$

on aggregation level 233 is given by

$$\text{profit}^{233}(2011, \text{Spain}) = \text{revenues}^{233}(2011, \text{Spain}) - \text{costs}^{233}(2011, \text{Spain}).$$

A *directed acyclic graph* $G(M) = (V, E)$ is associated with M (Feelders 1993). The vertex set $V(M)$ contains as elements all variables appearing in the model. The edge set $E(M)$ contains a directed edge from vertex x_i to x_j iff:

$$x_j = f(\ldots, x_i, \ldots) \in M.$$

We assume that the modeled graph $G(M)$ is acyclic. This restriction excludes business models that contain simultaneous equations. Nodes in the business model graph, with zero indegree, represent variables that cannot be explained in M. M^p denotes the level p in the business model, where $p = 0, 1, \ldots, d$. The root of the graph y is on level 0 (M^0), the children of the root $x_1, x_2, \ldots, x_n$ are on level 1 (M^1), the grandchildren of the root are on level 2 (M^2), and so on, until the deepest level d ($M^{p=d}$) where the nodes do not have children. The *depth of the business model* d is defined as the total number of levels in M or the associated directed graph.

The business model graph of the business model represented in Table 1.1, is depicted in Figure 2.7. In general, fully-additive measures in the business model M can be associated with each cube C in the aggregation lattice, because measures are defined as functions on cubes. A business analysis in M from the LHS to the RHS of Definition 2.22 for some cube C, results in a "drill-down in the business model" from $y^{\mathbf{i}}(C)$ (M^0) to $x^{\mathbf{i}}(C)$ (M^1).

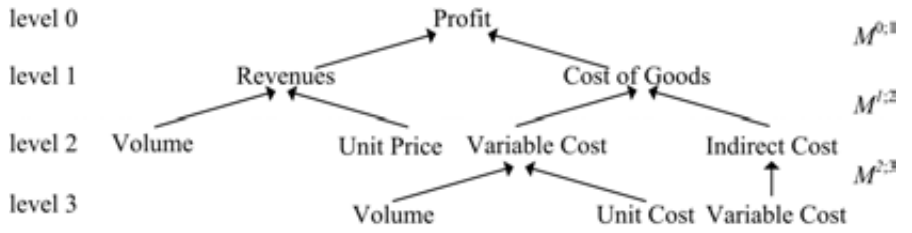


Figure 2.7: Business model graph $G(M)$ of measures in a financial database with depth $d = 3$.

In summary, various types of business analysis paths are possible in the OLAP database, involving a) only drill-down equations, b) only business model equations, or c) both. In a) the analysis is associated with a single lattice L and in b) the analysis is associated with a single business model M. In c), the situation of a mixed analysis, drill-down and business model equations are alternated. The result is a structure where multiple lattices are connected via the business model.

Example 2.3.3. In Figure 2.7, the analysis could start in the cell (2011, United States) of the cube C = Year $\times$ Country for the measure profit on level 233 in L. Subsequently, a drill-down equation (1), a business model equation (2), and a drill-down equation (3), are involved in the (mixed) analysis:

1. $\text{profit}^{233}(2011, \text{United States}) = \sum_{j=1}^{4} \text{profit}^{133}(2011.\text{Q}_j, \text{United States})$,

2. $\text{profit}^{133}(2011.\text{Q1}, \text{United States}) =$
   $\text{revenues}^{133}(2011.\text{Q1}, \text{United States}) - \text{costs}^{133}(2011.\text{Q1}, \text{United States})$,

3. $\text{revenues}^{133}(2011.\text{Q1}, \text{United States}) =$
   $\sum_{k=1}^{9} \text{revenues}^{123}(2011.\text{Q}_1, \text{United States.City}_k)$.

In the example, the analysis starts at $\text{profit}^{233}(2011, \text{United States})$ and ends at $\text{revenues}^{123}(2011.\text{Q}_1, \text{United States.City})$ in the lattice, via drill-down R^{-1}_{Time}, the first business model equation in Table 1.1, and drill-down R^{-1}_{Location}.

2.4 Related work

In addition to the star schema, many different formal notations and definitions of multi-dimensional data schemata are found in the literature (Kimball 1996; Agrawal et al. 1997; Gyssens and Lakshmanan 1997; Cabibbo and Torlone 1998; Lehner 1998; Vassiliadis 1998; Datta and Thomas 1999; Pedersen et al. 2001; Thalhammer et al. 2001; Caron and Daniels 2007; Kuznetsov and Kudryavtsev 2009; Ciferri et al. 2013). An in-depth comparison of multi-dimensional data models is provided by Vassiliadis and Sellis (1999), Pedersen et al. (2001), and Ciferri et al. (2013). Most of these models are developed for the design and technical implementation of multi-dimensional databases and not for the analysis of data cubes from a business user perspective, as in our case. The formal notations show a development from being purely focused on the description of technical database concepts to a focus on concepts that are important from a user analysis perspective. We particular introduce drill-down and business model equations, concepts which are absent in the other notations, for the purpose of diagnostic analysis.

2.5 Conclusion

In this chapter, we introduced a mathematical notation for the basic components of the multi-dimensional model: dimensions, dimension hierarchies, full cubes, subcubes, base cube, top cube, cells, and measures. The notation is coupled with navigational operators as roll-up, drill-down, slice, and dice.

In addition, we defined a structure in the multi-dimensional model, formed by the application of aggregation functions of a certain measure: the lattice structure of all aggregation levels L. The lattice L is formed by aggregating a measure y over all its associated dimensions and their dimension hierarchies in the data cube. In this lattice we defined the concepts: sublattice, upset, downset, and analysis path.

Lastly, we discussed two types of equations: drill-down equations for a single measure and relations between multiple measures. Drill-down equations are formed by the application of an aggregation function on a measure. Relations between measures are part of a business model M, representing, for example, financial or sales variables,

and relations between them.

These concepts lay the foundation for the research objectives in Chapter 1 and the results in the remainder of the thesis.

Chapter 3

Identification of exceptional values

3.1 Introduction

In this chapter, we consider the problem of finding exceptional cell values in multi-dimensional databases. In practice, multi-dimensional databases are often too large, in terms of the number of records in the fact table, and have too many dimensions and dimension hierarchies for business analysts to browse efficiently and effectively, and spot exceptional cells in the lattice of cubes manually. Notice that the number of cell contexts, see Equation (2.3), the number of cubes in the lattice, see Equation (2.5), and the number of lattice analysis paths, see Equation (2.2.3), grow exponentially fast when the number of dimensions and dimension hierarchies increase in the analysis. To deal with this, we develop a method and design an algorithm to detect exceptions automatically so that analysts can easily identify them, even when the data cube is very large.

This chapter is organised as follows. In the remainder of this section we introduce the topic of exception identification in multi-dimensional databases and list the basic concepts related to this topic. In particular, we introduce two specific classes of normative models: managerial and statistical models, that can be used in multi-dimensional databases for this purpose. In Section 3.2 we elaborate on various managerial models. In Section 3.3 we describe the general statistical model for OLAP exception identification and propose a statistical hypothesis test. Subsequently, in the next two sections we focus on two classes of statistical models. In Section 3.4 we discuss multi-way ANOVA models and in Section 3.5 we discuss contingency table

models. In Section 3.6 a general algorithm for exception identification is proposed for an n-dimensional data cube C. We briefly discuss related work on statistical outlier detection and outlier detection in OLAP databases in Section 3.7. Finally, in Section 3.8 we draw some conclusions.

3.1.1 Definition of exceptional values

Exception identification is a comparison activity by business analysts, based on the general diagnosis task, as depicted in Figure 1.2. The *actual cell data* $y^a(c)$ in some context cube C is compared with *reference cell data* $y^r(c)$ in the same cube in order to detect exceptions. The reference value for the cell is based on some *normative model*, which describes or predicts the reference values in C. The normative model specifies the appropriate *reference class* R which should be used to compare and the variables with respect to which the comparison should be made. The reference class R might describe, for example, the statistical normal case or the temporally normal case (Feelders and Daniels 2001). The reference object r represents one element from R.

The process of looking for exceptional cell values is equivalent to the process of looking for exceptional cell residuals, also known as problem identification or management by exception reporting (Judd et al. 1981). We now define a cell residual.

Definition 3.1. The *residual of a cell* $\partial y(c)$ in some context cube C is defined as the difference between its *actual value*, $y^a(c)$, and some *reference value* based on a normative model $y^r(c)$, i.e.,

$$\partial y(c) = y^a(c) - y^r(c).$$

Intuitively, an exception in a data cube is a cell with a value that is significantly different from the value we expected for this cell based on some normative model. The size of $\partial y(c)$ is the exception score for that cell. To determine the exceptions we have to apply a threshold to the exception scores. If the exception score is significant, i.e. larger than some threshold, it is viewed as an exceptional value that must be explained (see Chapter 4). The normative model under consideration and the domain knowledge related to the origin of the OLAP data, specify the appropriate *threshold* δ.

Definition 3.2. If the cell residual $\partial y(c) > \delta$, an exception score $\partial y(c) = high$ is added to the list of exceptional cells. Likewise, if the value of $\partial y(c) < -\delta$, an exception score $\partial y(c) = low$ is added. Otherwise, $\partial y(c) = normal$.

In this definition, and further in this thesis, the term exceptional cell denotes an exceptional cell value. The expression $\partial y(c) = y^a(c) - y^r(c) = q$ where $q \in \{low, normal, high\}$, specifies an event, i.e. a symptom, in the data cube. Notice that for the purpose of cell explanation (see Chapter 4), it is not interesting to explain events with the label $\partial y(c) = normal$, since it is only required to explain why a cell value deviates significantly from its reference value.

In conclusion, we combine the above definitions in an algorithm that identifies exceptions in an OLAP cube. Algorithm 1 lists the basic steps in the *exception identification process*.

Algorithm 1 Basic OLAP exception identification algorithm

Consider the cube C on some level $[i_1 i_2 \dots i_n]$ in the lattice L (see Chapter 2, Definition 2.6 and 2.15).

1. Compute/Determine the reference values $y^r(c)$ for all cells, based on some normative model R, to obtain $y^r(C)$.
2. Compute the residuals $\partial y(c)$ for all cells, as specified in Definition 3.1, to obtain $\partial y(C)$.
3. Compare the residual with the threshold values δ and $-\delta$, as specified in Definition 3.2, to determine the exceptional cells in the cube.
4. Mark the exceptional cells in the cube.

In diagnostic problem solving, the exception identification process is usually followed by an explanation process. This is described in Chapter 4.

3.1.2 Normative models

The normative behaviour in a multi-dimensional database, supporting business decision-making in a sales, financial or accountancy department, is usually defined by goals that have been formulated by the management. We will show that suitable

normative models can be incorporated in a multi-dimensional database, and applied as a reference class R. In this chapter we discuss two classes of normative models for exception identification in multi-dimensional databases:

1. R is a *managerial normative model*. In a study of Pounds (1969), it was found that managers use several types of managerial models to define their business goals:

   - *Planning and budget models*, the plan or determined budget is the expectation;
   - *Historical models*, expectation based on extrapolation of past experience and trends;
   - *Extra-organizational models*, models where expectations are derived from competition, customers, professional organizations, industry and branch averages, etc.

2. R is a *statistical normative model*. Decision-makers may also apply more abstract normative models in the form of statistical models. In this case the expected behaviour represents the statistically normal case (Feelders and Daniels 2001). We distinguish between two broad classes of statistical models that can be used in an OLAP database:

   - *Multi-way ANOVA models*, expectations for *continuous measures* are computed by multi-way ANOVA models;
   - *Contingency table models*, the expectations for *discrete measures* are computed by the independency model or the log-linear model.

Obviously, different normative models calculate the reference value and the threshold in different ways.

Furthermore, one can distinguish between *external* and *internal* normative models. External normative model are not directly available in the multi-dimensional database. These models first have to be stored in, or connected with, the multi-dimensional database to be applied as a reference object for exception identification. Planning, inter-organizational, and extra-organizational models, refer to norm values

that are derived from external sources (e.g. the planning system or the census office). Conversely, historical and statistical models are internal normative models. These models can be directly based on the data in the multi-dimensional database to form internal reference objects.

It is clear that the selection of the proper normative model in the OLAP context for which comparison should be made is fairly situation dependent. The choice for a particular normative model should be made by the management. Notice that it is not uncommon to apply multiple types of normative models for exception identification on the same data cube. Therefore, we have chosen to make the presentation of the normative model as general as possible, and to allow the model builder to specify and adapt the parameters of the selected normative model. In the next sections we discuss how the various normative models can be used in the OLAP context.

3.2 Managerial models

3.2.1 Planning and budget models

When the OLAP analyst is a firm's manager, the norm values may be the result of an explicit planning or budgetary control process. A significant difference between the firm's actual and planned performance will attract the attention of management, and will lead to the search for the underlying causes (Feelders 1993).

To apply planning and budget models in OLAP databases, the budgetary control process must determine reference values for all cells in some cube C, to obtain $y^r(C)$. For example with a simple budget model, the management might impose a budget decrease of 5% on the cube's actual results, then reference cells are computed straightforward with the formula $y^r(C) = 0.95 \cdot y^a(C)$. Moreover, when some planning or budget model is applied to all cubes in the (sub) lattice L, the budgetary control process must be as detailed as the values in the base cube C_b. Often it is desired that $y^r(C)$ is an additive measure, as specified in Definition 2.20. From a practical point this means that planning and budgetary information, from for example accountancy information systems, should be coupled with the ETL process and incorporated in

the star model. In this way, budget values are available for all cubes in the aggregation lattice by definition. For example in Table 3.1, the actual and budget figures are available for the cell (2001.Q2, department X) in the cube 2001.Quarter × department X on level 110 in the aggregation lattice for the measures in the business model relation: total costs$^{110}(C)$ = wages$^{110}(C)$ + travel$^{110}(C)$ + advertising costs$^{110}(C)$ + other costs$^{110}(C)$. The difference between the actual and budget is stored in the variance cell. When the firm operates a budgetary control system of management by exception, the attention of managers is focused on those departments and account items in the OLAP cube, when there is a significant variance from budget.

Table 3.1: Budget, actual and variance values for the financial variables of department X in Quarter 2 of the year 2001.

	Budget	Actual	Variance
wages^{110}(2001.Q2, department X)	11,100	13,100	2,000
travel^{110}(2001.Q2, department X)	3,100	3,700	600
advertising costs^{110}(2001.Q2, department X)	3,000	23,100	20,100
other costs^{110}(2001.Q2, department X)	800	3,100	2,300
total costs^{110}(2001.Q2, department X)	18,000	43,000	25,000

3.2.2 Extra/Intra-organizational models

The industry average of companies operating within the same industry or branch is often used as norm for the individual company in the area of *competition benchmarking* or *interfirm comparison* (IFC). By comparing the financial variables of a company with those of other companies, the company can assess its performance against objective standards and see where the company is strong or weak.

With respect to IFC in financial models a distinction is made between two types (Verkooijen 1993): (1) *ratio models* and (2) *nominal value models*. Ratio models fully consists of ratios between financial measures, whereas nominal value models consist of both ratios and pure nominal financial measures, such as inventory or cash. The business model equation from a financial cube: total assets turnover(C) = net sales(C)/total assets(C), is a typical example of measures in a ratio model. In this example, a financial analyst can compare some company with its branch

average, e.g. the cell total assets turnover(ABC-Company, 2008, Germany) might be compared with the cell Branch_AVG(total assets turnover(All-Companies, 2008, Germany)). The branch averages are computed by taking the average value of the business measures for a set of similar companies in the data set. Only ratio models are suitable to diagnose the financial results of two different firms, because the firm's size effect is eliminated by the ratios, which makes the ratios of different firms comparable. Normally, nominal value models can only be used when comparisons are made with previous recorded data of the same firm.

Similarly, as with the application of planning and budget models in the OLAP database, the information in the extra-organizational models must be on the base cube level. In Chapter 6, Section 6.2, an extra-organizational model is applied for interfirm comparison.

In addition, it is also possible to develop *inter-organizational normative models*. In such models a comparison is made with internal reference objects within the same company. The internal objects are based on the available dimensions of the database. For example, we might compare the results of business A unit with business unit B or we might compare the sales figures in different countries where the company is active, and so on.

3.2.3 Historical models

Here the norm value for a particular variable in the OLAP cube is its value in one or more previous time periods. Feelders (1993) notices that the number of previous time periods considered in the comparison should not be too large, because of the possibility of "structural changes", such as a shift in the macro-economic circumstances due to a financial crisis. Historical comparisons result in a judgement that the current period did better or worse than the previous period. Obviously, it does not enable one to say that "the judgement is good or bad in an absolute sense" (Feelders 1993). For example, it might be that a company has a declining profitability compared to last year, but that the branch on average is doing even worse.

There are many ways to construct historical reference objects in the OLAP database, because the Time dimension is nearly always present in the OLAP cube. The simplest way is manual *pairwise comparison* between two cells (Sarawagi 2001), where

the analyst selects an actual cell - representing the actual period - and a reference cell in the cube - representing the previous period - for comparison. In general, only the cells on the same aggregation levels will be used in the comparison task for obvious reasons, like the measurement scale of the variable. For example, the analysts could compare the actual cell profit^a(2011.Q1, Germany, Golf Equipment), in a financial cube from Example 2.1.1, with the profit of the first quarter in the previous year, the reference cell profit^r(2010.Q1, Germany, Golf Equipment), or with the profit in the previous quarter, the reference cell profit^r(2010.Q4, Germany, Golf Equipment). Obviously, it is also possible to compare the actual period with the average of previous periods, e.g. the actual cell in the latter example could be evaluated against AVG(profit^r(Previous Years.Q1, Germany, Golf Equipment)). Besides more complex historical reference object could be developed by time series models. How such models can be applied for regression analysis in OLAP databases is described by Chen et al. (2002). The choice for the application of a certain historical model is made by the analyst.

3.3 Statistical models

In multi-dimensional databases it is natural to use formal statistical models to automate, at least partly, exception detection. These models avoid subjective and error prone manual exception detection approaches in large data cubes. An exceptional value can also be defined as a large deviation of the expected value of the cell computed by a statistical model. A simple statistical model is given by the average value of cells in some context cube, computed over a single dimension D_q.

Definition 3.3. The average over the cells in $D_q^{i_q}$, denoted by $\bar{y}^{i_1 \ldots i_q \ldots i_n}(d_1, \ldots, \cdot, \ldots, d_n)$, in context cube $D_1^{i_1} \times \ldots \times D_q^{i_q} \times \ldots \times D_n^{i_n}$ is defined as

$$\bar{y}^{i_1 \ldots i_q \ldots i_n}(d_1, \ldots, \cdot, \ldots, d_n) = \frac{1}{J} \sum_{j=1}^{J} y^{i_1 \ldots i_q \ldots i_n}(d_1, \ldots, a_j, \ldots, d_n),$$

where $J = |D_q^{i_q}|$.

When a statistical model is used as a normative model, we usually write $\hat{y}(c)$ for $y^r(c)$. For convenience, we introduce a dot notation here: a dot $(\cdot)$ in place of a dimension

means averaging over that dimension. Notice that if y is an additive cube measure, the RHS of Definition 3.3 can be replaced by

$$\frac{1}{J}\sum_{j=1}^{J} y^{i_1\ldots i_q\ldots i_n}(d_1,\ldots,a_j,\ldots,d_n) = \frac{1}{J} y^{i_1\ldots(i_q+1)\ldots i_n}(d_1,\ldots,a,\ldots,d_n), \tag{3.1}$$

where $a \in D_q^{i_q+1}$ is a parent of $a_j \in D_q^{i_q}$. This follows directly from Definition 2.20. Moreover, Definition 3.3 can simply be generalized to multiple dimensions of the cell's context, to form a more sophisticated statistical model. For example, by computing the average of the cell measure over multiple associated dimensions.

Example 3.3.1. Consider a cell value $y^a(d_1, d_2, d_3)$ in a 3-dimensional base cube $C_B = D_1^0 \times D_2^0 \times D_3^0$. We now construct a reference cell value $\hat{y}^{000}(d_1, d_2, d_3)$ for the cell by averaging measures values over all dimensions in the base cube, as follows

$$\hat{y}^{000}(d_1, d_2, d_3) = \bar{y}^{000}(\cdot,\cdot,\cdot) = \frac{1}{JKL}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L} y^{000}(a_j, b_k, c_l). \tag{3.2}$$

More complex statistical models, as we shall see in Sections 3.4 and 3.5, take into consideration the position of the cell in the cube and the variation pattern over every dimension. For example, in a 3-dimensional cube the overall effect over the cube and the effects of rows, columns, and layers can be taken into account by more advanced models.

Definition 3.4. The *scaled residual* is defined as the normalization of $\partial y(c)$ by the standard deviation σ of the cell

$$s(c) = \frac{\partial y(c)}{\sigma},$$

where $\hat{y}(c)$ is computed with the statistical model applied to a certain context cube C of the cell and σ^2 is the variance in the same cube.

In statistical models it is assumed that the data cube is generated by some parametric distribution, for example, the Gaussian distribution. The parameters of the distribution are estimated from the given cube data. In Sections in Sections 3.4 and 3.5, we discuss how to estimate the parameters, such as the standard deviation σ in a cube C.

3.3.1 Statistical hypothesis test

We can now formulate a *statistical hypothesis test* to identify exceptions in a cube C. The null hypothesis (H_0) for such a test states that the actual data instance $y^a(c)$ has been generated from the estimated distribution. If the statistical test rejects H_0, $y^a(c)$ is declared to be an exception. A statistical hypothesis test is associated with a test statistic, which can be used to obtain an exception score for $y^a(c)$.

Here we use a straightforward test statistic, where we assume that $y^a(c)$ has a Gaussian distribution. All cells that are more than $\delta \cdot \sigma$ distance away from the distribution mean of all the cells in C are declared to be an exception. Typically, we select $\delta = 1.645$ (or 2.326) corresponding to a probability of 95% (or 99%) in the standard normal distribution. In general, the appropriate δ is based on the domain knowledge of the analyst, and therefore is user-defined. When this value is known, the software, as described in Chapter 6, can automatically determine the exceptional values in some cube based on a series of statistical tests.

Definition 3.5. For each cell residual, as specified in Definition 3.4, the following statistical tests are defined

$$\begin{cases} \text{if } \partial y(c)/\sigma > \delta \text{ (one-tailed test) then the cell is labelled } \partial y(c) = \text{“}\textit{high}\text{”}; \\ \text{if } \partial y(c)/\sigma < -\delta \text{ (one-tailed test) then the cell is labelled } \partial y(c) = \text{“}\textit{low}\text{”}; \\ \text{if } -\delta \leq \partial y(c)/\sigma \leq \delta \text{ (two-tailed test) then the cell is labelled } \partial y(c) = \text{“}\textit{normal}\text{”}. \end{cases}$$

Obviously, for the first two tests H_0 is rejected and for the last test H_0 is accepted.

3.3.2 General statistical model

Intuitively, an appropriate statistical model should capture the relation of a measure with its related dimensions and dimension hierarchies. A variety of appropriate statistical models exists for exception identification in two-way tables, three-way tables, four-way tables, etc., in the statistical literature; see, for example, Scheffé (1959) and Hoaglin, Mosteller, and Tukey (1988). Statistical problem identification in this thesis is mainly inspired by the work of Sarawagi et al. (1998) and the table analysis methods of Hoaglin et al. (1988) used in statistics. Later in this chapter we review three statistical models for problem identification in multi-dimensional databases, namely:

the multi-way ANOVA model for continuous data, the contingency model and log-linear model for category data. Naturally, these models calculate the distance to the reference value and the threshold in different ways.

Here we introduce the *general statistical model* without any concern for the type of measure y, discrete or continuous.

Definition 3.6. For a cell value $y^{i_1 i_2 \ldots i_n}(c)$ in the context cube C an *expected cell value* $\hat{y}^{i_1 i_2 \ldots i_n}(c)$ is defined as

$$\hat{y}^{i_1 i_2 \ldots i_n}(c) = f(C),$$

where f is some function defined in C.

The *general statistical model* for the cell is given by $y^{i_1 i_2 \ldots i_n}(c) = f(C) + \varepsilon(c)$, where $\varepsilon(c) = \partial y(c)$. The function f in the general statistical model can have any of the following functional forms:

- *Additive*: f returns the sum of its arguments. Models of this kind are called linear models or multi-way ANOVA models and are usually associated with *continuous data*. These models are appropriate for continuous measures (Definition 2.17), and are further discussed in Section 3.4;
- *Multiplicative*: f returns the product of its arguments. Models of this kind are the multinomial and the log-linear model and are usually associated with positive *discrete data*. These models are appropriate for discrete measures (Definition 2.17), and are further discussed in Section 3.5.

In the description of these models we postpone till Section 3.6 discussions regarding data transformation to improve the model fit and the checking of assumptions as constant variance, linearity and normality.

3.4 ANOVA models

The ANOVA model that can be used in OLAP databases with continuous measures, is the *multi-way ANOVA model with one observation per cell* (Scheffé 1959; Hoaglin,

Mosteller, and J. W. Tukey 1983). The dependent variable in the ANOVA model corresponds with a continuous numeric measure $y^{i_1 i_2 \ldots i_n} : C \to \mathbb{R}$, where a single data point is stored in the cube's cells, and the independent variables in the ANOVA model correspond with the cube's categorical dimensions $D_1, D_2, \ldots, D_n$. In this section we explain the application of multi-way ANOVA models for exception identification.

3.4.1 Main-effects ANOVA models

The first step in ANOVA model construction is usual to start with the *simple additive* or *main-effects model*, where the function f in Definition 3.6 is assumed to be additive. For the cube C, the expected value for a cell c, estimated with the simple additive model takes the following form

$$y^{i_1 i_2 \ldots i_n}(c) = \mu + \lambda_1(d_1) + \lambda_2(d_2) + \ldots + \lambda_n(d_n), \tag{3.3}$$

where μ is the *overall effect* in the whole context, $\lambda_1(d_1)$ is the *main effect* for dimension D_1, $\lambda_2(d_2)$ is the main effect for dimension D_2, and so on. This model has a simple interpretation because the separate contributions of the dimensions are just added together. This is consistent because it is assumed that there are no interactions between the dimensions in the context. Moreover, the usual assumption for this model is

$$\sum_{d_1}^{|D_1^{i_1}|} \lambda_1(d_1) = \sum_{d_2}^{|D_2^{i_2}|} \lambda_2(d_2) = \ldots = \sum_{d_n}^{|D_n^{i_n}|} \lambda_n(d_n) = 0. \tag{3.4}$$

This assumption states that the means for the different dimension instances are all equal. Additional assumptions for this model are the Gauss-Markov conditions as: statistical independence, normality, and equality of cell variances, thus $\varepsilon(c) \sim N(0, \sigma^2)$ (Scheffé 1959).

The coefficients of the model can be estimated by ordinary least-squares (OLS) (Scheffé 1959). The sum of squares of the residuals (SSR) to be minimized by OLS under the above assumptions is

$$SSR = \sum_{d_1}^{|D_1^{i_1}|} \sum_{d_2}^{|D_2^{i_2}|} \cdots \sum_{d_n}^{|D_n^{i_n}|} (y^{i_1 i_2 \ldots i_n}(d_1, d_2, \ldots, d_n) - \hat{y}^{i_1 i_2 \ldots i_n}(d_1, d_2, \ldots, d_n))^2. \tag{3.5}$$

The additive model that fits the data "best" is defined as one that determines estimates for μ, $\lambda_1(d_1)$, $\lambda_2(d_2)$, ..., $\lambda_n(d_n)$ so that the SSR is the smallest.

3.4.2 Full-effects ANOVA models

The simple additive model can be generalized to the *full-effects* ANOVA model, which is a model that includes degrees of freedom for non-additivity. This model can describe possible interaction between, for example, two dimensions D_1 and D_2 in the cube $D_1 \times D_2$. The interaction term $\lambda(d_1, d_2)$ is interpreted as that part of the main effect not captured in the additive effects of $\lambda_1(d_1)$ and $\lambda_2(d_2)$. For example, you may enjoy beer and nuts individually, but the combination is superior. In contrast, you may like beer and ice cream but not together.

In general, the expected value for a cell c, estimated with the full-effects model is

$$\begin{aligned} y^{i_1 i_2 \ldots i_n}(c) = \quad & \mu + \\ & \lambda_1(d_1) + \lambda_2(d_2) + \ldots + \lambda_n(d_n) + \\ & \lambda_{12}(d_1, d_2) + \lambda_{23}(d_2, d_3) + \ldots + \lambda_{(n-1)n}(d_{n-1}, d_n) + \\ & \lambda_{123}(d_1, d_2, d_3) + \lambda_{234}(d_2, d_3, d_4) + \ldots + \lambda_{(n-2)(n-1)n}(d_{n-2}, d_{n-1}, d_n) + \\ & \ldots + \ldots + \ldots + \ldots, \end{aligned} \tag{3.6}$$

where μ is the overall effect in the whole context; $\lambda_1(d_1)$, $\lambda_1(d_2)$, ..., $\lambda_n(d_n)$ are the main effects for dimension D_1, D_2, ..., D_n; $\lambda_{12}(d_1, d_2)$, $\lambda_{23}(d_2, d_3)$, ..., $\lambda_{(n-1)n}(d_{n-1}, d_n)$ are the *first-order effects* for each pair of dimensions $D_1 \times D_2$, $D_2 \times D_3$, ..., $D_{n-1} \times D_n$; $\lambda_{123}(d_1, d_2, d_3)$, $\lambda_{234}(d_2, d_3, d_4)$, ..., $\lambda_{(n-2)(n-1)n}(d_{n-2}, d_{n-1}, d_n)$ are the *second-order effects* for each triplet of dimensions $D_1 \times D_2 \times D_n$, $D_2 \times D_3 \times D_4$, ..., $D_{n-2} \times D_{n-1} \times D_n$; and so on for higher-order effects. In this model it is assumed that

$$\sum_{d_1}^{|D_1^{i_1}|} \lambda_1(d_1) = \sum_{d_2}^{|D_2^{i_2}|} \lambda_2(d_2) = \ldots = \sum_{d_n}^{|D_n^{i_n}|} \lambda_n(d_n) =$$

$$\sum_{d_1}^{|D_1^{i_1}|} \sum_{d_2}^{|D_2^{i_2}|} \lambda_{12}(d_1, d_2) = \sum_{d_2}^{|D_2^{i_2}|} \sum_{d_3}^{|D_3^{i_3}|} \lambda_{23}(d_2, d_3) = \ldots = \sum_{d_{n-1}}^{|D_{n-1}^{i_{n-1}}|} \sum_{d_n}^{|D_n^{i_n}|} \lambda_{(n-1)n}(d_{n-1}, d_n) =$$

$$= \ldots = 0.$$

In addition, the Gauss-Markov conditions are also assumed to hold for the full-effects model. For investigating non-additivity (interaction) in Equation (3.6), we apply

a generalization of Tukey's test of additivity Tukey (1949) to all first-order effects, second-order effects, etc. For details on this test, we refer to Scheffé (1959).

The OLS estimates for all the model coefficients in Equation (3.3) and Equation (3.6) are found by minimizing the SSR. The following *mean-based estimates*, generalized from Hoaglin et al. (1983) and Scheffé (1959), yield the OLS estimates that minimize the SSR. The estimate for the overall-effect is

$$\hat{\mu} = \bar{y}(\cdot, \cdot, \ldots, \cdot). \tag{3.7}$$

The estimates for the main-effects are

$$\begin{aligned} \hat{\lambda}_1(d_1) &= \bar{y}(d_1, \cdot, \ldots, \cdot) - \hat{\mu}, \\ \hat{\lambda}_2(d_2) &= \bar{y}(\cdot, d_2, \cdot, \ldots, \cdot) - \hat{\mu}, \\ &\ldots, \\ \hat{\lambda}_n(d_n) &= \bar{y}(\cdot, \ldots, \cdot, d_n) - \hat{\mu}. \end{aligned} \tag{3.8}$$

The estimates for the first-order effects are

$$\begin{aligned} \hat{\lambda}_{12}(d_1, d_2) &= \bar{y}(d_1, d_2, \cdot, \ldots, \cdot) - \hat{\lambda}_1(d_1) - \hat{\lambda}_2(d_2) - \hat{\mu}, \\ \hat{\lambda}_{23}(d_2, d_3) &= \bar{y}(\cdot, d_2, d_3, \cdot, \ldots, \cdot) - \hat{\lambda}_2(d_2) - \hat{\lambda}_3(d_3) - \hat{\mu}, \\ &\ldots, \\ \hat{\lambda}_{(n-1)n}(d_{n-1}, d_n) &= \bar{y}(\cdot, \ldots, \cdot, d_{n-1}, d_n) - \hat{\lambda}_{n-1}(d_{n-1}) - \hat{\lambda}_n(d_n) - \hat{\mu}. \end{aligned} \tag{3.9}$$

And so on for the estimation of the higher-order effects.

In Equation (3.7) $\hat{\mu}$ represents the overall mean in the whole cube C.

Moreover, when $y^{i_1 i_2 \ldots i_n}(C)$ is a fully-additive measure (Definition 2.20), the estimates for the coefficients in Equations (3.7), (3.8), (3.9), and so on, are directly available in the various cubes of the lattice L. Similarly, the mean-based estimates are determined in L as follows (where $j \neq k$)

$$\begin{aligned} \bar{y}(\cdot, \cdot, \ldots, \cdot) &= \tfrac{1}{|C|} \cdot y(C_T) \\ \bar{y}(\cdot, \ldots, \cdot, d_j, \cdot, \ldots, \cdot) &= \tfrac{1}{|R^{+1}_{D_j}(C)|} \cdot y(R^{-1}_{D_j}(C_T)) \\ \bar{y}(\cdot, \ldots, \cdot, d_j, \cdot, \ldots, \cdot, d_k, \cdot, \ldots, \cdot) &= \tfrac{1}{|R^{+1}_{D_j} \circ R^{+1}_{D_k}(C)|} \cdot y(R^{-1}_{D_j} \circ R^{-1}_{D_k}(C_T)) \\ \ldots &\quad . \end{aligned} \tag{3.10}$$

The mechanism behind these formulas can be understood by defining the concept of a complement cube. The *complement cube* $\overline{C} = [j_1, j_2, \ldots, j_n]$ of a cube $C = [i_1, i_2, \ldots, i_n]$ in the lattice L is defined as $[j_1, j_2, \ldots, j_n] = [\max_1 - i_1, \max_2 - i_2, \ldots,$

$\max_n - i_n]$. Every cube C has an unique complement cube $\overline{C}$. In Equation (3.10), C_T is the complement of $C = C_B$, $R^{-1}_{D_j}(C_T)$ is the complement of $R^{+1}_{D_j}(C)$, $R^{-1}_{D_j} \circ R^{-1}_{D_k}(C_T)$ is the complement of $R^{+1}_{D_j} \circ R^{+1}_{D_k}(C)$, and so forth. The general idea in Equation (3.10) is that in the RHS of the equations the total of some complement cube $y(\overline{C})$, is divided by the cells of the cube $|C|$, to obtain the average value for some cell c'

$$\bar{y}^{i_1 i_2 \ldots i_n}(c') = \frac{1}{|C|} \cdot y(\overline{C}),$$

where c' is a cell with one or more averaged values.

3.4.3 Standard deviation, quality of fit, and significance of effects

After fitting a multi-way ANOVA model and obtaining the cell residuals, we need to scale them (Definition 3.4), where the standard deviation of each cell in the cube is required for the computation. The general assumption in ANOVA models is the assumption of equal variances within the cells in one-way or higher table layouts (Scheffé 1959). Therefore, we assume that $\sigma^2(C) = \sigma^2(c)$.

Suppose that $s^2(C)$ denotes the sample variance of a random sample from a cube C with variance σ^2. The sample variance for a cell in a cube C is a generalization of Scheffé (1959), and is given by

$$s^2(c) = \frac{\sum_{d_1=1}^{|D_1^{i_1}|} \sum_{d_2=1}^{|D_2^{i_2}|} \cdots \sum_{d_n=1}^{|D_n^{i_n}|} (y(d_1, d_2, \ldots, d_n) - \hat{y}(d_1, d_2, \ldots, d_n))^2}{(|D_1^{i_1}||D_2^{i_2}| \cdots |D_n^{i_n}| - 1)}. \tag{3.11}$$

Then $E(s^2(C)) = \sigma^2(C)$. In words, the variance $s^2(C)$ is estimated as the SSR divided by approximately the number of cells, i.e. the degrees of freedom, in the cube.

A measure for the size of residuals is the SSR. The multi-way ANOVA model uses the "fraction of the sum of squared variation explained by the fit" to judge the quality of the fit of the ANOVA model. The fraction may be written as (Snedecor and C.Cochran 1980)

$$R^2 = 1 - \frac{SSR}{\sum_{d_1} \sum_{d_2} \cdots \sum_{d_n} (y(d_1, d_2, \ldots, d_n) - \hat{\mu})^2}. \tag{3.12}$$

This expression, using the estimated mean of the data $\hat{\mu} = \bar{y}(\cdot, \cdot, \ldots, \cdot)$, arises naturally in a least-squares framework.

For both the main-effects and full-effects ANOVA model it has to be verified whether the effects are significant and therefore should be included in the ANOVA model or not. For each effect in the model we have to examine the matching hypothesis from the list:

$H_{0;D_i}$:	There are no main effects for dimension D_i, i.e. $\lambda_i(d_i) = 0$ for all d_i;
$H_{0;D_iD_j}$:	There are no first order effects between pairs of dimensions $D_i \times D_j$, i.e. $\lambda_{ij}(d_i, d_j) = 0$ for all d_i and d_j;
$H_{0;D_iD_jD_k}$:	There are no second order effects between triplets of dimensions $D_i \times D_j \times D_k$, i.e. $\lambda_{ijk}(d_i, d_j, d_k) = 0$ for all d_i, d_j and d_k;
$H_{0;\ldots}$:	....

For each null hypothesis that is rejected we accept the presence of the (main, first order, second order, etc.) effect and include this effect in the ANOVA model. For a cube $C = [i_1 i_2 \ldots i_n]$, the total number of hypotheses that is tested is equal to the number of cubes in its upset $\{\uparrow C\}$. For example, for a cube with 3 dimensions without hierarchies, we can formulate 7 hypothesis, given by $H_{0;D_1}$, $H_{0;D_2}$, $H_{0;D_3}$, $H_{0;D_1D_2}$, $H_{0;D_2D_3}$, $H_{0;D_1D_3}$, and $H_{0;D_1D_2D_3}$. These hypothesis are tested with the standard F-test. An F-test is a statistical test in which the test statistic has an F-distribution under the null hypothesis. For $H_{0;D_i}$ the test statistic is given by

$$f = \frac{s_b^2(D_i)}{s_w^2(\epsilon)} = \frac{\frac{\mathrm{SS}(D_i)}{|D_i|-1}}{\frac{\mathrm{SS}(\epsilon)}{(|D_1|-1)(|D_2|-1)\ldots(|D_n|-1)}}, \tag{3.13}$$

where s_b^2 is the variance between dimensions, s_w^2 the variance within dimensions, $\mathrm{SS}(D_i)$ is the sum of squares of dimension D_i, and $\mathrm{SS}(\epsilon)$ is the sum of squares of the residuals. This test statistic is compared with the F-distribution with $[(|D_i| - 1), (|D_1| - 1)(|D_2| - 1) \ldots (|D_n| - 1)]$ degrees of freedom at a certain significance level α. The F-test is formulated by the equality $\Pr\{f > F_{\mathrm{df};\alpha}\} = \alpha$. The null hypothesis is rejected at level α if $f > F_{\mathrm{df};\alpha}$. Notice that the test statistic needs to be adapted for a null hypothesis with higher-order effects.

3.4.4 Example

We investigate the measure revenues on the cube $C = 1997.\text{Month} \times \text{Products}$, obtained from the Foodmart data warehouse[1], a realistic sales database from an American supermarket. In this example we scale the data, with the natural logarithms as $y(C) = \log(\text{revenues}(C))$. The data in $\{\uparrow C\}$ for the measure y is created by summarizing $y(C)$, resulting in an additive measure (Definition 2.20). With Algorithm 1 we identify exceptional values and compute $y^r(C)$ with the following additive ANOVA models

1. $\hat{y}^{00}(\text{Month, Products}) = \hat{\mu}$;
2. $\hat{y}^{00}(\text{Month, Products}) = \hat{\mu} + \hat{\lambda}_1(\text{Month})$;
3. $\hat{y}^{00}(\text{Month, Products}) = \hat{\mu} + \hat{\lambda}_2(\text{Products})$;
4. $\hat{y}^{00}(\text{Month, Products}) = \hat{\mu} + \hat{\lambda}_1(\text{Month}) + \hat{\lambda}_2(\text{Products})$.

In this example we refrain from testing the significance of effects with the F-test, to show the outcomes of various ANOVA models, even from models that might omit significant effects. This is done to examine the model's effect on the sets of exceptional cells that are identified. The estimates for an arbitrary cell $y^{00}(\text{Month}, \text{Products})$ in C obtained by model 4 is given by

$$\begin{aligned}
&\hat{\mu} = \bar{y}^{00}(\cdot,\cdot),\\
&\hat{\lambda}_1(\text{Month}) = \bar{y}^{01}(\text{Month},\cdot) - \bar{y}^{00}(\cdot,\cdot),\\
&\hat{\lambda}_2(\text{Products}) = \bar{y}^{10}(\cdot,\text{Products}) - \bar{y}^{00}(\cdot,\cdot).
\end{aligned}$$

If the estimates are plugged into the model, we derive

$$\hat{y}^{00}(\text{Month}, \text{Products}) = \bar{y}^{01}(\text{Month},\cdot) + y^{10}(\cdot,\text{Products}) - \bar{y}^{00}(\cdot,\cdot).$$

The various means in the above model can now be obtained by using Equation (3.10)

$$\begin{aligned}
\bar{y}^{00}(\cdot,\cdot) &= \tfrac{1}{|C|} \cdot y^{11}(C_T),\\
\bar{y}^{01}(\text{Month},\cdot) &= \tfrac{1}{|R^{+1}_{D_1}(C)|} \cdot y^{01}(R^{-1}_{D_1}(C_T)),\\
\bar{y}^{10}(\cdot,\text{Products}) &= \tfrac{1}{|R^{+1}_{D_2}(C)|} \cdot y^{10}(R^{-1}_{D_2}(C_T)).
\end{aligned}$$

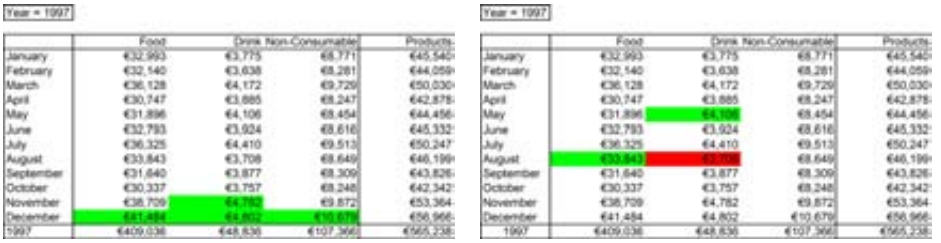
[1] Available from http://www.emielcaron.nl

In exception identification we take $\delta = 1.645$ ($p = .95$) as a threshold value. If we apply the first and the third ANOVA model, no exceptional cells in the cube are found. If we apply model 2, the following cells are labelled as high exceptions (November, Drink), (December, Food), (December, Drink), and (December, Non-Consumable), see Figure 3.1 a). And if we apply model 4, the exceptional cells with label high are (May, Drink) and (August, Food) and the exceptional cell with label low is (August, Drink), see Figure 3.1 b). For example in model 4, we have the following data for cell c = (August, Drink): $y^a(c) = 8.218 = \ln(3,708)$, $\hat{y}(c) = 8.266$, and $\partial y(c) = -0.048$. $\partial y(c)$ is scaled with $\sigma(C) = 0.0225$, computed with $(12 \cdot 3) - 1 = 35$ d.f., to produce the scaled residual $\partial y(c)/0.0225 = -2.142$. If we compare the scaled residual with the threshold, $-2.142 < -1.645$, the cell c is labelled as a low exception.

In Figure 3.1 a), the revenues figures for products in the Month December are identified as exceptional high with model 2. This reference model only includes the overall effect μ and the month-effect λ_1(Month). The economic explanation behind the exceptions is given by additional revenues related to increased sales in the Christmas period, which occur every year in December. If also the product-effect λ_2(Product) is included in the reference model, as in model 4, a different set of exceptional cells in identified, as depicted in Figure 3.1 b). Here the month-effect is weakened and the combined effects shows that the cells (May, Drink), (August, Food), and (August, Drink) are remarkable. For example, the relatively low revenues in the cell (August, Drink) might be explained by low temperatures in that month, resulting in low revenues compared to other summer months and low revenues compared to other product categories. We conclude that the choice for a particular statistical reference model might result in different sets of exceptional values.

3.5 Contingency table models

The contingency table models that can be used in OLAP databases for measures with discrete values, given by $y^{i_1 i_2 \ldots i_n} : C \to \mathbb{N}$, are the *multinomial* and the *log-linear models* for multi-dimensional tables. These models were proposed by statisticians to model contingency or frequency tables where the cell entries are positive discrete values, i.e. count data. Detailed descriptions of these statistical models and their



a) b)

Figure 3.1: Identification of exceptions in the example cube 1997.Month × Products with ANOVA model $\hat{y}$(Month, Products)= $\mu + \lambda_1$(Month) (a) and model $\hat{y}$(Month, Products)= $\mu + \lambda_1$(Month) + λ_2(Products) (b). The color green indicates a high exception and the color red a low exception in the figure.

analysis are given in, for example, in Bishop, Fienberg, and Holland (1975) and Everitt (1994). In this section we generalize these contingency models for the common two-way tables to multi-way tables, i.e. the n-dimensional cube C.

3.5.1 Multinomial models for contingency tables

The general form of the n-dimensional contingency table is a positive discrete measure, classified with respect to n qualitative variables (dimensions) $D_1^{i_1}$, $D_2^{i_2}$, ..., $D_n^{i_n}$, similar to a cube $y^{i_1 i_2 \ldots i_n} : C \to \mathbb{N}$. The count in the a_j category of dimension d_1, the b_k category of dimension variable d_2, ..., and the z_q category of dimension variables d_n, that is the frequency in the $(d_1^{i_1}, d_2^{i_2}, \ldots, d_n^{i_n})$-th cell of the cube, is represented by $y(a_j, b_k, \ldots, z_q)$. The total number of observations in the a_jth category of d_1 is denoted by $y(a_j, +, \ldots, +)$ and the total number of observations in the b_kth category of d_2 is denoted by $y(+, b_k, \ldots, +)$, and so on. These are known as *marginal totals*. $N = y(+, +, \ldots, +)$ represents the overall total of cell values in the cube C.

Now suppose that each of the N cell values is classified independently in one of the cells of the cube C, i.e. a multi-way table, and suppose that the probability that an observation falls in the $(d_1, d_2, \ldots, d_n)$-th cell is $p(d_1, d_2, \ldots, d_n)$. Let $Y(D_1, D_2, \ldots, D_n)$ denote a random variable representing the number of cell values in dimensions D_1, D_2, ..., D_n of the cube, and let $y(d_1, d_2, \ldots, d_n)$ denote the actual

observed cell frequency. It is often assumed that the actual cell frequencies follow a multinomial or Poisson distribution, with probability values $p(d_1, d_2, \ldots, d_n)$, where $\sum_{d_1}^{|D_1|} \sum_{d_2}^{|D_2|} \ldots \sum_{d_n}^{|D_n|} p(d_1, d_2, \ldots, d_n) = 1$, see for more detailed information Everitt (1994).

In general, the most important question in the analysis of contingency cubes (tables) is whether the dimensions forming the cube are independent or not. From the multiplication law of probability, independence between dimensions, implies that: $p(c) = p(d_1, +, \ldots, +)p(+, d_2, \ldots, +) \cdots p(+, +, \ldots, d_n)$. Therefore, the hypothesis of mutual independence of the dimensions $D_1, D_2, \ldots, D_n$, where we assume that the actual cell values $y(d_1, d_2, \ldots, d_n)$ follow a multinomial distribution, is formulated as

$$H_0 : p(c) = p(d_1, +, \ldots, +)p(+, d_2, \ldots, +) \cdots p(+, +, \ldots, d_n), \tag{3.14}$$

where $p(c)$ represent the probability of a cell value being in the cell c of the cube C, and $p(d_1, +, \ldots, +)$, $p(+, d_2, \ldots, +)$, …, $p(+, +, \ldots, d_n)$, are the marginal probabilities of dimension D_1, D_2, and so on. The question now is how to test the hypothesis and how to estimate the probabilities. Therefore, estimates of the frequencies to be expected when H_0 is true have to be computed. In the case that Hypothesis (3.14) holds, the expected value for an actual cell $y^{i_1 i_2 \ldots i_n}(c)$ in the context cube C, is given by the multinomial model

$$\hat{y}^{i_1 i_2 \ldots i_n}(c) = N\hat{p}(d_1, +, \ldots, +)\hat{p}(+, d_2, \ldots, +) \cdots \hat{p}(+, +, \ldots, d_n), \tag{3.15}$$

where $\hat{p}(d_1, +, \ldots, +)$, $\hat{p}(+, d_2, \ldots, +)$, …, $\hat{p}(+, +, \ldots, d_n)$ are estimates of the corresponding probabilities. It can be shown that the best estimates are derived from the marginal totals of the cube's dimensions, namely

$$\begin{aligned} \hat{p}(d_1, +, \ldots, +) &= \tfrac{y(d_1, +, \ldots, +)}{N}, \\ \hat{p}(+, d_2, \ldots, +) &= \tfrac{y(+, d_2, \ldots, +)}{N}, \\ &\cdots, \\ \hat{p}(+, +, \ldots, d_n) &= \tfrac{y(+, +, \ldots, d_n)}{N}. \end{aligned} \tag{3.16}$$

Note that these are the maximum likelihood estimates (Bishop, Fienberg, and Holland 1975). In a cube $C = [i_1 i_2 \ldots i_n]$ in L, the estimates of the marginal probabilities in

Equation (3.16) are obtained by the following operations

$$\begin{array}{c} N = y(R^{+1}_{D_1} \circ R^{+1}_{D_2} \circ \ldots \circ R^{+1}_{D_n}(C)), \\ y(d_1, +, \ldots, +) = y(R^{+1}_{D_2} \circ R^{+1}_{D_3} \circ \ldots \circ R^{+1}_{D_n}(C)), \\ y(+, d_2, \ldots, +) = y(R^{+1}_{D_1} \circ R^{+1}_{D_3} \circ \ldots \circ R^{+1}_{D_n}(C)), \\ \cdots, \\ y(+, +, \ldots, d_n) = y(R^{+1}_{D_1} \circ R^{+1}_{D_2} \circ \ldots \circ R^{+1}_{D_{n-1}}(C)). \end{array} \tag{3.17}$$

3.5.2 Log-linear models for contingency tables

An alternative model for contingency table data is the log-linear model, where the actual values are treated as realizations of independent Poisson random variables with probability values $p(c)$. Equation (3.15) specifies a multiplicative model for the data. However, it is also possible to rearrange this model so that $p(c)$ can be expressed as the sum of the marginal probabilities. By taking the natural logarithms of Equation (3.15), the model is rearranged to

$$\log y(c) = \\ \log N + \log p(d_1, +, \ldots, +) + \log p(+, d_2, \ldots, +) + \cdots + \log p(+, +, \ldots, d_n).$$

The model of complete independence can now be rewritten in a form equivalent with the main-effects ANOVA model (Equation 3.3) (Bishop, Fienberg, and Holland 1975). Accordingly, the logarithm of the expected value for a cell $\hat{y}^{i_1 i_2 \ldots i_n}(c)$ in the context cube C, estimated with the log-linear model under the assumption of mutual independence, is given by

$$\log y^{i_1 i_2 \ldots i_n}(c) = \mu + \sum_{j=1}^{n} \lambda_j(d_j), \tag{3.18}$$

where μ is the overall-effect and the $\lambda_j(d_j)$'s are the main-effects for each dimension D_j. This model is known as the *Poisson additive model* or the *main-effects log-linear model*. The main-effect parameters of this model are measured as deviations from the dimension means of the log-frequencies from the overall mean, and it is assumed that Equation (3.4) holds for all dimension means.

What is of interest, is to extend the main-effect log-linear model (Equation (3.18)), to the common situation where the dimensions of the cube C cannot be assumed to be (completely) independent. To do this, extra terms representing interactions between

the dimensions are introduced into Equation (3.18), resulting in the *complete log-linear model* (Bishop et al. 1975; Hoaglin et al. 1988; Everitt 1994). This model is given by

$$y^{i_1 i_2 \ldots i_n}(c) = \gamma + \prod_{j=1}^{n} \gamma(d_j) + \prod_{j \neq k}^{n} \gamma(d_j, d_k) + \prod_{j \neq k \neq l}^{n} \gamma(d_j, d_k, d_l) + \ldots \ , \tag{3.19}$$

where γ is the overall-effect, the $\gamma(d_j)$'s are the contributions from each individual dimension, the $\gamma(d_j, d_k)$'s are the interactions among two dimensions, the $\gamma(d_j, d_k, d_l)$'s are the interactions among three dimensions, and so on for the higher-order effects. The multiplicative form can simply be transformed into a linear additive form by taking the log of the original data values giving

$$\begin{gathered} l(c) = \log y(c) = \\ \log N + \sum_{j=1}^{n} \log \gamma_j(d_j) + \sum_{j \neq k}^{n} \log \gamma_{jk}(d_j, d_k) + \sum_{j \neq k \neq l}^{n} \log \gamma_{jkl}(d_j, d_k, d_l) + \ldots = \\ \mu + \sum_{j=1}^{n} \lambda_j(d_j) + \sum_{j \neq k}^{n} \lambda_{jk}(d_j, d_k) + \sum_{j \neq k \neq l}^{n} \lambda_{jkl}(d_j, d_k, d_l) + \ldots \ . \end{gathered} \tag{3.20}$$

Estimates of the parameters in the log-linear models are obtained as a function of the logarithm of $\hat{y}(c)$ and the form of such estimates is very similar to those used for the parameters in ANOVA models. Setting $l(d_1, d_2, \ldots, d_n) = \log y(d_1, d_2, \ldots, d_n)$ and again adopting the 'bar' and 'dot' notation for means, that is

$$\bar{l}(\cdot, \cdot, \ldots, \cdot) = \frac{1}{|D_1||D_2| \cdots |D_n|} \sum_{d_1}^{|D_1|} \sum_{d_2}^{|D_2|} \cdots \sum_{d_n}^{|D_n|} \log y(d_1, d_2, \ldots, d_n), \text{ etc.}$$

Then the following mean-based estimates, written in the form taken by parameter estimates in multi-way ANOVA, yield the following estimates for the main-effects log-linear model (Equation (3.18) and the complete log-linear model (Equation 3.20). The estimate for the overall-effect is given by

$$\hat{\mu} = \bar{l}(\cdot, \cdot, \ldots, \cdot). \tag{3.21}$$

The estimates for the main-effects are given by

$$\begin{aligned} \hat{\lambda}_1(d_1) &= \ \bar{l}(d_1, \cdot, \ldots, \cdot) - \hat{\mu}, \\ \hat{\lambda}_2(d_2) &= \ \bar{l}(\cdot, d_2, \cdot, \ldots, \cdot) - \hat{\mu}, \\ &\ldots, \\ \hat{\lambda}_n(d_n) &= \ \bar{l}(\cdot, \ldots, \cdot, d_n) - \hat{\mu}. \end{aligned} \tag{3.22}$$

The estimates for the first-order effects are given by

$$\begin{aligned}
\hat{\lambda}_{12}(d_1, d_2) &= \bar{l}(d_1, d_2, \cdot, \ldots, \cdot) - \hat{\lambda}_1(d_1) - \hat{\lambda}_2(d_2) - \hat{\mu}, \\
\hat{\lambda}_{23}(d_2, d_3) &= \bar{l}(\cdot, d_2, d_3, \cdot, \ldots, \cdot) - \hat{\lambda}_2(d_2) - \hat{\lambda}_3(d_3) - \hat{\mu}, \\
&\ldots, \\
\hat{\lambda}_{(n-1)n}(d_{n-1}, d_n) &= \bar{l}(\cdot, \ldots, \cdot, d_{n-1}, d_n) - \hat{\lambda}_{n-1}(d_{n-1}) - \hat{\lambda}_n(d_n) - \hat{\mu}.
\end{aligned} \tag{3.23}$$

And so forth for the estimation of the higher-order effects.

Notice that due to the logarithmic form of Equation (3.18) and Equation (3.20) the reference measure $y^r(C)$ is not fully-additive (Definition 2.20). For example, $\log \bar{y}(\cdot, \cdot, \ldots, \cdot) \neq \frac{1}{|C|} \cdot \log y(C_T)$. Therefore, we cannot use the computations as formulated in Equation (3.10). In other words, in the estimation of the parameters of the log-linear models, we cannot (re-)use the other cubes in L, for each cube C under consideration we have to estimate the parameters in Equations (3.21), (3.22), and (3.23) separately. Obviously, such computations are computationally more demanding than the parameter estimations in multi-way ANOVA models.

Notice that the expected values corresponding to some deviant log-linear models cannot be obtained directly from particular marginal totals of the actual cell values. This is so because in such cases the maximum likelihood equations have no explicit solution. In these situations, the expected values are obtained alternatively, for example, by the algorithmic method of iterative proportional fitting (Bishop et al. 1975).

For correct application of the multinomial model (Equation (3.15)) and the log-linear models (Equations (3.18) and (3.20)), we need to test for independence between sets of dimensions, i.e. we need to test the truth of Hypothesis (3.14). This test is based upon comparing the actual cell values $y(c)$ with the estimated cell values $\hat{y}(c)$ in some cube $y(C)$, under a particular hypothesis of independence. Two well-known tests are the Pearson X^2 statistic and the likelihood ratio statistic X^2_L; we refer to Everitt (1994) for details on these test statistics. Both statistics follow approximately a *chi-square distribution* when the hypothesis tested is true (Bishop et al. 1975). Testing the hypothesis of independence is performed by comparing the calculated X^2 with the values in the chi-square distribution, with some significance level α, often some low probability value of $\alpha = .05$ or $\alpha = .01$. Notice that the degrees of freedom the test statistic, X^2, depend upon on the number of instances of each dimension forming

the cube. A straightforward way of determining the degrees of freedom of the X^2 statistic for cell values in a multi-dimensional cube C is by the use of the formula (Everitt 1994)

$$\text{d.f.} = (|C| - 1) - (|D_1^{i_1}| - 1) - (|D_2^{i_2}| - 1) - \ldots - (|D_n^{i_n}| - 1).$$

In general, these test statistics can also be used as goodness-of-fit criteria.

In the analysis of residuals in cubes with a discrete measure, it is often not appropriate to scale the cell residuals with an estimate of the standard deviation as in cubes with a continuous measure. Hence, for a cell c in a cube $y^{i_1 i_2 \ldots i_n}(C) \rightarrow \mathbb{N}$ the scaled residuals (Everitt 1994), adapted for multiple dimensions, can be used. The use of scaled residuals for examination of a contingency table may often give conservative indications of cells having lack of fit. A more precise analysis of the residuals is proposed by Haberman (1973), by means of adjusted residuals. When the dimensions forming the cube are independent, these adjusted residuals are approximately normally distributed with mean zero and standard deviation one. Moreover, in Sarawagi et al. (1998) an alternative method for scaling the residuals is proposed.

3.6 Algorithm for statistical exception identification

In this section a *general algorithm for statistical exception identification* in multi-dimensional databases is presented. The algorithm can be adapted for both multi-way ANOVA models to handle continuous measures, and contingency table models to handle positive discrete measures. The input of the algorithm is a cube $C = [i_1 i_2 \ldots i_n]$ and its upset $\{\uparrow C\}$ with measure values $y(C)$ somewhere in the lattice L. The output of the algorithm is a set of exceptions if any. The basic steps of the algorithm are listed in Algorithm 2.

In the data transformation step (Step 1), the analyst might decide to transform the measure values $y^{i_1 i_2 \ldots i_n}(C)$, to create a common measurement scale if desired, by some appropriate scaling operator SCALE($y(C)$). Transformations of the measure values might improve the fit of the statistical model and correct for violations of model assumptions. Typically, the natural logarithms are taken of the cube's cell

Algorithm 2 Statistical exception identification algorithm

1. *Data transformation*;
2. *Statistical modeling*;
3. *Diagnostics*;
   (a) Test for the significance of model effects;
   (b) Test for the presence of interaction effects;
   (c) Test for the normality of residuals;
   (d) Test for the homogeneity of variance;
4. *Exception identification.*

values by $\log(y(C))$. Obviously, in the application of the log-linear model (Equation (3.18)), the measure values need to be scaled by taken the natural logarithm, by definition. Next, the empty cells $N_{\text{empty\_cells}}$ in the cube C are determined, and the appropriate method to deal with the incomplete data cube is selected. Section 3.6.2 presents more details.

In the statistical modeling step (Step 2), we execute the first three steps of the basic exception identification algorithm (Algorithm 1), where the normative model is an advanced statistical model. If y is a continuous measure, R is selected to be a multi-way ANOVA model in the form of a simple additive model. See Equation (3.3), or a full-effects model, see Equation (3.6), to identify exceptional cells. The coefficients of those models are estimated with Equations (3.7), (3.8), and (3.9), and so on. Subsequently, the expected values $\hat{y}(C)$ and cell residuals $\partial y(C)$ are computed. Finally, the variance $\sigma^2(C)$ is estimated with Equation (3.11), and the scaled residuals $s(C)$ are determined. Furthermore, if y is a counting measure, R is selected to be a contingency table model in the form of a complete independency model, see Equation (3.15), or a log-linear model, see Equation (3.20). The model coefficients for the independency model are estimated with Equation (3.16), and the model coefficients for the log-linear model with Equations (3.21), (3.22), and (3.23), and so on. Next the expected values $\hat{y}(C)$ and cell residuals $\partial y(C)$ are computed. Finally, the cell

residuals are standardized to obtain the scaled or adjusted residuals $s(C)$ (Haberman 1973; Everitt 1994).

In the diagnostics step (Step 3), a series of statistical tests is performed automatically or guided by the analyst, to determine the quality of the statistical model's fit (a) and to check the statistical model's assumptions as independence (b), normality (c), and homoscedasticity (d). Firstly, for multi-way ANOVA models we determine the quality of the model with Equation (3.12) and we check whether the dimension effects are significant or not, by using an F-test under the appropriate degrees of freedom (see Equation 3.13). Secondly, the assumption of independence between the cube's dimensions is tested formally with Tukey's non-additivity test (Hoaglin, Mosteller, and J. W. Tukey 1983) and graphically by the analyst with interaction plots between the dimensions. Thirdly, it is checked whether the residuals are distributed normally, graphically by the analyst with Quantile-Quantile (Q-Q) plots and partly automated with the Shapiro-Wilk normality test and/or the Kolmogorov-Smirnov test, where the null hypothesis is that the residuals come from a normal distribution[2]. Fourthly, the assumption of homoscedasticity is verified with Bartlett's test and/or Fligner-Killeen test of homogeneity of variance, where the null hypothesis is that the variances in each dimension are the same.

Furthermore, for both the independency and log-linear model we test the hypothesis of independence (Equation (3.14)) with the Pearson statistic. Notice that testing the hypothesis of independence in the multinomial model is equivalent with testing the goodness of fit of the Poisson additive model (Everitt 1994).

In the exception identification step (Step 4), the exceptional cells in C are labelled as specified in Definition 3.5. In the software, the cells with high or low exceptions are highlighted with colors, and presented to the analyst. Obviously, the number of exceptional classes can be increased in the software, if the analyst wants to discriminate between more than 2 classes.

Notice that the analyst can return to a previous step in the method if desired. For example, if in the diagnostic step (Step 3) it is shown that the selected model has a poor fit, the analyst can decide to return to the statistical modeling step (Step 2).

[2]These formal tests are quite strict and sensitive to the presence of outliers, therefore from a mild rejection of the null hypothesis we do not directly assume that the residuals are not normally distributed.

In summary, this algorithm is an iterative and interactive process, where the analyst has to configure the parameters in the consecutive steps.

3.6.1 Algorithm for statistical model fitting

In this section we focus on the statistical modeling step (Step 2) of the algorithm. Typically this algorithm is applied on a single cube C on some level $i_1 i_2 \dots i_n$ in the lattice L. Therefore, only a single statistical model is fitted in Step 2. However, an analyst usually explores multiple cubes in L to identify exceptions in a combined analysis. For example, the analyst might create an analyse path P of increasingly specialized cubes from top cube C_T to base cube C_B. For each cube on this path a separate statistical model has to be fitted. Therefore, it is often beneficial, from a computational viewpoint, to reuse computations from cubes in L that are analysed by the analyst previously on the path. Here we review and describe an algorithm for this purpose.

The general idea behind such an algorithm, is to fit a separate but similar statistical model for each cube C in the lattice L, in the form of an ANOVA model (Equation (3.6)) or a log-linear model (Equation (3.20)), and to reuse intermediate modeling results from earlier computations in later ones. Here a model-fitting algorithm is described where a statistical model is fitted on the base cube C_B of some (sub) lattice. This method is inspired by the Up-Down algorithm from Sarawagi et al. (1998).

The algorithm is composed out of three main steps. In the first step (1) the various means are computed, in the second step (2) the statistical model is fitted on the cube C, and in the third step (3) the reference values $\hat{y}(c)$ for all cells in C are computed. The input of the algorithm is a cube $C = [i_1 i_2 \dots i_n]$ with measure values $y^a(C)$ in the lattice L, and the output of the algorithm is $y^r(C)$ computed with an advanced statistical model. The outline of the algorithm for statistical model fitting is presented in Algorithm 3.

Remark 3.6.1. If y is a counting measure and hypothesis (3.14) holds, then R is selected to be the multinomial model (Equation (3.15)). In that case Algorithm (3) is simplified by skipping the first step and by modifying the second step. In the second step we just have to compute the model coefficients with estimates based on Equation

Algorithm 3 Algorithm for statistical model fitting

Initialization:
Consider the cube C on some level $[i_1 i_2 \dots i_n]$ in the lattice L;
Set $C = C_B = [00 \dots 0]$ in the sub lattice L', where $L' = C +$ upset of C;
Computation:

1. Consider the cubes in the lattice L'
   - For each cube in L' cell means $\bar{y}^{i_1 \dots i_q \dots i_n}(c)$ as defined in Equation (3.10)
2. For each cube in L' starting from the top level to the base level do
   - If R is a multi-way ANOVA model (Equation (3.6)) then compute the model coefficients with estimates based on Equations (3.7), (3.8), (3.9), etc.
   - If R is a log-linear model (Equation (3.20)) then compute the model coefficients with estimates based on Equations (3.21), (3.22), (3.23), etc.
3. Consider the cube C
   - Add up all the model coefficients obtained in step (2) to obtain Equation (3.6) or Equation (3.20)
   - Compute $\hat{y}(c)$ for all cells in C, to obtain $y^r(C)$

(3.17). Subsequently, we proceed with the third step.

Remark 3.6.2. In comparison with the algorithm developed by Sarawagi et al. (1998), Algorithm 3 can handle empty cells in the calculation of the coefficients of model (3.6) or model (3.20).

In addition, Algorithm 3 can be extended for fitting a separate statistical model for each cube C in the complete lattice simultaneously, as described in Sarawagi et al. (1998). However, this task is, in general, computationally rather intensive, because of the large number of possible cubes in L (Equation (2.5)). In the first step of the extended method we apply the first step of the algorithm on the complete lattice and form a minimum spanning graph for it. In the second step all the necessary statistical model coefficients for all the cubes in L are computed. This step is the most computational intensive, because the coefficients for each model have to be computed and the computation of each coefficient involves the substraction of $|L| - 1$ coefficients

from higher level cubes. In the third step for each cube in L a statistical model is fitted and the reference values are computed. In conclusion, we argue that the extended algorithm for simultaneously fitting statistical models for all cubes in the lattice has serious practical implementation problems, due to computational complexity. Therefore, we have implemented Algorithm 3 without the extension in our software for exception identification.

3.6.2 Dealing with empty cells

Empty cells are very common in OLAP cubes (Thomsen 1997). Sparsity in an OLAP cube refers to the proportion of cube cells that are empty. Empty cells in the cube obviously need to be taken into account when applying statistical models upon them for exception identification. In general, there are three classes of methods for dealing with incomplete data sets in statistical analysis (van Buuren et al. 1994):

- Discard records that have one or more missing values in the data set;
- Adapt the statistical analysis method;
- Impute (i.e. fill in) unknown entries by "reasonable" values.

A simple method for dealing with incomplete data is to discard records that have missing values from the data set. This method is usually not applicable in OLAP cubes, because empty cells are often the result of the multi-dimensional representation of the data, i.e. a complete fact table might result in incomplete cubes in the lattice. Consequently, discarding records from the fact table would be equal to deleting information from the cube. The other two methods might be used in an OLAP cube. Next we review how the multi-way ANOVA and the log-linear model are adapted when empty cells are present.

In general, our application of ANOVA models simply ignores missing values in the calculation of the model coefficients, by computing the averages only over the values that are actually present. When some cells in C are empty we simply ignore them, and do not count them when computing the effects for the corresponding ANOVA model coefficients, and adjust the formulas for it accordingly. Therefore, we determine the number of empty cells $N_{\text{empty\_cells}}$ for some cube C in the lattice, and adjust the

denominator of Definition 3.3 by subtracting the $N_{\text{empty\_cells}}$ from the total number of cells $|C|$ in the cube. For example, when computing $\bar{y}(\cdot,\cdot,\cdot)$ in Equation (3.2), we divide by $|C| - N_{\text{empty\_cells}}$, where $|C| = |D_1^0| \times |D_2^0| \times |D_3^0|$, to correct for empty cells.

Similarly, the analysis of incomplete contingency tables involves the use of log-linear models from which parameters referring to cells containing structural zeros are excluded since they are known *a priori* to be zero (Everitt 1994). Expected values for such models may be obtained by using modifications of the *iterative proportional fitting* algorithm of Deming and Stephan (1940). The computation of the correct degrees of freedom for the test is, however, complicated by the presence of empty cells. The formula for determining the degrees of freedom in a data cube with empty cells is, equivalent with (Everitt 1994)

$$\text{d.f.} = |C| - N_{\text{model\_parameters}} - N_{\text{empty\_cells}}, \tag{3.24}$$

where $N_{\text{model\_parameters}}$ represents the number of parameters in the model that need to be estimated. It is important to determine the correct number of parameters to be estimated, since those referring to the empty cells are known a priori to be zero and must therefore be excluded. Difficulties arise when fitting log-linear models to contingency table data in the occurrence of zero cell entries. They arise because the logarithm of zero is minus infinity. We solve this problem by adding a small positive constant (e.g., 0.5 or so) to each cell in the base cube.

3.7 Related work

Outlier detection is an important problem within various research areas and application domains; see Chandola, Banerjee, and Kumar (2009) for an overview. Obviously, this topic is closely related to the topic of exception identification in OLAP databases. Outlier detection refers to the problem of finding values in a data set that do not conform to expected behavior. These nonconforming values are referred to as outliers, symptoms, exceptions, surprise values, discordant values, etc. Examples of application areas are the detection of fraudulent credit cards or insurance claims, fault detection in production systems, intrusion detection for cyber-security, finding surprise values in management reports, and so on (Chandola et al. 2009). Researchers

have adopted techniques from diverse areas such as statistics, machine learning, data mining, information theory, and have applied them to specific formulations of the outlier detection problem. In this thesis we formulate the outlier detection problem in statistical terms (Section 3.3).

In statistics, an *outlier* is often defined as a value that lies very far from the middle of the statistical distribution of the variable under consideration in either direction (Mendenhall et al. 1993). This definition is limited to continuous variables. In the identification of of continuous outliers, the frequency of occurrence is also signicant. This is stressed in a different definition: "An outlier is a single, or very low frequency, occurrence of the value of a variable that is far away from the bulk of the values of the variable (Barnett and Lewis 1994)". Therefore, in detecting outliers in categoric data, which are always part of OLAP data, the frequency of occurrence is an important aspect. A general definition of an outlier in a set of continuous or categoric data is: "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data (Barnett and Lewis 1994)".

The phrase 'appears to be inconsistent' in the latter definition is crucial. Because it is a matter of subjective judgement on the part of the observer (i.e., the OLAP analyst) whether or not some observation (i.e., some cell in the cube) is picked out for scrutiny. This judgement is also driven by different aims of the analyst in examining outliers. In statistics, it is often the aim to detect and remove outliers in a data set due to human error and ignorance. However, we define an outlier value in the sense that a value is *surprisingly* high or low in relation to the others, and therefore interesting to the analyst regardless of its cause. Notice that the actual cause of an outlier is often not known to the data analyst. Sometimes, this is an erroneous value resulting from a poor quality data set. In spite of this, we assume that an outlier expresses valid, albeit "exceptional information" (Mendenhall et al. 1993), to the business analyst working with an OLAP cube. The analyst would like to be informed about such exceptional information, because this information might point him to some business problem or opportunity. Based on this information the analyst can decide on further analysis, to find the actual causes of the exception within the OLAP structures, and determine the appropriate business action. Therefore, we do not use the term outlier value but rather the term exceptional value or surprise value.

In the literature numerous techniques are identified for outlier detection. See Barnett and Lewis (1994) for a comprehensive overview. In this work, two classes of outlier detection methods are distinguished: univariate methods, where analysis is performed on an individual variable, and multivariate methods, which analyse more than one variable at a time. Outlier detection methods for univariate data are of little concern for outlier detection in OLAP databases because measures are multivariate (Definition (2.17)). There are multiple methods for outlier detection in multivariate data, usually dependent on the structure of the data (Barnett and Lewis 1994). For example, the outlier may be a value in a regression analysis, a time series analysis, unstructured multivariate data, etc. Besides, informal and formal methods are developed for outlier detection in the statistical literature. In principle both informal and formal methods can be used in OLAP databases.

Manual inspection of scatter plots is the most common informal analysis (Pyle 1999). Here data analysts have to use their own intuitition to decide on parameters to single out outliers. Obviously, manual inspection of scatter plots for every variable is time-consuming and therefore not applicable in large multi-dimensional databases, containing millions of numeric and categoric values. In Barnett and Lewis (1994), an informal, unsupervised method for identification of numeric outliers is explained. This method is based on the construction of a boxplot, which represents data via their quartiles. In the boxplot, most values are assumed to be in the interquartile range (H) The authors label values lying outside the $\pm 1.5H$ range as mild outliers and values outside the boundaries of $\pm 3H$ as extreme outliers. In Chen (1999) a method is outlined to construct box plots for OLAP cube data. In some practices like monitoring a manufacturing process, the 3σ rule is generally adopted. The 3σ rule is: calculating the mean μ and the standard deviation σ, and if one observation lies outside the $(\mu - 3\sigma, \mu + 3\sigma)$ range, we say it an outlier. Some researchers suggest using the median and the MAD scale instead of the mean and the standard deviation for detecting outliers (Hoaglin et al. 1983).

In the research on outliers detection, a number of formal statistical tests were developed, called *discordancy tests* (Barnett and Lewis 1994; Hawkins 1994). In a discordancy test, potential outliers are tested with the prospect of rejecting it from the data set, or of identifying it as a feature of special interest. Specific discordancy tests

are developed for specific statistical distributions (e.g., normal samples, exponential samples, and Pareto samples) and data structures (e.g., regression, the linear model, and designed experiments). For example, the notion of outlying variables in regression analysis is related to the examination of residuals. For the simple linear model the maximum absolute standardized residual is used to detect and test the discordancy of a single outlier (Barnett and Lewis 1994). Basically, our method for exception identification is a discordancy test for multi-dimensional data. Furthermore, our method can be seen as a multivariate version of Grubb's test (Grubbs 1969).

Finally, two related specific works on exception identification in OLAP databases are mentioned. The first work applies a statistical model for exception identification and the second a data mining model.

Important early research work on statistical exception identification is the work by Sarawagi, Agrawal, and Megiddo (1998) in the i^{3}Cube project[3]. In this project the authors developed a discovery-driven exploration paradigm that explores the multi-dimensional data for exceptions and summarizes the exceptions at appropriate aggregation levels in advance, by applying a log-linear model. The discovery-driven method is guided by pre-computed indicators of exceptions at various levels of detail in the cube. By this method the analyst is guided by the model to interesting data regions using pre-computed indicators. In Cariou et al. (2007), a similar approach is taken. In this work the Chi-square contribution and test-value are used to discover interesting cells. Our method for problem identification is quite similar, however our approach is based on both the multi-way ANOVA model as the independency model, dependent on the type of measure (continuous or discrete). In contrast, our algorithm for exception identification (Section 3.6) pays specific attention to the diagnostics related to the statistical model, for example, to check the statistical model assumptions. Diagnostics are nearly absent in Sarawagi et al. (1998).

In the literature, multiple data mining methods have been developed to discover informative parts of the OLAP data cube. In Lin and Brown (2003) and Lin and Brown (2006), an OLAP-outlier-based data association method is proposed. This method integrates both outlier detection concepts in data mining and ideas from the OLAP field. An outlier score function is defined on OLAP cube cells which measures

[3]http://www.cse.iitb.ac.in/∼sunita/icube/index.htm

the extremeness of the cell. They associate the data points in the cell when the cell is unusual. Their method is applied to the problem of associating criminal incidents. Result shows that this combination of OLAP and data mining provides a promising solution to the problem. Furthermore in Usman et al. (2013b), a method is presented to discover association rules in OLAP databases. This method supports the singling out of "informative dimension and fact variables". And in Usman et al. (2013a), data mining methods based on principal component analysis combined with agglomerative hierarchical clustering are applied on multi-dimensional data sets to "discover cubes of interest". Similar to these works, we also focus on OLAP cube cells in outlier detection. However, we do not apply a data mining model for this purpose but various classical statistical models.

3.8 Conclusion

In this chapter, we extended the functionality of the multi-dimensional database with exception identification. Exceptional cell values are determined based on a normative model R. We discussed two broad classes of normative models for OLAP databases: managerial models and statistical models, and we discussed how they can be used in an OLAP cube. In the case of normative models we differentiate between planning and budget models, historical models, and extra-organizational models. In the case of statistical normative models we distinguish between simple and advanced statistical models. Important advanced statistical models are the multi-way ANOVA model for continuous measures and independency models for discrete measures. For statistical models we developed a hypothesis test to identify exceptional values in a cube C. Moreover, we showed how the estimates for statistical models can be determined by operations on the aggregation lattice. Finally, we provided an algorithm for statistical exception identification and presented an algorithm for statistical model fitting. In Chapter 6 this algorithm is applied and illustrated in a number of practical business case studies.

Chapter 4

Explanation of exceptional values*

4.1 Introduction

In this chapter, we describe a method to automatically generate causal explanations of exceptional cell values. In the multi-dimensional database the actual model in the general diagnosis task (Chapter 1) is determined by a system of drill-down equations and a system of business model equations (Chapter 2). The OLAP analyst determines the object of a diagnostic task by selecting an exceptional cell (Chapter 3). The normative model is determined by a reference class R used for exceptional cell identification. The reference class specifies the reference objects in the explanation formalism. Consider a lattice L of OLAP cubes, a cube $C = [i_1 i_2 \dots i_n]$, and an exceptional cell value $\partial y(c) = q$ with $c \in C$. We can explain the exceptional values by:

1. purely business equations from the model M related to the measure y, or
2. purely drill-down equations from the exceptional cell's downset $\{\downarrow c\}$, where $c \in C$, or
3. combinations of drill-down and business equations.

The remainder of this chapter is organised as follows. In Section 4.2, we review the most important concepts of the explanation methodology for automated business

*This chapter is mainly based on two articles by Caron and Daniels (2007) and Daniels and Caron (2009).

diagnosis as developed by Feelders (1993) and Feelders and Daniels (2001). Here explanation is discussed in the context of business model M and based on a generic explanation formalism. In addition, we deal with the appropriate conditions under which the explanation formalism produces valid explanations. In Sections 4.3, 4.4, and 4.5 we apply the explanation formalism to explain an exceptional cell value in a multi-dimensional OLAP database. In Section 4.3, we extend the explanation method with a procedure to deal with cancelling-out effects in data sets. This is a common phenomenon in financial and other data sets. This procedure is implemented in a look-ahead algorithm. In Section 4.4, we discuss explanation in systems of purely drill-down equations. In Subsection 4.4.1, we describe a general top-down explanation method for these type of systems. In Subsection 4.4.2, we particularly focus on systems of additive equations that exhibit the property of transitivity. We use this property to construct a greedy algorithm for explanation. In Section 4.5, we discuss explanation in hybrid systems of equations, i.e. in systems with both OLAP drill-down and business model equations. For this purpose, we propose a general algorithm for explanation. In Section 4.6, we develop filter methods to reduce the number of explanations that are generated by the algorithms for explanation. In this way explanation trees can be pruned to a manageable size. In Section 4.7, we discuss how to construct consistent chains of reference objects for various types of normative models applicable in the OLAP context. In Section 4.8, we discuss related work on computerized diagnosis in the field of business and management. In Section 4.9, we draw some conclusions.

4.2 Overview of theory on explanation*

4.2.1 Explanation formalism

The explanation formalism applied to multi-dimensional databases is largely based on Feelders and Daniels' method for explanations, which in turn is essentially based on Humphreys' notion of aleatory explanations (Humphreys 1989) and the theory of explaining differences by Hesslow (1983). Causal influences can appear in two forms:

* With the exception of subsection 4.2.4, this section gives an overview of the work presented in (Feelders and Daniels 2001; Heckman 2000; Feelders 1993; Kosy and Wise 1984).

contributing and *counteracting*. The canonical formalism for causal explanations is given by

$$\langle a, F, r\rangle \text{ because Cb, despite Ca,} \tag{4.1}$$

where $\langle a, F, r\rangle$ is the event to be explained, Cb is a non-empty set of contributing causes, and Ca a set of counteracting causes, which can be empty. The explanation itself consists of the causes to which Cb refers. Ca is not part of the explanation, but gives a clearer notion of how the members of Cb actually brought about the symptom.

In words, the explanandum is a three-place relation between object a that shows the actual behaviour, a property F, that shows the deviation for a particular variable from its norm value, and a reference object r, obtained from the normative class R. In the OLAP context, for example, the actual object a might be the cell c =(2010, Germany, beer) from the cube $C = \{2010\} \times$ Country $\times$ Product, and the reference object r might be the cell c' =(2009, Germany, beer) from the cube $C = \{2009\} \times$ Country $\times$ Product. The property F is that the measure profit in the year 2010 for the cell c is relatively low compared to profit in the previous 2009 for the cell c'. The task now is not to explain why a has property F, but rather to explain why a has property F *when the other members of* r *do not*. For example, when r is selected as the statistically normal case, the explanatory cause must be abnormal.

If $\partial y(c) = q$ is identified as an exceptional cell value, by the methodology as discussed in Chapter 3, we can subsequently try to explain the difference $\partial y(c) = y^a(c) - y^r(c)$ based on the internal structures of the multi-dimensional database, that are described in Chapter 2. By using (4.1) the event to be explained can be rewritten as

$$\langle y^a(c), \partial y(c) = q, y^r(c)\rangle \text{ because Cb, despite Ca.} \tag{4.2}$$

In this expression it can be the case that $y^r(c) = \hat{y}(c)$, where $\hat{y}(c)$ is computed by a statistical normative model described in Chapter 3 and $\hat{y}(c)$ is associated with the same cell c as the actual value. Besides it can be the case that $y^r(c) = y(c')$, where $y(c')$ is computed by a managerial normative model described in Chapter 3 and $y(c')$ is associated with a different cell c' in comparison with the actual cell c.

4.2.2 Causality

Explanations generated with the explanation formalism based on (4.1) are based on general laws expressing relations between events: such as cause-effect relations or constraints between variables. In a multi-dimensional database, these general laws are represented in two internal structures: the system of drill-down equations (see Equation (2.11)) and the system of business model equations M. The system of drill-down equations can be represented in a semilattice SL and the system of business equations can be represented in a business model graph $G(M)$ (Section 2.3.2). The vertices in both graphs, which represent variables in the drill-down equations and variables in the business model, indicate the direction of influence, or *causal direction*. Interpreting the $=$ in both systems of equations as a $\leftarrow$, the causal direction is given as used by economists, accountants or financial analysts. Thus, in both systems of equations the effects appear on the left-hand side (LHS) of the equations and the causes on the right-hand side (RHS). The direction of explanation is the opposite of the causal direction. In other words, the explanation generation process proceeds from the whole, the LHS variables, to the parts, the RHS variables.

Generally speaking, there is not a single notion of causality in economics and business. In principle one would say that the cause precedes the effect in time. For example, rain is the cause of a wet street or after an increase in the price of a product demand will decrease (c.p.). Even this simple notion might be tricky, for example, a child may think that closure of a railway crossing barrier will cause a train to arrive. Another example where correlation between the data does not imply a causal relation was found in a database with data on traffic accidents (Feelders et al. 2000). This notion of causality has been discussed extensively in de Kleer and Brown (1986), de Kleer et al. (1992), and Reiter (1987).

Other less intuitive notions of causality are also known in economics. For example, X causes Y if information on X leads to better prediction of Y. This definition from econometrics is due to Granger (2001). It is also well known that economists use the implicit intuitive notion of causality to reason about static economic models (Berndsen and Daniels 1990). In these cases it is assumed that some effects happen instantaneously, like the clearing of the market, whereas in reality there is a small time lag between cause and effect. However static models are often preferred because

they are much simpler and provide a comprehensive framework to answer questions in comparative statics (Samuelson 1941). Nevertheless this type of "cause-effect" reasoning can be confusing for novices.

In multi-dimensional databases the direction of causality is obvious, because the reasoning is always from the whole, e.g. a parent cell, to the constituent parts, e.g. the parent's child cells. In general, causes with greater influence are considered more important by the analysts. Of course the measure of influence has to be chosen in correspondence with the notion of significance of the analyst.

4.2.3 Measure of influence

Suppose $y = f(\mathbf{x})$ is an equation of the business model M, then we define a *measure of influence* as follows

$$\inf(x_i, y) = f(\mathbf{x}^r_{-i}, x^a_i) - y^r, \tag{4.3}$$

where $f(\mathbf{x}^r_{-i}, x^a_i)$ denotes the value of $f(\mathbf{x})$ with all variables evaluated at their reference values, except the measure x_i. In words, $\inf(x_i, y)$ indicates what the difference between the actual and reference value of y would have been if *only* x_i would have deviated from its reference value. The *inf-measure* represents a form of ceteris paribus reasoning where the x_i's play the role of causes that produced y. For computational purposes we store for each equation in the business model a change in the actual, reference, and influence measure values in an *influence table*; see Table 4.2 for an example.

The inf-measure enables the operationalisation of the concepts of contributing and counteracting causes in expression (4.2). The *set of contributing* (*counteracting*) *causes* Cb (Ca) consists of variables x_i with

$$\inf(x_i, y) \times \partial y > 0 \ (< 0). \tag{4.4}$$

In words, the contributing causes are those variables whose influence values have the same sign as ∂y, and the counteracting causes are those variables whose influence values have the opposite sign.

Insignificant influences are left out in the explanation by means of a *reduction measure* or *method* (RM). Other reduction methods are described in detail in Section

4.6. RM_1 reduces the set of causes reduced to the so-called *parsimonious* or *significant set of causes*. The *parsimonious set of contributing causes* Cb_p is defined as the smallest subset of the set of contributing causes Cb, such that its influence on y exceeds a particular fraction T^+ of the influence of the complete set, i.e.

$$\frac{\inf(\mathrm{Cb}_p, y)}{\inf(\mathrm{Cb}, y)} \geq T^+. \tag{4.5}$$

The definition regarding parsimonious counteracting causes, Ca_p, is similar. The fractions T^+ and T^- are numbers between 0 and 1 and are determined empirically by the analyst.

4.2.4 Consistency and Conjunctiveness

A correct interpretation of the influence measure, i.e. the generation of valid explanations for a symptom, is only possible if and only if the following two constraints are fulfilled:

1. the actual and reference values satisfy the *consistency constraint* (Definition 4.1 below), and

2. the function f satisfies the *conjunctiveness constraint* (Definition 4.2 below).

Definition 4.1. The consistency constraint states that the reference values must satisfy the same functional requirements as the actual values, i.e. $y^a = f(\mathbf{x}^a)$ and $y^r = f(\mathbf{x}^r)$, where the reference objects are obtained by a normative model R.

This is not always the case, because in some situations, $y^r \neq f(\mathbf{x}^r)$ due to the form of the function f or the type of normative model R applied. If this is the case, the explanation procedure described in this section is questionable, because then $\partial y = y^a - y^r \neq \sum_{i=1}^{n} \inf(x_i, y)$.

Example 4.2.1. A straightforward example of a violation of the consistency constraint is given in Table 4.1. In this table we observe the actual values of business variables in the equation $y = x_1 \times x_2$ for two different firms. The column average (Avg) is the reference value. From the last column in this table we infer that $y^r \neq x_1^r \times x_2^r$, where $y^r = \frac{1}{2}(y^a(\text{Firm 1}) + y^a(\text{Firm 2}))$. Here taking reference vales and applying f do not commute: $y^r = \mathrm{Avg}(y^a) \neq f(\mathrm{Avg}(x^a))$.

Table 4.1: Actual and norm values for $y = x_1 \times x_2$.

Variables	Firms		
	1	2	Avg
y	8	10	9
x_1	2	5	3.5
x_2	4	2	3

In Section 4.7, we explain in detail under what conditions the reference values satisfy the functional equations of OLAP or business model equations. Furthermore, we describe for managerial and statistical normative models, that are applicable in the OLAP context, how to construct a consistent chain of reference values.

Definition 4.2. A model equation satisfies the conjunctiveness constraint if for all subsets $X \subseteq \{x_1, \ldots, x_n\} \backslash \{x_i\}$ the following holds

$$\begin{aligned} \inf(x_i, y) \geq 0 \;&\Rightarrow\; \inf(X \cup \{x_i\}, y) \geq \inf(X, y), \\ \inf(x_i, y) \leq 0 \;&\Rightarrow\; \inf(X \cup \{x_i\}, y) \leq \inf(X, y). \end{aligned}$$

This constraint captures the intuitive notion that the influence of a single variable x_i should not turn around when it is considered in conjunction with the influence of a number of other variables. Only under this condition can significant causes be joined together as a total set (Section 4.6, Equation 4.5).

Two large classes of functions satisfy the conjunctiveness constraint, namely *additive* and *monotonic functions* (Feelders and Daniels 2001). By monotonicity we mean monotonicity in all variables separately, on the domain under consideration. Relations in financial models are almost always monotone. Additivity and monotonicity can also be easily checked in the business model. For example, the financial model presented in Chapter 1, Table 1.1 consists of 2 additive relations and 3 monotonic relations.

If f satisfies the consistency constraint and the conjunctiveness constraint then the following holds (assuming $y^a > y^r$):

- $f(x_1, x_2, \ldots, x_n)$ increases if $x_i \in$ Cb is changed from x_i^r to x_i^a, and
- $f(x_1, x_2, \ldots, x_n)$ decreases if $x_i \in$ Ca is changed from x_i^r to x_i^a.

Consequently, $y^r = f(x_1^r, x_2^r, \ldots, x_n^r)$ gradually changes to $y^a = f(x_1^a, x_2^a, \ldots, x_n^a)$ by replacing the reference values to the actual values of x_i. Notice that *for the remainder of this thesis it is assumed that the consistency and conjunctiveness constraints are satisfied.*

4.2.5 Interpretation of the influence measure

The interpretation of the inf-measure (expression 4.3) is dependent on the functional form of the function f. For additive measures that are common in OLAP, we show that $\partial y = y^a - y^r = \sum_{i=1}^{n} \inf(x_i, y)$.

Theorem 4.2.1. If f is an additive function such that $y^a = \sum_{k=1}^{n} s_k(x_k^a)$, where the s_k, $k = 1, \ldots, n$, are arbitrary functions, and $y^r = \sum_{k=1}^{n} s_k(x_k^r)$ then $\partial y = y^a - y^r = \sum_{i=1}^{n} \inf(x_i, y)$.

Proof.

$$
\begin{aligned}
& y^a - y^r = \sum_{k=1}^{n} s_k(x_k^a) - \sum_{k=1}^{n} s_k(x_k^r) \\
& \inf(x_i, y) = \sum_{k \neq i}^{n} s_k(x_k^r) + s_i(x_i^a) - y^r = s_i(x_i^a) - s_i(x_i^r) \qquad (4.6) \\
& \text{and therefore: } \sum_{i=1}^{n} \inf(x_i, y) = y^a - y^r \ . \square
\end{aligned}
$$

Correspondingly, in the situation that the function f is the average, the inf-measure is given by $\inf(x_i, y) = (x_i^a - x_i^r)/n$, where n is the number of RHS elements in the function. In this case $\partial y = \sum_{i=1}^{n} \inf(x_i, y) = \sum_{i=1}^{n} (x_i^a - x_i^a)/n$ and Theorem 4.2.1 applies.

Moreover, in the case that the function f is differentiable, and holds for both the set of actual values as for the set of reference values, then $\inf(x_i, y)$ is also correctly interpreted as a quantitative specification of the change in y that is explained by a relatively small change in x_i.

Lemma 4.2.2. If f is possibly non-additive but differentiable, $y^r = f(\mathbf{x}^r)$ and $\delta_i = x_i^a - x_i^r$ is small then $\partial y \approx \sum_{i=1}^{n} \inf(x_i, y)$.

Proof. This follows immediately from the Taylor series expansion of f around x_i^r.□

Remark 4.2.1. Notice that in general ∂y is not necessarily equal to $\sum_{i=1}^{n} \inf(x_i, y)$, even in the case when f has an additive form, but when $y^r \neq f(\mathbf{x}^r)$. In such cases, the influence measure is difficult to interpret as shown in the example in Table 4.1. For firm 1 in this table we infer that $-1 = \partial y \neq \inf(x_1, y) + \inf(x_2, y) = 2$. In this example we can interpret the sign of the inf-measure but not its value.

4.2.6 Maximal explanation

So far, we have discussed "one-level" explanations, explanations based on a single equation from the business model M. For diagnostic purposes, however, it is meaningful to continue an explanation of $\partial y = q$, by explaining the quantitative differences between the actual and norm values of its contributing causes in the business model. Causes can be chained together, from one level to the next in the business model, until a *maximal explanation* is obtained, (Feelders 1993; Feelders and Daniels 2001).

The idea behind the method of maximal explanation is to construct an *explanation tree* or *tree of causes* T with $\partial y = q$ on level M^0 as the root, the children of the root are contributing causes on level M^1, the grandchildren of the root are contributing causes on level M^2, and so forth, until the contributing causes on level M^d. Usually we only add parsimonious causes (see RM_1) to the tree.

In Figure 2.7 on page 45 in Section 2.3.2, a multi-level business model with measures from a financial database is depicted. In the tree the counteracting causes are not explained any further, because they are not seen as part of the explanation itself. The explanation process is continued until a contributing cause is encountered that cannot be explained within the business model M. In the explanation tree, T^p denotes the level p in T, where $p = 0, 1, \ldots, d$ and $T^0 = \partial y = q$ is the root of tree.

4.3 Cancelling-out effects and look-ahead explanation

A shortcoming of the method of maximal explanation is that it cannot deal with *cancelling-out* or *neutralisation effects*. Cancelling-out is the phenomenon that the

effects of two or more lower-level variables in the business model M cancel each other out, so that their joint influence on a higher-level variable in the business model is partly or fully neutralized. For example, the first half-year positive financial results could partially cancel out the negative financial results of the second half-year in a financial model. This cancelling-out pattern would not be visible on the aggregated year level in the model. These *hidden causes* are quite common in business data. Hidden causes are significant causes that are not visible in the explanation tree, because they are cancelled-out by other variables. The problems with these patterns were first mentioned by Kosy and Wise (1984), however no solution was presented in their article. In this section, we present a look-ahead explanation method that deals with the presence of cancelling-out effects.

4.3.1 Making hidden causes visible by substitution

In theory, cancelling-out effects may occur at every level in the business model. Of course, business and financial analysts would like to be informed about these hidden causes, and would consider an explanation tree without mentioning these causes as incomplete. We develope a method that can identify hidden causes if present.

Suppose that we are explaining a symptom $\partial y = q$ and the following equations from the business model M

$$y = f(\mathbf{x}) \in M^{p;(p+1)}, \tag{4.7}$$

$$x_i = g_i(\mathbf{z}) \in M^{(p+1);(p+2)}, \tag{4.8}$$

where $\mathbf{x} = (x_1, \ldots, x_i, \ldots, x_n)$ and $\mathbf{z} = (z_1, \ldots, z_m)$ denote n and m-component vectors. In the above equations, $M^{p;(p+i)}$ represents a subset of equations from the business model M. In this notation the variables on level p appear on the *left hand side of the business model equations* on level M^p. These are expressed in terms of the variables on a higher level $(p+i)$, that appear on the *right hand side of the business model equations* on level $M^{(p+i)}$. For variables that cannot be expressed in variables at level $(p+i)$, i.e. leaf nodes in M on intermediate levels, we use variables at a lower level closest to the level $(p+i)$.

Now suppose that explanation generation with Equation (4.7) on level $M^{p;(p+1)}$ results in sets of parsimonious causes where *variable x_i does not belong to*, thus

$x_i \notin \mathrm{Cb}_p(y)$ or $x_i \notin \mathrm{Ca}_p(y)$. In words, the variable x_i is not significant because it has a marginal influence on the root y. An extreme situation occurs when $\inf(x_i, y) = 0$, then the variable x_i has no influence on ∂y. To make sure that the explanation is complete, all successors of x_i have to be investigated for possible cancelling-out effects. Therefore, all children of x_i, i.e. the elements of $\mathbf{z}$ in the RHS of Equation (4.8), are substituted into the RHS of Equation (4.7) to derive the new equation

$$y = h_i(\mathbf{x}, \mathbf{z}) \in M^{p;(p+2)}. \tag{4.9}$$

$M^{p;(p+2)}$, the result of *substituting* jointly all equations on level $M^{(p+1);(p+2)}$ into the parent equation(s) on level $M^{p;(p+1)}$, is *added to the business model*. Now Equation (4.9) is used for explaining the symptom $\partial y = q$ and we obtain causes at level $p + 2$, possibly not captured by straightforward application of maximal explanation. This procedure is called *one-step look-ahead*.

Example 4.3.1. Here we consider the business model M composed out of the equations $y = f(x_1, x_2) \in M^{0;1}$ and $x_2 = g(z_1, z_2) \in M^{1;2}$. $G(M)$, the graph of M, is depicted in Figure 4.1 on the left-hand side. In this example we apply the one-step look-ahead method on the equations. We substitute equation $M^{1;2}$ into equation $M^{0;1}$ to derive the equation $y = f(x_1, g(z_1, z_2)) = h(x_1, z_1, z_2) \in M^{0;2}$. On the right-hand side of Figure 4.1, this equation is depicted as the explanatory graph $G(M^{0;2})$. Notice that in Section 6.2 an extensive example is given.

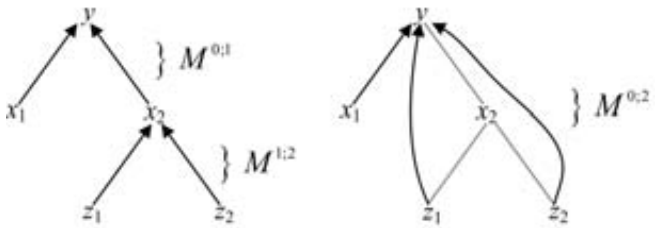


Figure 4.1: Explanatory graphs for $G(M)$ (left) and for one-step look-ahead $G(M^{0;2})$ (right).

We now define contributing and counteracting hidden causes and their influence on a symptom ∂y.

Definition 4.3. Variable z_j of Equation (4.9) is a *contributing hidden cause* when $z_j \in \mathrm{Cb}_p(y)$ and $x_i \notin \mathrm{Cb}_p(y)$, where z_j is a successor of x_i.

Definition 4.4. Variable z_j of Equation (4.9) is a *counteracting hidden cause* when $z_j \in \mathrm{Ca}_p(y)$ and $x_i \notin \mathrm{Ca}_p(y)$, where z_j is a successor of x_i.

Here the influence of z_j on y is given by:

$$\inf(z_j, y) = f(\mathbf{x}^r_{-i}, g_i(\mathbf{z}^r_{-j}, z^a_j)) - f(\mathbf{x}^r_{-i}, g_i(\mathbf{z}^r)), \tag{4.10}$$

and the influence of x_i on y is given by:

$$\inf(x_i, y) = f(\mathbf{x}^r_{-i}, x^a_i) - f(\mathbf{x}^r) = f(\mathbf{x}^r_{-i}, g_i(\mathbf{z}^a)) - f(\mathbf{x}^r_{-i}, g_i(\mathbf{z}^r)). \tag{4.11}$$

This means that the effect of z_j is neutralized by the effects of other variables in the vector $\mathbf{z}$. It is assumed that the function h_i satisfies the conjunctiveness constraint. In the special case that the functions f and g_i from Equations (4.7) and (4.8) are both additive we have $\inf(x_i, y) = \sum_{j=1}^{m} \inf(z_j, y)$ (see Theorem 4.2.1). From this relation it immediately follows that when $x_i \notin \mathrm{Cb}_p(y)$ (or $x_i \notin \mathrm{Ca}_p(y)$) and $z_j \in \mathrm{Cb}_p(y)$, at least one variable z_j is in the set of counteracting causes $\mathrm{Ca}(y)$. Or vice versa, when $x_i \notin \mathrm{Cb}_p(y)$ (or $x_i \notin \mathrm{Ca}_p(y)$) and $z_j \in \mathrm{Ca}_p(y)$, at least one variable z_j is in the set of contributing causes $\mathrm{Cb}(y)$.

4.3.2 Algorithm for look-ahead explanation

The one-step look-ahead method can simply be extended to *multi-step look-ahead* to visualize hidden causes at deeper levels in M, in the following way. *Two-step look-ahead* is defined as explanation in $M^{p;(p+3)}$, the result of substituting all equations at level $M^{(p+2);(p+3)}$ into $M^{p;(p+2)}$, and so on for $M^{p;(p+4)}$, $M^{p;(p+5)}$, ..., $M^{p;(p+d)}$. In general, for a business model M with depth d, the maximal number of look-ahead steps is $d-1$. In the multi-step look-ahead method, we generalize Definitions 4.3 and 4.4 to other levels in M as follows: a *successor of variable* x_i on level $(p+i)$ is

a *hidden cause* if its influence on y is significant after $i-1$ substitutions, when the influence of variable x_i of Equation (4.7) on y is not significant.

Here a look-ahead algorithm is proposed which is composed out of two consecutive phases: an *analysis* (1) and a *reporting* (2) phase. In the analysis phase the explanation generation process starts for a symptom ∂y, similar as for maximal explanation, with the root equation in the business model by determining parsimonious causes (RM_1). However, instead of proceeding with strictly parsimonious causes, all non-parsimonious causes are investigated for possible cancelling-out effects at a specific level in M. In this phase, hidden causes are made visible by means of substituting equations. The derived equations are added to M and considered for explanation generation. In the reporting phase the explanation tree is updated when hidden causes are identified. In updating the tree new parsimonious causes are added and causes that have become non-parsimonious are removed. In Algorithm 4, the pseudo code of the algorithm is presented, where q is the number of selected look-ahead steps.

Example 4.3.2. In Section 6.2, the working of the algorithm for look-ahead explanation is shown in a case study on interfirm comparison at Statistics Netherlands (2009).

Remark 4.3.1. When Algorithm 4 is executed with $q = 0$ (i.e. no look-aheads) the algorithm reduces to maximal explanation, as discussed in Section 4.2.6.

4.4 Explanation in a system of drill-down equations

If an exceptional cell value $\partial y(c)$ is identified in a cube C, the next step is to explain this exception within the internal structures of the OLAP database, i.e. the system of drill-down equations and/or the system of business model equations. To do this we propose:

1. explanation methods for systems of purely drill-down equations:
   (a) a *top-down explanation method* (Section 4.4.1);
   (b) a *greedy explanation method* (Section 4.4.2), if only certain types of drill-down equations apply;

Algorithm 4 Multi-level explanation algorithm with look-ahead

Initialization:
  $\partial y = q$: a symptom;
  M: a business model, where d is the number of levels in M;
  RM_1: with pre-determined values for T^+ and T^-;
  q: the number of look-ahead steps, where $0 \leq q < d$;
Computation:
  $p := 0$;
  y is the root node of the explanation tree T^0;
  repeat {Maximal explanation}
    if a node corresponds to a parsimonious contributing cause **then**
      determine parsimonious causes $\mathrm{Cb}_p(y)$ and $\mathrm{Ca}_p(y)$ for equation(s) $M^{p;(p+1)}$;
      add parsimonious causes to T^{p+1} as child nodes;
      $p := p + 1$;
    end if
  until a node corresponds to a variable that cannot be explained on $M^{p;(p+1)}$ **or** $p := (d-1)$;
  if $q > 0$ **then**
    for $k = 0$ to $d - 1$ **do** {Analysis phase}
      for $p = 1$ to q **do**
        substitute jointly all equations on $M^{(p+k);(p+k+1)}$ into equation $M^{(k);(p+k)}$;
        add new equations $M^{(k);(p+k+1)}$ to M;
        determine parsimonious causes $\mathrm{Cb}_p(y)$ and $\mathrm{Ca}_p(y)$ for equation $M^{(k);(p+k+1)}$;
        if causes on level M^{p+k+1} are parsimonious **then** {Reporting phase}
          add parsimonious causes to T^{p+k+1} as successor nodes;
          remove causes from T^{p+k+1} that have become non-parsimonious;
        end if
      end for
    end for
  end if

2. and an *explanation method for hybrid systems of equations* (Section 4.5), composed out of both drill-down and business model equations.

4.4.1 Top-down explanation

In this section, we apply the method of maximal explanation (Section 4.2.6) in a system of drill-down equations. Here a symptom is explained, top-down, level-by-level, in the symptom's downset. In this manner, the explanation approach for symptoms in multi-dimensional data seems quite similar to the data mining process at multiple dimension levels. Especially, the idea of *progressive deepening* (Han 1995) is very "natural" in explanation; start symptom detection on an aggregated level in the symptom's downset and progressively deepen it to find the causes for that symptom at lower levels in its downset. This idea we adopt for *top-down explanations in OLAP*.

In this case the explanation process of an exceptional cell value $\partial y^{i_q}(c) = q$[1] in a cube $C = [i_1 \dots i_q \dots i_n]$, can be continued top-down over some analysis path p in L, to the base cube C_B. Here the explanation generation procedure is based on the computation of the inf-measure for the *same measure* $y(c)$ *for different cells in the downset* $\{\downarrow c\}$.

Formally, this procedure can be written as follows. If $\partial y^{i_q}(c) = q$ is an exceptional cell value in a cube C then the causes one level deeper in its down set $\{\downarrow c\}$, can be computed by using the following expression for the inf-measure:

$$\inf(y^{a;i_q-1}(c'), y^{a;i_q}(c)) = y^{a;i_q-1}(c') - y^{r;i_q-1}(c'), \tag{4.12}$$

where y is an additive measure, as defined in Definition 2.20. This is a direct result of Equation 4.6. In the case that y is an average drill-down measure, as defined in Definition 2.21, then

$$\inf(y^{a;i_q-1}(c'), y^{a;i_q}(c)) = \frac{1}{|R_q^{-1}(C)|}(y^{a;i_q-1}(c') - y^{r;i_q-1}(c')). \tag{4.13}$$

In addition, the above expressions for the inf-measure can directly be used in expression (4.4) to determine sets of contributing and counteracting causes.

[1] where $y^{i_q}(c)$ is the shorthand notation of $y^{i_1 \dots i_q \dots i_n}(c)$.

Example 4.4.1. Suppose the event to be explained is $\langle \text{profit}^a(2010, \text{Germany, Beer}), \partial\text{profit}=\text{“high”}, \text{profit}^r(2009, \text{Germany, Beer})\rangle$, where the results from the previous year are used as reference objects. The increase in profit on the Year level is examined on the Quarter level of the Time dimension. The corresponding additive drill-down equation is

$$\text{profit}(2010, \text{Germany}, \text{Beer}) = \sum_{j=1}^{4} \text{profit}(2010.\text{Q}_j, \text{Germany}, \text{Beer}).$$

The method yields the following results, taking RM_1 (see expression (4.5)) with fractions $T^+ = T^- = 0.9$. In Table 4.2 a comparison is made between profit(2010, Germany, Beer) and profit(2009, Germany, Beer) (norm). From the data in the table it follows that Cb = {profit(2010.Q3, ., .), profit(2010.Q4, ., .)} and Ca = {profit(2010.Q1, ., .)}. Cb_p = {profit(2010.Q4, ., .)} since only profit(2010.Q4, ., .) is needed to explain the desired fraction on inf(Cb, profit(2010, Germany, Beer)) and Ca_p = {profit(2010.Q1, ., .)}.

Table 4.2: Data for explanation of ∂profit(2010, Germany, Beer)=“high”.

	norm (2009)	actual (2010)	inf
profit(2010, Germany, Beer)	100	150	
profit(2010.Q1, ., .)	25	0	-25
profit(2010.Q2, ., .)	25	25	0
profit(2010.Q3, ., .)	25	26	+1
profit(2010.Q4, ., .)	25	99	+74

The result of top-down explanation is an explanation tree of causes T, where the root of the tree is $\partial y(c) = q$ with two types of children, corresponding to its parsimonious contributing and counteracting causes respectively. A node in T that corresponds to a parsimonious contributing cause is a new symptom on a lower level that can be explained further. A node that corresponds to a parsimonious counteracting cause has no successors. The corresponding algorithm is Algorithm 5. The output of this algorithm is a tree of causes.

Moreover, there are numerous explanation paths from the root to the leaf nodes in the explanation trees generated by Algorithm 5. This implies that in general many

Algorithm 5 Algorithm for top-down explanation in drill-down equations

Initialization:
  $y^{\mathbf{i}}(c) = q$: a symptom in the context cube C;
  L: the symptoms's downset $\{\downarrow c\}$ with actual and reference values;
  analysis path $p(C, C')$ in L;
  reduction methods taken from {RM_1,RM_2,RM_3};
Computation:
  $t := 1$;
  $k := |p(C, C')|$;
  $A \leftarrow C$ where $A = [i_1 i_2 \dots i_n]$ and $c \in A$;
  $\partial y^{\mathbf{i}}(c) = q$ is the root of the explanation tree T^0;
  repeat {Maximal explanation in a downset L via path p}
    if a node in T^t corresponds to parsimonious contributing cause **then**
      $A' := R^{-1}(A)$ where $A' = [j_1 j_2 \dots j_n]$ and $c' \in A'$ and drill-down R^- specified by p;
      determine parsimonious causes Cb_p and Ca_p for equation $y^{\mathbf{i}}(c) = \sum_{c' \in R^{-1}(c)} y^{\mathbf{j}}(c')$
      add parsimonious causes to T^t as child nodes;
      $k := k - 1$, $t := t + 1$, $A \leftarrow A'$ where $A = [i_1 i_2 \dots i_n]$ and $c \in A$;
    end if
  until a node corresponds to a child cell that cannot be explained in L **or** $k := 0$;

different explanations can be generated for a symptom $\partial y(c) = q$. In most practical cases one would therefore apply additional reduction or pruning methods, next to RM_1, yielding a comprehensive tree T of the most important causes, by judgement of the analyst. In this case, Algorithm 5 can also be configured with reduction methods RM_2 and RM_3 (Section 4.6).

Hidden causes might also be present in $\{\downarrow c\}$. The approach for detecting them is similar with the detection of hidden causes in the business model, by substituting equations from a lower level in the system into a higher level. By slightly modifying Algorithm 4 we can identify hidden causes in a system of drill-down equations. Instead of substituting equations in the business model we now substitute drill-down equations over an analysis path p in the lattice as follows. Suppose that $y^{i_q}(c)$ is an additive drill-down measure and that $c \in C$, $c' \in R_q^{-1}(c)$, and $c'' \in R_p^{-1}(c')$. Hidden causes on level $[i_q - 2]$ are identified with one-step look ahead in the following equation

$$y^{i_q}(c) = \sum_{c'' \in R_p^{-1} \circ R_q^{-1}(c)} y^{(i_q - 2)}(c''), \tag{4.14}$$

which is the result of substituting

$$y^{(i_q-1)}(c') = \sum_{c'' \in R_p^{-1}(c')} y^{(i_q-2)}(c'')$$

into equation

$$y^{i_q}(c) = \sum_{c' \in R_q^{-1}(c)} y^{(i_q-1)}(c').$$

The substitution of drill-down equations related to cells in $\{\downarrow c\}$ can be continued, level by level in p, until the drill-down equations related to C_B. Here a similar reasoning is applied as in the proof of Theorem 2.3.1.

In general, notice that different analysis paths in the same lattice L of drill-down equations, corresponding to the same set of drill-down equations in a different order, may produce explanatory trees with slightly different structures. A reason for this phenomenon is the use of specific reduction measures and/or the presence of cancelling-out effects in the data. For example, when parsimonious sets of causes are constructed for a symptom with RM_1, only parsimonious causes are examined further by the algorithm. In that case, a cause might become parsimonious in one analysis path and non-parsimonious in the other.

4.4.2 Greedy explanation

In this section, an exceptional cell $\partial y(c)$ in some cube $C = [i_1 i_2 \dots i_n]$ is explained. This is done in a case where only additive drill-down equations from the exceptional cell's downset $\{\downarrow c\}$ are applied. For this purpose, a greedy method of explanation is proposed that utilizes the transitivity property, a feature which is present in additive systems of drill-down equations. This method is implemented in an algorithm and illustrated in an OLAP sales database. Finally, greedy explanation is discussed in systems of average and maximum/mimimum drill-down equations.

System of additive drill-down equations

If $\partial y(c)$ is an exceptional cell value in a cube C, and y is an additive drill-down measure then this cell value can also be expressed in the cell values at lowers values in the cube, by repeatedly applying additive drill-down equations (see Theorem 2.3.1). For the

determination of the influence of a cell in the downset of c on $y(c)$ we derive expression 4.15, which is similar to expression 4.12. This property is called transitivity:

Theorem 4.4.1. (*Transitivity*). If $C_p = [i_1, i_2, \ldots, i_n]$ and $C_q = [j_1, j_2, \ldots, j_n]$ are cubes in L where $C_q \leq C_p$, $c \in C_p$ and $c' \in C_q$, and y is an additive drill-down measure, then

$$\inf(y^{a;\mathbf{j}}(c'), y^{a;\mathbf{i}}(c)) = y^{a;\mathbf{j}}(c') - y^{r;\mathbf{j}}(c'), \tag{4.15}$$

where $\mathbf{i} = i_1 i_2 \ldots i_n$ and $\mathbf{j} = j_1 j_2 \ldots j_n$, under the conditions of Section 4.2.4.

Proof.

We define $S = R_1^{i_1 - j_1} \circ R_2^{i_2 - j_2} \circ \cdots \circ R_n^{i_n - j_n}$ and $x^{\mathbf{j}} = y^{\mathbf{j}}(c')$ then

$$\begin{array}{l}
\inf(x_i^{a;j}, y^{a;\mathbf{i}}(c)) = f(\mathbf{x}_{-i}^{r;\mathbf{j}}, x_i^{a;\mathbf{j}}) - y^{r;\mathbf{i}}(c) = \\
\text{(by applying Theorem 2.3.1)} \\
\sum\limits_{c' \in S(c)} \mathbf{x}_{-i}^{r;\mathbf{j}} + x_i^{a;\mathbf{j}} - y^{r;\mathbf{i}}(c) = \\
\sum\limits_{c' \in S(c)} \mathbf{x}_{-i}^{r;\mathbf{j}} + x_i^{a;\mathbf{j}} - \sum\limits_{c' \in S(c)} \mathbf{x}^{r;\mathbf{j}} = \\
\sum\limits_{c' \in S(c)} \mathbf{x}_{-i}^{r;\mathbf{j}} + x_i^{a;\mathbf{j}} - (\sum\limits_{c' \in S(c)} \mathbf{x}_{-i}^{r;\mathbf{j}} + x_i^{r;\mathbf{j}}) = x_i^{a;\mathbf{j}} - x_i^{r;\mathbf{j}} = \\
y^{a;\mathbf{j}}(c') - y^{r;\mathbf{j}}(c'). \square
\end{array}$$

In general, the computation of influence values for a symptom $\partial y(c) = q$ with Equation (4.15), is based on the same measure $y(c)$ computed for different cells in $\{\downarrow c\}$. For the application of Equation (4.15) we therefore need both the actual values y^a and the reference values y^r for the symptom's downset $\{\downarrow c\}$. In other words, we need a sublattice L' with base cube C_q on level $[j_1 j_2 \ldots j_n]$ and top cube c on level $[i_1 i_2 \ldots i_n]$ with actual and reference values for all cubes in L'. The actual values y^a are directly available in L' by the application of roll-up operations on the base cube. However, the availability of reference values y^r depends on the type of normative model R that is selected for exception identification. If the normative model is internal then the reference values are simply obtained by roll-up operations on the reference values in base cube C_q and if the normative model is external then the reference values have to be computed for (part of) the symptom's downset. In Section 4.7, we discuss how to construct chains of reference values in L' for each type of normative model R that is applicable.

Theorem 4.4.1 implies that in a system of drill-down equations *the influence of a variable* $y^{\mathbf{j}}(c')$ *on any ancestor variable in its upset* $\{\uparrow c'\}$ *is given by* $y^{a;\mathbf{j}}(c') - y^{r;\mathbf{j}}(c')$. Transitivity greatly simplifies the computation of influence values in the upset of a cell, because we only have to compute the difference between the actual and reference value of a cell, to obtain the influence values on any of its ancestors in its upset, instead of repeatedly applying Equation (4.3). This is illustrated in the following example.

Example 4.4.2. Using the data in Section 4.4.2 the transitivity property reads:

$$\begin{aligned}
&\inf(\text{profit}^{230}(.,.,\text{Golf Equip.Irons.Titan Irons}), \text{profit}^{231}(.,.,\text{Golf Equip.Irons})) = \\
&\inf(\text{profit}^{230}(.,.,\text{Golf Equip.Irons.Titan Irons}), \text{profit}^{232}(.,.,\text{Golf Equip})) = \\
&\inf(\text{profit}^{230}(.,.,\text{Golf Equip.Irons.Titan Irons}), \text{profit}^{233}(.,.,\text{All-Products})) = \\
&5,959 - 2,507 = 3,452.
\end{aligned}$$

With Theorem 4.4.1 we can simplify the definition of causes, as formulated in 4.4, for a lattice system of additive drill-down equations. The set of contributing (counteracting) causes Cb (Ca) for a symptom $\partial y(c)$ in L where y is an additive drill-down measure and $C_p = [i_1 i_2 \dots i_n]$ and $C_q = [j_1 j_2 \dots j_n]$ are cubes in L where $C_q \leq C_p$ and $c \in C_p$ and $c' \in C_q$, consists of the set of successors from the downset $\{\downarrow c\}$ such that

$$\inf(y^{\mathbf{j}}(c'), y^{\mathbf{i}}(c)) \times \partial y^{\mathbf{i}}(c) > 0 \ (< 0). \tag{4.16}$$

An example is given in Section 4.4.2.

Additionally, the transitivity property has implications for the construction of parsimonious sets of causes. Again assume that $C_p = [i_1 i_2 \dots i_n]$, $C_q = [j_1 j_2 \dots j_n]$ and $C_r = [k_1 k_2 \dots k_n]$ are cubes in L where $C_r <= C_q <= C_p$, $c'' \in C_r$, $c' \in C_q$ and $c \in C_p$. The implications are formulated as

$$\begin{aligned}
&y(c') \in \text{Cb}_p(\partial y(c)) \ \wedge \ y(c'') \in \text{Cb}_p(\partial y(c)) \rightarrow y(c'') \in \text{Cb}_p(\partial y(c')), \\
&y(c') \in \text{Ca}_p(\partial y(c)) \ \wedge \ y(c'') \in \text{Ca}_p(\partial y(c)) \rightarrow y(c'') \in \text{Ca}_p(\partial y(c')).
\end{aligned} \tag{4.17}$$

In words, if the variables $y(c')$ and $y(c'')$ are in the parsimonious contributing set of causes for the symptom $\partial y(c)$, then the variable $y(c'')$ is also in the parsimonious set of contributing (counteracting) causes for the symptom $\partial y(c')$. With these implications we can connect (disconnected) causes to form an explanation tree T.

At last, we design an algorithm for the explanation of a symptom $\partial y(c)$. In this algorithm we utilize Equation 4.15 (Theorem 4.4.1) to compute influence values and expression 4.16 for the determination of the symptom's causes. The inputs for the algorithm are a symptom $\partial y(c)$ and an *aggregated influence table*. This table is a generalization of the influence table, for one or more drill-down paths in the symptom's downset, with entries for the actual, norm and influence values. For a database with actual values y^a and reference values y^r and $c \in C$ the influence values are computed as $\inf(y^a(c'), y^a(c)) = y^a(c') - y^r(c')$ for all $c' \in \{\downarrow c\}$. The aggregated table is composed out of a separate column for each dimension level and columns for all its actual, norm, and influence values. Each dimension level corresponds to a record in the table. The general form of the aggregated influence table for a drill-down path within dimension D_q is given in Table 4.3. In this table we fill in the dimension

Table 4.3: General form of the aggregated influence table for dimension D_q.

$D_q^{i_q-1}$	$D_q^{i_q-2}$	...	D_q^0	Norm values	Actual values	Inf. values
...	...	...	...	...	...	...

level instances in the appropriate columns level by level. For an aggregated dimension level instance we fill in the term 'All' for its successors in the corresponding columns on its right hand side in the table. In Table 4.4, an example is presented of an aggregated table for the Product dimension and the measure profit. The main reason to construct this table is to have one joint structure with all the influence values for cells in (part of) the symptom's downset. By ranking the influence values in this table we can easily determine significant causes for the symptom under consideration.

Basically, the algorithm for explanation is composed of three main steps. In the first step the aggregated table is constructed for the symptom's downset. In the second step the causes are determined greedily in the aggregated table by selecting the n largest causes and filtered by application of heuristics. In the final step the explanation tree is constructed possibly, as the algorithm's output. The pseudo code for the algorithm is given in Algorithm 6.

Compared with top-down explanation (Section 4.4.1), greedy explanation always identifies the largest causes - independently from their level in the aggregation lattice

Algorithm 6 Greedy algorithm for explanation in a system of additive drill-down equations

Initialization:
  $\partial y^{i_1 \cdots i_q \cdots i_n}(c) = q$: a symptom in context cube C from L;
  downset $\{\downarrow c\}$ with actual and reference values, making up L' with top (c) and base (C_B);
  analysis path p in L';
  reduction methods taken from $\{\mathrm{RM}_1, \mathrm{RM}_2, \mathrm{RM}_3, \mathrm{RM}_4\}$;
Computation:
  Construct the aggregated table for the symptom's downset based the path p;
  For each record in the aggregated table compute the influence value with Equation (4.15);
  Determine contributing and counteracting causes with Equation (4.16);
  Sort the influence values in the table with a sorting algorithm;
  repeat {Greedily determine the largest causes for $\partial y^{\cdots i_q \cdots}(c) = q$ the root of T}
    Add to the root node (T^0) the successor contributing and counteracting variables with the highest influence values;
  until the n-th largest cause is determined *or* the T^+ (T^-) is explained on each cube in P;
  repeat {Representation of explanation tree: Reorganise T in line with Equation (4.17)}
    Add an edge between a descendant and its most direct ancestor;
    Remove the edge between the descendant and its most distant ancestor.
  until For each cause its ancestry is determined in the downset.

- because significant causes are determined globally over the whole (or at least a large part of) symptom's downset, instead of locally per drill-down equation. We illustrate this notion with the following example. Suppose that in Figure 4.1 all equations are additive, where $y = x_1 + x_2$ and $x_2 = z_1 + z_2$, and that the $\inf(x_1, y) = 10$, $\inf(x_2, y) = 1$, $\inf(z_1, y) = 100$, and $\inf(z_1, y) = -99$. In this situation, greedy explanation first identifies z_1 as the largest contributing cause, after that it identifies z_2 as the largest counteracting cause, etc. It is obvious that $z_1 \in \mathrm{Cb}_p(y)$ is determined when $T^+ = 0.9$.

Example: Greedy explanation in financial data

In this section, we present an example of the explanation of a symptom in the Product dimension of an OLAP database with sport equipment financial figures, with greedy explanation (Algorithm 6). The star schema of the database is depicted in Figure 2.1. The specification of the symptom to be explained is: $\langle \mathrm{profit}^a(2001, \text{Netherlands}, \text{All-Products}),\ \partial\mathrm{profit}^{233} = -9,803 = \text{“low”},\ \mathrm{profit}^r(2000, \text{Netherlands}, \text{All-Products}) \rangle$. The algorithm is configured with $n = 20$ (RM_4), to present the 20 largest contributing and counteracting causes to the business analyst. Table 4.4, presents the aggregated

table for the explanation of this symptom. The table contains all the 141 combinations of instance elements of the Product dimension's hierarchy, ordered over the absolute values of their influence values from high to low. Because we analyse a low exception in this example, a negative influence value indicates a contributing cause and a positive influence value indicates a counteracting cause in the table.

Figure 4.2 shows two intermediate substeps of Algorithm 6 after 3 rounds, i.e. the algorithm is configured for $n = 3$. The upper figure shows the first step that determines in this case the first 3 largest causes: Golf Equipment of the ProductLine level as a contributing cause $(-7,958)$, Golf Equipment.Woods on the ProductType level as a contributing cause $(-3,277)$, and Golf Equipment.Irons.Titanium Iron on the Product level as a counteracting cause $(+3,452)$. In notation, Cb= {Golf Equipment.Woods, Golf Equipment} and Ca= {Golf Equipment.Irons.Titanium Iron}. The second step depicted in the lower figure reorganises the explanation tree consistent with the structure of the dimension hierarchy with Equation (4.17). Here an edge is connected in the explanation tree between the node Woods and the node Golf Equipment and the direct edge between the symptom and the node Woods is removed. Besides, an edge is added between Titanium Iron and Golf Equipment, and the edge between the symptom and Titanium Iron is removed. In this manner, the causes identified by greedy explanation are mapped again to the hierarchy of the Product dimension to show their ancestry. Accordingly, the identified causes are presented intuitively to the business analyst and are accessible for drill-down operations.

Figure 4.3 shows the explanation tree T for the symptom with the 20 largest causes identified in the Product dimension. The algorithm identified 14 significant contributing causes and 6 significant counteracting causes. The sets of causes are given by Cb= {Golf Equipment.Woods, Golf Equipment, ..., Golf Equipment.Golf Acc.Pro Golf Bag} and Ca= {Golf Equipment.Irons.Titanium Iron, Camping Equipment.Tents, ..., Golf Equipment.Irons}. In the explanation tree contributing causes are indicated with a straight line and counteracting causes are indicated with a dotted line. Symptoms that have the reverse direction compared to the root symptom are indicated with an uparrow $\uparrow$ for a high symptom and with a downarrow $\downarrow$ for a low symptom.

The complete tree in the figure describes why the sales in the Netherlands were

Table 4.4: Aggregated table for the Product dimension where the actual object is the year 2001, the norm is the year 2000, and the influence values for instances within the Product dimension are related to the exceptional cell profit[233](2001, The Netherlands, All-Products).

Nr.	ProductLine P^2	ProductType P^1	Product P^0	Norm (2000)	Actual (2001)	Influence
	All	All	All	156,658	146,855	
1	Golf Equip.	Woods	All	34,493	26,116	-8,377
2	Golf Equip.	All	All	54,999	47,041	-7,958
3	Golf Equip.	Irons	Titan. Ir.	2,507	5,959	3,452
4	Golf Equip.	Woods	St. Woods	10,440	7,420	-3,020
5	Camp. Equip.	Tents	All	28,685	31,256	2,571
6	Mount. Equip.	All	All	21,735	19,235	-2,500
7	Golf Equip.	Woods	Hail. T. Wds	7,134	4,830	-2,304
8	Golf Equip.	Woods	Lady Hail. T. Wds	9,152	7,198	-1,954
9	Camp. Equip.	All	All	57,521	59,090	1,569
10	Golf Equip.	Irons	Lady Hail. T. Ir.	4,236	2,800	-1,436
11	Mount. Equip.	Rope	All	11,252	9,998	-1,254
12	Camp. Equip.	Tents	Star Gazer 6	5,483	6,620	1,137
13	Golf Equip.	Irons	Hail. St. Ir.	5,500	4,400	-1,100
14	Golf Equip.	Woods	Lady Hail. St. Wds	7,767	6,668	-1,099
15	Camp. Equip.	Tents	Star Gazer 2	5,400	6,322	922
16	Mount. Equip.	Tools	All	6,062	5,289	-773
17	Golf Equip.	Irons	All	15,764	16,503	739
18	Mount. Equip.	Rope	Husky Rope 200	4,425	3,757	-668
19	Pers. Acc.	All	All	21,727	21,104	-623
20	Golf Equip.	Golf Acc.	Pro Golf Bag	1,990	1,432	-558
...	...	...	...	...	...	...
141	Mount. Equip.	Tools	Gran. Extreme	1580	1664	84

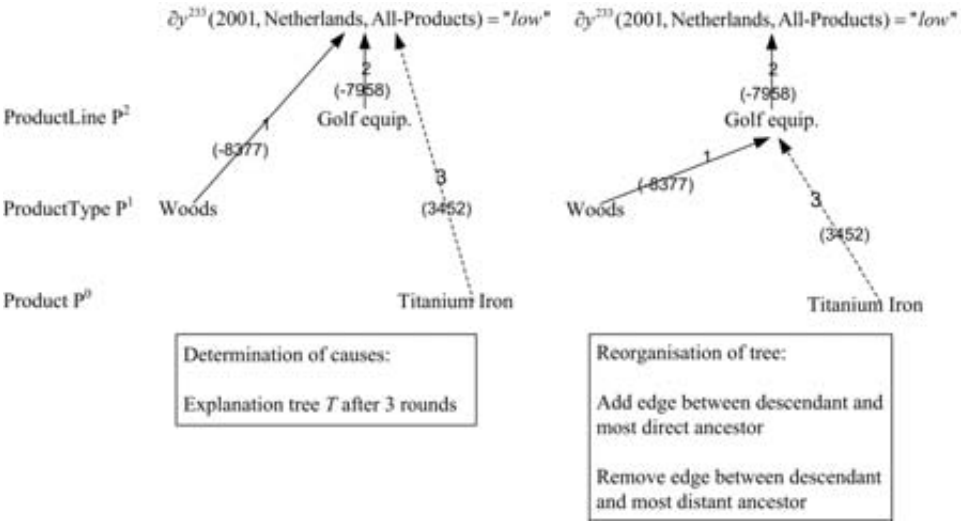


Figure 4.2: Illustration of two intermediate steps of the greedy algorithm for explanation in the Product dimension. The right figure shows the determination step, where the 3 largest causes are identified. The left figure shows the reorganisation step, where the causes are re-organised based on the structure in the dimension hierarchy.

quite low in the year 2001 compared to the previous year within this dimension. A brief business interpretation of the explanation tree reads as follows. The largest causes that explain the difference are identified as elements of the Golf Equipment ProductLine. Striking causes are products in the ProductType Woods, that have performed rather badly $(-8,377)$. Conversely, the ProductType Irons performed relatively well - depicted in the three as a relatively large counteracting cause $(+739)$. Although, the ProductType Irons as a whole performed positively, indicated with an uparrow, because of the large contributing cause Hailstorm Titanium Irons $(+3,452)$, it does have two large counteracting causes associated with it namely Lady Hailstorm Titanium Irons $(-1,436)$ and Hailstorm Steel Irons $(-1,100)$.

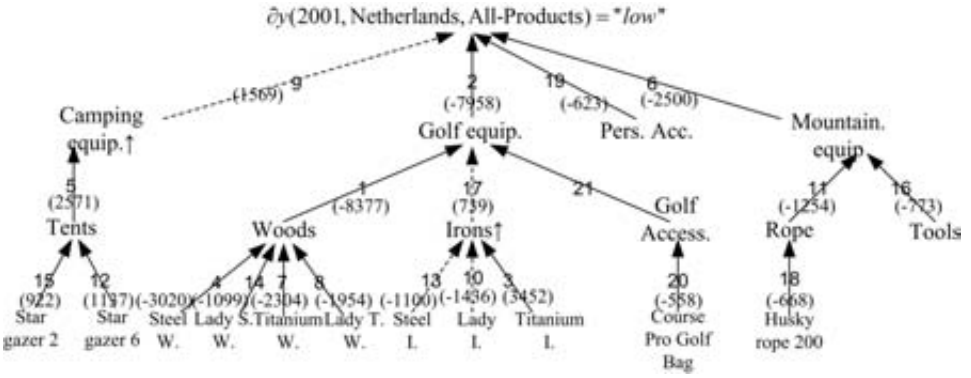


Figure 4.3: Illustration of the explanation tree T for a symptom using the Product dimension with the algorithm for greedy explanation. This explanation tree presents the 20 most largest contributing and counteracting causes to the analyst.

System of average or maximum/minimum drill-down equations

For the greedy explanation of an exceptional cell value $\partial y(c) = q$, in a system of average drill-down equations, we determine $\inf(y^{a;\mathbf{j}}(c'), y^{a;\mathbf{i}}(c))$:

Theorem 4.4.2. (*Influence measure for systems of average drill-down equations*). If $C_p = [i_1 i_2 \dots i_n]$ and $C_q = [j_1 j_2 \dots j_n]$ are cubes in L where $C_q \leq C_p$, $c \in C_p$, $c' \in C_q$, and $S = R_1^{i_1 - j_1} \circ R_2^{i_2 - j_2} \circ \dots \circ R_n^{i_n - j_n}$, $\mathbf{i} = i_1 i_2 \dots i_n$, $\mathbf{j} = j_1 j_2 \dots j_n$, and

$$y^{a;\mathbf{i}}(c) = \tfrac{1}{|C_q|}\Big(\sum_{c' \in S(c)} y^{a;\mathbf{j}}(\mathbf{c}')\Big) \text{ and}$$
$$y^{r;\mathbf{i}}(c) = \tfrac{1}{|C_q|}\Big(\sum_{c' \in S(c)} y^{r;\mathbf{j}}(\mathbf{c}')\Big),$$

where y is the average drill-down measure (2.3.2), then

$$\inf(y^{a;\mathbf{j}}(c'), y^{a;\mathbf{i}}(c)) = \frac{1}{|C_q|}(y^{a;\mathbf{j}}(c') - y^{r;\mathbf{j}}(c')). \tag{4.18}$$

Proof. We define $x^{\mathbf{j}} = y^{\mathbf{j}}(c')$.

$$\begin{array}{l}
\inf(x_i^{a;\mathbf{j}}, y^{a;\mathbf{i}}(c)) = f(\mathbf{x}_{-i}^{r;\mathbf{j}}, x_i^{a;\mathbf{j}}) - y^{r;\mathbf{i}}(c) = \\
\frac{1}{|C_q|} \sum\limits_{c' \in S(c)} \mathbf{x}_{-i}^{r;\mathbf{j}} + \frac{1}{|C_q|} x_i^{a;\mathbf{j}} - y^{r;\mathbf{i}}(c) = \\
\frac{1}{|C_q|} \sum\limits_{c' \in S(c)} \mathbf{x}_{-i}^{r;\mathbf{j}} + \frac{1}{|C_q|} x_i^{a;\mathbf{j}} - \frac{1}{|C_q|} \sum\limits_{c' \in S(c)} \mathbf{x}^{r;\mathbf{j}} = \\
\frac{1}{|C_q|} \sum\limits_{c' \in S(c)} \mathbf{x}_{-i}^{r;\mathbf{j}} + \frac{1}{|C_q|} x_i^{a;\mathbf{j}} - (\frac{1}{|C_q|} \sum\limits_{c' \in S(c)} \mathbf{x}_{-i}^{r;\mathbf{j}} + \frac{1}{|C_q|} x_i^{r;\mathbf{j}}) = \\
\frac{1}{|C_q|} x_i^{a;\mathbf{j}} - \frac{1}{|C_q|} x_i^{r;\mathbf{j}} = \\
\frac{1}{|C_q|} (y^{a;\mathbf{j}}(c') - y^{r;\mathbf{j}}(c')). \square
\end{array}$$

Here transitivity does not hold, because of the form of expression 4.18. The influence of a variable $y^{a;\mathbf{j}}(c')$ on elements in its upset $\{\uparrow c'\}$, will usually decreases per roll-up in the lattice L, because the number of cells in the denominator of the expression will increase, while its numerator $(y^{a;\mathbf{j}}(c') - y^{r;\mathbf{j}}(c'))$ remains constant. However, because of Theorem 4.4.2 we can use Algorithm 6, albeit with an extra computation, for the explanation of an exceptional cell value $\partial y(c)$, where y is an average drill-down equation, within the exceptional cell's downset $\{\downarrow c\}$. An important result is that we can construct an aggregated influence table for the symptom, where the influence values from elements in its downset can be sorted from high to low, and that significant causes can be determined in the sorted table accordingly. To construct this table we need to store the number of cells of each cube C in the analysis path to compute Equation (4.18).

In addition, for the greedy explanation of an exceptional cell in a system of maximum drill-down equations (see Equation 2.18) we give the expression for the influence measure. Suppose that y is a maximum drill-down measure and $x^{\mathbf{j}} = y^{\mathbf{j}}(c')$ then $\inf(x^{a;\mathbf{j}}, y^{a;\mathbf{i}}(c))$ is given by the following two cases,

1. if $x_i^{r;\mathbf{j}}$ was the maximum, denoted by $x_i^{r;\mathbf{j}}(c) = y^{r;\mathbf{i}}(c)$, then

$$\inf(x_i^{a;\mathbf{j}}, y^{a;\mathbf{i}}(c)) = \begin{cases} x_i^{a;\mathbf{j}} - x_i^{r;\mathbf{j}} & \text{if } x_i^{a;\mathbf{j}} = \max(\mathbf{x}_{-i}^{r;\mathbf{j}}, x_i^{a;\mathbf{j}}), \\ x_l^{r;\mathbf{j}} - x_i^{r;\mathbf{j}} & \text{if } x_l^{r;\mathbf{j}} = \max(\mathbf{x}_{-i}^{r;\mathbf{j}}, x_i^{a;\mathbf{j}}), \end{cases} \tag{4.19}$$

2. if $x_i^{r;\mathbf{j}}$ was not the maximum, but $x_l^{r;\mathbf{j}} = y^{r;\mathbf{i}}(c)$, then

$$\inf(x_i^{a;\mathbf{j}}, y^{a;\mathbf{i}}(c)) = \begin{cases} x_l^{r;\mathbf{j}} - x_l^{r;\mathbf{j}} = 0 & \text{if } x_l^{r;\mathbf{j}} = \max(\mathbf{x}_{-i}^{r;\mathbf{j}}, x_i^{a;\mathbf{j}}), \\ x_i^{a;\mathbf{j}} - x_l^{r;\mathbf{j}} & \text{if } x_i^{a;\mathbf{j}} = \max(\mathbf{x}_{-i}^{r;\mathbf{j}}, x_i^{a;\mathbf{j}}). \end{cases} \tag{4.20}$$

The expression for the influence measure for a minimum drill-down equation (see Equation 2.19) is defined similarly. Because of Equations (2.20) and (2.21) we can use Algorithm 6 for explanation generation in systems of maximum (minimum) equations. The result of the algorithm is an explanation tree where all the dimension attributes have the maximum (minimum) value.

4.5 Explanation in a system of mixed equations

In this section, explanation in a system of mixed equations is discussed. For this purpose we develop a combined approach. Where multi-level explanation (see Algorithm 4) is applied, configured with or without look-ahead, if a business model equation is evaluated, and top-down explanation (see Algorithm 5) or greedy explanation (see Algorithm 6) is applied, if a drill-down equation is evaluated.

In the application of multi-level explanation in the business model M of a multi-dimensional database, the explanation process of a symptom $\partial y(c) = q$ is continued *top-down* from M^0 until M^d. Formally, this procedure is stated as

$$\inf(x_i(c), y(c)) = f(\mathbf{x}^r_{-i}(c'), x^a_i(c)) - y^r(c'), \tag{4.21}$$

where the inf-measure is evaluated on the actual cell c and the reference cell c' from cube $C = [i_1 \dots i_q \dots i_n]$. Here the explanation procedure is based on the computation of the inf-measure for *different measures* $x_i(c)$ *from* M *that are evaluated on the same cell* c within the context cube C. In the case that R is a statistical normative model then $c = c'$ and in the case that R is a managerial normative model then c differs from c' in the single dimension attribute d_q that is selected for reference, such that $c = (d_1, \dots, d_q, \dots, d_n)$ and $c' = (d_1, \dots, d'_q, \dots, d_n)$. Naturally, if the cell c in the cube C is selected by the OLAP analyst, the above expression for the inf-measure simply reduces to expression (4.3).

For explanation in a mixed system of equations a combined analysis path is required over a path in the aggregation lattice and the business model. Such a path is composed out of series of drill-down operations over the exceptional cell's downset alternated with series of "drill-down operations" in the exceptional cell's business model. Such a combined analysis path can be represented in a 3-dimensional analysis cube, a

straightforward extension of the 2-dimensional analysis table, as presented on page 33. In this cube the columns represent the dimensions of the cube from $D_1, D_2, \ldots, D_n$, the rows represent the levels of the lattice L from level $(i_1 + i_2 + \ldots + i_n)$ to level $(j_1 + j_2 + \ldots + j_n) + 1$, and the layers represent the levels in the business model from level M^0 to level M^d. A cell with the value -1 in the analysis cube represents a drill-down in the exceptional cell's downset or a drill-down in its associated business model from level M^p to level M^{p+1}.

Example 4.5.1. The symptom to be explained is $\text{profit}^{111}(c)$. Here the drill-down equations are derived from the lattice of cubes in Figure 2.4 and the business model equations are derived from the sales model given in Table 1.1. The combined analysis path is given by $[1,1,1] \rightarrow [0,1,1] \rightarrow [0,0,1] \rightarrow [0,0,0] \rightarrow M^{0;1} \rightarrow M^{1;2} \rightarrow M^{2;3}$ (see Figure 4.4).

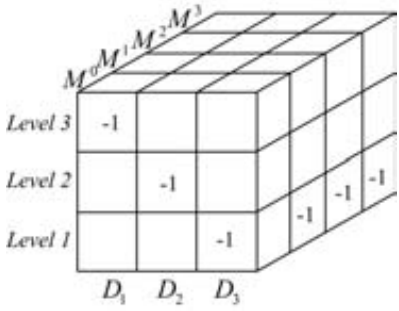


Figure 4.4: Example of an combined analysis path over both drill-down and business model equations.

Additionally, we propose a straightforward algorithm for explanation in a mixed system of equations (Caron and Daniels 2005). In this algorithm, Algorithm 4 is invoked if a business model equation is considered in the combined analysis path and Algorithm 5 or 6 is invoked if a drill-down equations is considered in the combined analysis path. The pseudo-code for the algorithm is presented in Algorithm 7.

Algorithm 7 Algorithm for explanation in a mixed system

Initialization:
  S: $y^{i_1 i_2 \cdots i_n}(c) = q$: a symptom in cube C;
  analysis path $p[t]$ in L;
  L: symptom's downset $\{\downarrow c\}$ with actual and reference values;
  M: business model with actual and reference values;
  reduction methods taken from $\{\mathrm{RM}_1, \mathrm{RM}_2, \mathrm{RM}_3\}$;
Computation:
  t=1;
  repeat {Explanation in a hybrid system of equations}
    if Equation in $P[t]$ is a business model equation in M **then**
      invoke Algorithm 4 with parameters $(S, M, \mathrm{RM}, q = 0)$
    else if Equation in $P[t]$ is a drill-down equation in L **then**
      invoke Algorithm 5 or 6 with parameters (S, L, RM);
    end if
    t=t+1;
  until the last step in the analysis path $P[t]$;

4.6 Reducing information overload

Because every applicable equation in the multi-dimensional database yields a possible explanation, the number of explanations obtained for a single symptom can be very large. It is equal to the number of paths from the cell to the base, see formula (2.2.3) in Section 2.2.5. In order to avoid information overload, we can reduce the number of explanations, by applying one or more *reduction methods*, denoted in shorthand by RM. We propose five generic reduction methods (RM_1-RM_5) to reduce the number of explanations.

4.6.1 Parsimonious causes (RM_1)

Feelders and Daniels (2001) proposed a reduction method to construct parsimonious sets of causes, denoted by RM_1, as described in Section 4.2.3 and expression 4.5. Obviously, RM_1 can also be applied on explanations generated for symptoms in a system of drill-down equations. Additionally, the value for T^+ (T^-) has to be determined by the business analyst by adapting the value to the internal structure of the multi-dimensional database. By inspecting the generated explanation trees iteratively the analyst makes a selection between significant and insignificant causes. In our experiments with multi-dimensional data sets, described in Chapter 6, it was found that

fractions with values between 0.7 and 1 are often appropriate.

4.6.2 Specificity (RM_2)

The number of explanations is reduced by applying a *measure of specificity* for each applicable equation in the symptom's downset or business model. This measure, denoted by RM_2, quantifies the specificity or "interestingness" of the explanation step. The measure is defined as:

$$S = \text{specificity} = \frac{\#\text{ possible causes}}{\#\text{ actual causes}}. \tag{4.22}$$

The number of possible causes is the number of RHS elements of each equation, and the number of actual causes is the number of elements in the parsimonious set of causes. In general, we prefer explanation steps with a relatively high specificity value. Using this measure we can order the explanation paths from specific to general and if desired only list the explanation path in the most specific dimension(s). To do this each dimension has to be explored to compute S, only the explanation step in the dimension for which S is maximal is reported. In particular, if we explain a symptom $\partial y^{i_q}(c)$ solely by applying drill-down equations from the downset $\{\downarrow c\}$, we can write the measure of specificity as

$$S = \frac{|D_q^{i_q-1}|}{|\mathrm{Cb}_p| + |\mathrm{Ca}_p|}, \tag{4.23}$$

for a dimension D_q. Notice that this measure is to some extent similar to the rule evaluation measure "specificity" used in association rule learning (Lavrac et al. 1999).

4.6.3 Reduction heuristic (RM_3)

The number of explanations for an exceptional cell $\partial y(c)$ can also be limited simply by reducing the number of drill-down equations (i.e. cubes) in the analysis path by some criterium of interestingness. In other words, the total set of applicable drill-equations is reduced to a smaller set of "interesting" drill-down equations. This criterium of interestingness, denoted by RM_3, is formed by the application of a number of typical *reduction heuristics*:

RM_{3a} Explanation in an user-defined analysis path p in the downset $\{\downarrow c\}$. For example, the analyst wants to explain the symptom in the following sequence of dimensions over the downset: Time (on a high level), Location (on an intermediate level), and Product (on a low level).

RM_{3b} Explanation in a single dimension D_q from the downset $\{\downarrow c\}$. For example, the analyst is only interested in an explanation in the Product or Location dimension, to identify the causes in each of these dimensions separately. Here the algorithm for explanation has to be executed for each selected dimension.

4.6.4 Select the largest causes (RM_4)

In Algorithm 6 a specific reduction method can be applied, denoted by RM_4, that lets the analyst select the number of significant contributing (counteracting) causes he/she wants to explore for a particular symptom. In this way the analyst can simply select only the n largest contributing and/or counteracting causes. Algorithm 6 first identifies the largest cause, then that the second largest cause and so on, until the n largest causes are found. The reduction method can be configured for a single dimension D_q or for multiple dimensions at the same time. For example, the analyst can generate a top-10 list of largest causes for only the Product dimension or for all available dimensions in the exceptional cells's downset.

The drawback of RM_4 is that the choice for a certain n by the analyst is rather arbitrary. In this manner the analyst might miss a number of "large causes" that are just out of the selected set. To address this issue, we can combine the concept of parsimony in RM_1 with greedy explanation, to construct an alternative version of RM_4, as a compound reduction measure. In this approach we greedily explain the symptom for a certain dimension and stop explaining on a certain dimension level until the desired fraction T^+ is explained. The selected fraction can hold for a number of dimensions, a single dimension, or for a single level in the dimension hierarchy. For example, if we greedily explain a symptom in a dimension D_q, with three levels in the dimension hierarchy, with fraction $T^+ = 0.9$, the reduction will explain at least 90% of the difference on each of the three levels in the dimension hierarchy.

4.6.5 Similarity reduction (RM_5)

In the problem identification phase the analyst selects a set of symptoms from a context cube C. In general, only a fraction of the cells in a cube is taken into account for explanation. Sometimes it might be interesting from a business perspective to explain a whole range of similar exceptional cells. Quite often the explanation trees T will also be similar. In that case one could report the *similarity patterns* or *generic explanations* in the generated trees (RM_5). Similarity is defined as common significant contributing and counteracting causes, i.e. branches that the trees have in common. The common contributing causes for exceptional cell values $\partial y(c_1)$, $\partial y(c_2)$, ..., $\partial y(c_n)$, in the context cube $C = [i_1 i_2 \dots i_n]$ are given by

$$\mathrm{Cb}_{\text{similarity}} = \bigcap_{i=1}^{n} \mathrm{Cb}(\partial y(c_i)), \tag{4.24}$$

where n is the number of cells in the range. An equivalent approach, based on graph theory, is to simply determine the maximal common subtree for the set of generated explanation trees.

Similarity patterns might answer questions as whether we get generic explanations for corresponding symptoms. For example, if we explain low revenues in some sales cube, the question might be whether we see the same pattern for all countries where the company is active or for all products groups the company sells.

In contrast with the other reduction methods, that work on the level of an individual explanation for an exceptional cell, RM_5 works on the level of a group of explanations, to produce a generic explanation for a range of cells in a context cube. In order to be able to create generic explanation for a range of cells, we first need to give explanations for each individual cell with a top-down explanation (see Algorithm 5) or a greedy explanation (see Algorithm 6), configured with a fixed subset of reduction methods out of $\{RM_1, RM_2, RM_3, RM_4\}$. Subsequently, the detection of generic explanation is divide into three basic steps:

1. Determine the range of cells c_1, c_2, ..., c_n in the context cube C;

2. Generate for each cell in the range an explanation tree, with a specific algorithm for explanation and a fixed set of appropriate reduction methods;

3. Determine the similarity pattern between the generated explanation trees with expression (4.24).

In the above approach similarity is defined as structural similarity, i.e. similarity between the nodes of the explanation trees, independent of the weights (i.e., influence values) of the branches. Alternatively, for future research it might be interesting to develop a method to identify similarity patterns that are defined in terms of corresponding values for the influence values within some bandwidth. This approach would produce similarity patterns that give more information.

4.7 Consistency of reference values

In this section, we investigate under what conditions the reference values satisfy the functional equations of OLAP or business model equations, i.e. under what conditions the consistency constraint (see Definition 4.1) is satisfied. We discuss how consistent reference objects can be formed for the different types of normative models, that are discussed in Chapter 3. Actual values in the OLAP context are consistent because they satisfy the drill-down equations (Equation (2.12)) or business model equations (Equation (2.22)) by definition. However, for each type of normative model R, it has to be verified whether these equations also hold for the reference values. Often reference values are computed directly from the actual values in the multidimensional database. When the business model or OLAP equation (f) commute with the operator that computes the reference values (R), the consistency constraint holds, because then $y^r = R(y) = R(f(\mathbf{x})) = f(R(\mathbf{x})) = f(\mathbf{x}^r)$.

There is a natural canonical way to construct a *consistent chain of reference objects* if the above requirement is satisfied. If the chain is formed with strictly drill-down equations, we can create a path in the downset of $\{\downarrow c\}$ level by level, with actual and reference values for successors of c. And if the chain is formed with strictly equations from the business model M, we can obtain a business model with actual and reference values for the business measures. In the remainder of this section we discuss for each type of normative model R and for each type of equation, drill-down and business model, how consistent reference values can be constructed.

4.7.1 R is a planning/budget model

If R is selected to be a planning/budget model (Section 3.2.1), $y^r(C)$ is usually computed directly from the cube with actual values $y^a(C)$. If we explain an exceptional value by a drill-down equation y (Definition 2.20), $y^{r;i_q}(c)$ is determined by

$$y^{r;i_q}(c) = p \cdot y^{a;i_q}(c),$$

where p is some constant value that specifies the budget increase/decrease, then the reference values for $y^{i_q-1}(c')$ are determined by the model

$$y^{r;i_q-1}(c') = p \cdot y^{a;i_q-1}(c').$$

And if an exceptional value is explained by a business model equation (Definition 2.22), the formation of reference values depends obviously on the functional form of the function f (Section 4.2.4).

4.7.2 R is an extra/intra-organizational model

If R is an extra-organizational model (Section 3.2.2), we need to construct consistent reference values, typically composed out of branch averages, for business equations in M. Whether the consistency constraint holds, depends on the form of the function f in the business equation. If the function f is additive, consistent reference values are obtained. Consider the additive business model equation on level i_q

$$y^{i_q}(c) = \sum_{i=1}^{n} x_i^{i_q}(c).$$

The branch average for the measure $y^{i_q}(c)$ is given by

$$y^{r;i_q}(c) = \frac{1}{|D_q^{i_q-1}|} \sum_{c' \in R_{D_q}^{-1}(c)} y^{i_q-1}(c'),$$

and the branch average for the measure $x_i^{i_q}(c)$ is given by

$$x_i^{r;i_q}(c) = \frac{1}{|D_q^{i_q-1}|} \sum_{c' \in R_{D_q}^{-1}(c)} x_i^{i_q-1}(c').$$

We want to show that:

$$y^{r;i_q}(c) = \sum_{i=1}^{n} x_i^{r;i_q}(c).$$

Now:

$$\sum_{i=1}^{n}\left(\frac{1}{|D_q^{i_q-1}|}\sum_{c'\in R_{D_q}^{-1}(c)} x_i^{i_q-1}(c')\right) = \frac{1}{|D_q^{i_q-1}|}\sum_{c'\in R_{D_q}^{-1}(c)} y^{i_q-1}(c') = y^{r;i_q}(c).\square$$

If the function f is non-additive, see Example 4.2.1, the consistency constraint can be violated.

In addition, if R is an intra-organizational model, reference objects are available internally in the database, and determined in a similar way as with an historical model, as described in the next section (see Section 4.7.3). However, here the slice operation is used in other dimensions than the Time dimension. For example, from the Location dimension we select a certain business unit or from the Product dimension we select a certain product group as an intra-organizational reference object.

4.7.3 R is a historical model

If R is selected as a historical model (Section 3.2.3), the reference objects are directly available in the cube. The historical reference objects are determined by a specific slice operation on the Time dimension, where, e.g. the previous year is selected as the norm. Because the reference objects are just cells in a cube C, the consistency of reference values in drill-down equations is guaranteed by definition. Here we assume that the first dimension in a cube C, $D_1^{i_1}$, represents the Time dimension T^{i_1}, and we write $D_1^{i_1} = T^{i_1}$. In general, the child reference cells in the case of pairwise comparison for a parent cell c are determined by

$$y^{r;i_q}(S^{\text{Time}=t}(c)) = \sum_{c'\in R_q^{-1}(c)} y^{r;i_q-1}(S^{\text{Time}=t}(c')).$$

In this case the actual cell values are identical to the reference cell values, except for the Time dimension.

For explanation in the business model M, the reference values must satisfy Equation (2.22), while maintaining the Slice operation on the Time dimension on the same

cell, written as $S^{\text{Time}=t}(c)$. Because historic reference objects for M are based on internal values in the database, Equation (2.22) is consistent for the reference values.

Notice that, when the historical model is selected as the average over a number of periods, which corresponds to the application of an additive ANOVA model with only one main effect included for a single dimension, Theorem 4.7.1 on page 119 can be applied directly.

4.7.4 R is a statistical model

Statistical models, in general, lead to non-consistent reference values, because many statistical models have multiplicative terms. An exception to this general rule are additive ANOVA models, which include main-effects ANOVA models (Section 3.4). Suppose that A_1 is an additive ANOVA model and A_2 is an additive ANOVA model. Reference values are computed by $y^r = A_1(y^a)$ and $\mathbf{x}^r = A_2(\mathbf{x}^a)$. The model is consistent if $A_1 \circ f = f \circ A_2$ because then $y^r = A_1(y^a) = A_1 \circ (f(\mathbf{x}^a)) = f \circ (A_2(\mathbf{x}^a)) = f(\mathbf{x}^r)$.

Here we state that, the reference values are consistent, if and only if, the ANOVA model used for the child cell is a *specialisation* of the ANOVA model used for the parent cell. With a specialized ANOVA model we mean a model that is the result of a drill-down operation on one effect $\lambda_q(D_q^{i_q})$ in the ANOVA model of the parent cell.

Theorem 4.7.1. (*Consistency of ANOVA models*). If reference values are computed with ANOVA models for $y^{i_q}(c)$ and $y^{i_q-1}(c')$, consistency holds if

1. the ANOVA model is additive, i.e. contains no interaction effects, and
2. the ANOVA models at both levels are the same in each dimension, or it is a model with a specialisation for dimension q to which the drill-down operator is applied, corresponding to the lower level $i_q - 1$ of aggregation.

The cube C in which exceptional values are identified by an ANOVA model, determines the context cube in which the reference values are computed for explanation by some equation. The constitution of the context cube depends on the type of equation selected for explanation. In the case of a drill-down equation, we might explain an exceptional cell $\partial y(c)$ in the context cube $C = [i_1 \dots i_q \dots i_n]$ with a main-effects

ANOVA model in the direction of dimension D_q. Now the reference values for the variables on the RHS of Equation (2.12) have to be determined in the context cube $C' = R^{-1}_{D_q}(C)$. In the case of an additive business model equation, we might explain an exceptional cell $\partial y(c)$ in a context cube C with a main-effects ANOVA model in the business model M. Now the reference values for the variables on the RHS of Equation (2.22) have to be determined in the same context cube C. However, the cube C is now associated with the measures present on the RHS of the business equation under consideration.

In the proof of Theorem 4.7.1 we distinguish between two typical cases:

Case 1) within a dimension $D_1^{i_1}$ which is not unfolded;

Case 2) within a dimension $D_1^{i_1}$ which is unfolded.

In these two cases we consider a cube $C = (D_1^{i_1})$ for an additive drill-down measure y and where $c \in C$, $c = (d_1^{i_1})$, and R is a main-effects ANOVA model. Notice that without loosing generality the proof holds for an arbitrary dimension D_q in a cube $D_1^{i_1} \times \ldots \times D_q^{i_q} \times \ldots \times D_n^{i_n}$ and for a main-effects ANOVA model that consists of any proper subset of effects.

Proof.

Case 1) The specialization of a main-effects ANOVA model within a dimension $D_2^{\max_2}$ which is not unfolded.

Suppose we have a parent cube $C = (D_1^{i_1}, D_2^{\max_2})$, where $c \in C$, and a child cube $C' = R^{-1}_{D_2}(C)$, where $c' \in C'$, and the additive measure y with

$$y^{i_1 \max_2}(c) = \sum_{c' \in R^{-1}_{D_2}(c)} y^{i_1 (\max_2 - 1)}(c').$$

The expected value for $y^{i_1 \max_2}(c)$ is computed by the additive ANOVA model (see Equation (3.3))

$$\begin{aligned} \hat{y}^{i_1 \max_2}(c) = & \ \mu_+ \\ = & \ \frac{1}{|D_1^{i_1}|} \sum_{c \in R^{-1}_{D_1}(c_p)} y^{i_1 \max_2}(c), \end{aligned}$$

where $c_p \in R^{+1}_{D_1}(c)$, and the expected value for $y^{i_1 - 1}(c')$ is computed by the specialized ANOVA model

$$\hat{y}^{i_1 (\max_2 - 1)}(c') = \mu_- + \lambda(R^{-1}_{D_2}(d_2^{\max_2})) = \mu_- + \lambda(d_2^{\max_2 - 1}),$$

where

$$\begin{aligned}\mu_- &= \frac{1}{|D_1^{i_1}||D_2^{\max_2-1}|}\sum_{c\in R_{D_1}^{-1}(c_p)}\sum_{c'\in D_2^{-1}(c)} y^{i_1(\max_2-1)}(c') \\ &= \frac{1}{|D_1^{i_1}||D_2^{\max_2-1}|}\sum_{c\in R_{D_1}^{-1}(c_p)} y^{i_1\max_2}(c),\end{aligned}$$

and

$$\lambda(d_2^{\max_2-1}) = \frac{1}{|D_1^{i_1}|}\sum_{c\in R_{D_1}^{-1}(c_p)} y^{i_1(\max_2-1)}(c') - \mu_-.$$

We want to show that:

$$\hat{y}^{i_1\max_2}(c) = \sum_{c'\in R_{D_2}^{-1}(c)} \hat{y}^{i_1(\max_2-1)}(c').$$

Now:

$$\begin{aligned}\sum_{c'\in R_{D_2}^{-1}(c)} \hat{y}^{i_1(\max_2-1)}(c') &= \sum_{c'\in R_{D_2}^{-1}(c)} (\mu_- + \lambda(d_2^{\max_2-1})) \\ &= \sum_{c'\in R_{D_2}^{-1}(c)} (\tfrac{1}{|D_1^{i_1}|}\sum_{c\in R_{D_1}^{-1}(c_p)} y^{i_1(\max_2-1)}(c')) \\ &= \frac{1}{|D_1^{i_1}|}\sum_{c\in R_{D_1}^{-1}(c_p)} y^{i_1\max_2}(c) \\ &= \mu_+ = \hat{y}^{i_1\max_2}(c).\square\end{aligned}$$

Example 4.7.1. We illustrate the latter case with an example from the foodmart data warehouse. We identify exceptional values in the context cube C= Year × Products with the main-effects ANOVA model

$$\hat{y}(\text{Year}, \text{Products}) = \mu + \lambda_1(\text{Year}) + \lambda_2(\text{Products}).$$

If we now explain an exceptional cell in the direction of the Time dimension, we compute reference objects in the context cube $R_T^{-1}(C)$ = Year.Quarter × Products with the specialised main-effects ANOVA model $\hat{y}(\text{Year.Quarter}_i, \text{Products}) = \mu + \lambda_1(\text{Year.Quarter}_i) + \lambda_2(\text{Products})$. Notice that this model complies with Theorem 4.7.1. Therefore, we obtain consistent reference objects i.e.:

$$\hat{y}(\text{Year}, \text{Products}) = \sum_{i=1}^{4} \hat{y}(\text{Year.Quarter}_i, \text{Products}).$$

Case 2) The specialization of a main-effects ANOVA model within an unfolded dimension $D_1^{i_1}$.

Suppose we have a parent cube $C = (D_1^{i_1})$, where $c \in C$, and a child cube $C' = R_{D_1}^{-1}(C)$, where $c' \in C'$, and the additive measure y with

$$y^{i_1}(c) = \sum_{c' \in R_{D_1}^{-1}(c)} y^{i_1-1}(c').$$

The expected value for $y^{i_1}(c)$ is computed by the additive ANOVA model (Equation (3.3))

$$\begin{aligned} \hat{y}^{i_1}(c) = &\ \mu_+ \\ = &\ \frac{1}{|D_1^{i_1}|} \sum_{c \in R_{D_1}^{-1}(c_p)} y^{i_1}(c), \end{aligned}$$

where $c_p \in R_{D_1}^{+1}(c)$, and the expected value for $y^{i_1-1}(c')$ is computed by the specialized ANOVA model

$$\begin{aligned} \hat{y}^{i_1-1}(c') = &\ \mu_- \\ = &\ \frac{1}{|D_1^{i_1}||D_1^{i_1-1}|} \sum_{c \in R_{D_1}^{-1}(c_p)} \sum_{c' \in R_{D_1}^{-1}(c)} y^{i_1-1}(c') \\ = &\ \frac{1}{|D_1^{i_1}||D_1^{i_1-1}|} \sum_{c \in R_{D_1}^{-1}(c_p)} y^{i_1}(c). \end{aligned}$$

We want to show that:

$$\hat{y}^{i_1}(c) = \sum_{c' \in R_{D_1}^{-1}(c)} \hat{y}^{i_1-1}(c').$$

Now:

$$\begin{aligned} \sum_{c' \in R_{D_1}^{-1}(c)} \hat{y}^{i_1-1}(c') = &\ \sum_{c' \in R_{D_1}^{-1}(c)} \left(\frac{1}{|D_1^{i_1}||D_1^{i_1-1}|} \sum_{c \in R_{D_1}^{-1}(c_p)} y^{i_1}(c)\right) \\ = &\ |D_1^{i_1-1}| \cdot \mu_- \\ = &\ \mu_+ = \hat{y}^{i_1}(c). \square \end{aligned}$$

Example 4.7.2. We illustrate this case with an example from the foodmart data warehouse, similar as in Example 4.7.1. However, in this case, we explain an exceptional cell in the Location dimension, we compute reference objects in the context cube $R_L^{-1}(C)$ = Year × Products × Country with the specialised main-effects ANOVA model $\hat{y}$(Year, Products, Country)= $\mu + \lambda_1$(Year) + λ_2(Products) + λ_3(Country). This model is the result of a drill-down operation on the location dimension from the level All-Countries to the level Country. Notice that this model complies with Theorem 4.7.1. Therefore, we obtain consistent reference objects given by

$$\hat{y}(\text{Year}, \text{Products}) = \sum_{k=1}^{20} \hat{y}(\text{Year}, \text{Products}, \text{Country}_k).$$

Remark 4.7.1. Furthermore, we consider a special case, which is the specialization of an ANOVA model within an unfolded dimension and with slices over the drilled-down data. This special case is only applicable for explanation of an exceptional value in a dimension, that is a balanced tree, where each parent has the same number of children. This is an additional property for Theorem 4.7.1. From a practical viewpoint only the Time dimension has this property in general, e.g. each year is composed out of 4 quarters, and each quarter is composed out of 3 months.

Suppose we have a parent cube $C = (D^{i_1})$, where $c \in C$, and a matrix sliced child cube $C' = S^{D_1^{i_1}.D_1^{i_1-1}=D_1^{i_1}.d_1^{i_1-1}}(R_{D_1}^{-1}(C)) = (D_1^{i_1}.d_1^{i_1-1})$, where $c' \in C$. The number of matrix sliced child cubes is $|D_1^{i_1-1}|$. Moreover, we have an additive measure y (see Definition 2.20) given by

$$y^{i_1}(c) = \sum_{c' \in R_{D_1}^{-1}(c)} y^{i_1-1}(c').$$

The expected value for $y^{i_1}(c)$ is computed by the additive ANOVA model (see Equation (3.3))

$$\begin{aligned} \hat{y}^{i_1}(c) = &\ \mu_+ \\ = &\ \frac{1}{|D_1^{i_1}|} \sum_{c \in R_{D_1}^{-1}(c_p)} y^{i_1}(c), \end{aligned}$$

where $c_p \in R_{D_1}^{+1}(c)$, and the expected value for $y^{i_1-1}(c')$ is computed by the specialized ANOVA model

$$\begin{aligned} \hat{y}^{i_1-1}(c') = &\ \mu_- \\ = &\ \frac{1}{|D_1^{i_1}.d_1^{i_1-1}|} \sum_{c \in R_{D_1}^{-1}(c_p)} \sum_{c' \in S^{D_1^{i_1}.D_1^{i_1-1}=D_1^{i_1}.d_1^{i_1-1}}(R_{D_1}^{-1}(c))} y^{i_1-1}(c') \\ = &\ \frac{1}{|D_1^{i_1}|} \sum_{c \in S^{D_1^{i_1}.D_1^{i_1-1}=D_1^{i_1}.d_1^{i_1-1}}(R_{D_1}^{-1}(c_p))} y^{i_1}(c). \end{aligned}$$

We want to show that:

$$\hat{y}^{i_1}(c) = \sum_{c' \in R_{D_1}^{-1}(c)} \hat{y}^{i_1-1}(c').$$

Now:

$$\begin{aligned}\sum_{c' \in R_{D_1}^{-1}(c)} \hat{y}^{i_1-1}(c') = & \sum_{c' \in R_{D_1}^{-1}(c)} (\frac{1}{|D_1^{i_1}|} \sum_{c \in S^{D_1^{i_1}.D_1^{i_1-1}=D_1^{i_1}.d_1^{i_1-1}}(R_{D_1}^{-1}(c_p))} y^{i_1}(c)) \\ = & \frac{1}{|D_1^{i_1}|} \sum_{c \in R_{D_1}^{-1}(c_p)} y^{i_1}(c) \\ = & \mu_+ = \hat{y}^{i_1}(c). \square\end{aligned}$$

An example of this special case is given in the case study described in Section 6.3.

Remark 4.7.2. Notice that Theorem 4.7.1 only holds for additive ANOVA models. If R is an ANOVA model with interaction terms or a contingency table model, the chains of reference objects will usually become inconsistent because of the presence of multiplicative terms in the equations, see, for example, Equation (3.15) or Equation (3.19). For the application of an ANOVA model with non-additive terms or a contingency table model in explanation, consistency needs to enforced. How this can be done is described in the following procedure:

1. Compute the expected values for all RHS elements in drill-down Equation (2.12) or business model Equation (2.22) with the statistical model under consideration.

2. Compute $y^{r;i_q}(c)'$ by using the same equation as for the actual values, applied on the expected values determined in the previous step.

Obviously, this procedure creates a bias term in the explanation, because $y^{r;i_q}(c)'$ $\neq$ $\hat{y}^{i_q}(c)$. As long as this bias is relatively small, this procedure will not have a significant effect on the composition of the explanation.

4.8 Related work

In this section, we discuss some related work on the topic of computerized explanation and diagnosis in the domain of business and management. There are many contributions on technical diagnosis and medical diagnosis, see Appendix A for a brief overview and Verkooijen (1993) for a comprehensive overview. In contrast, there are

only a limited number of publications related to the automatic generation of explanations based on business or financial models (Binbasioglu and Zychowicz 1998; Bouwman 1983; Daniels and Caron 2009; Courtney et al. 1987; Feelders 1993; Feelders and Daniels 2001; Hamscher 1992; Hamscher 1994; Kosy and Wise 1984) and multi-dimensional models (Caron and Daniels 2007; Cariou et al. 2008; Sarawagi 2001). In Table A.1, comparison is made between four applications domains of model-based diagnosis on a number of characteristics.

An early work related to our approach is the work of Bouwman (1983). Bouwman studied the diagnostic reasoning of financial analysts and compared this to the problem solving behaviour of novices. He also developed computer programs that can mimic the behaviour of human analysts including the shortcomings and mistakes that occurred in their analysis. Bouwman uses a qualitative model of reasoning compared to the quantitative model of reasoning used in this thesis.

Kosy and Wise (1984) and Kosy (1989) describe a general system for generating explanations in financial models, not directed specifically at diagnostic problem solving. In their method no strict separation is made between contributing and counteracting causes, which leads to counterintuitive results in some cases and it may cause the system to leave out significant causes from the explanation.

Courtney et al. (1987) and Mohammed et al. (1988) describe a DSS directed at managerial problem diagnosis. Functional relations that are allowed to sustain explanations are restricted to linear functions however. The restriction to linear relationships is not very realistic in a financial context. A clear distinction is made in their system between contributing and counteracting influences similar as described in RM_1.

Hamscher (1992) discusses the motivations and foundations of model-based reasoning and diagnosis in the financial domain, and surveys several existing AI programs for explanation in this domain. Moreover, Hamscher (1994) proposes a method and develops a prototype system that automatically constructs explanations for financial results based on a quantitative model. Their method gives information about the relative likelihoods of individual explanations, and is opposed to our method related to probabilistic reasoning.

Binbasioglu and Zychowicz (1998) present a diagnostic knowledge-based system

for analyzing the financial "health" of a company. An important difference is that they do not have an explanation methodology that gives the underlying causes for a symptom, instead they document the interactions among the financial domain objects.

In Feelders (1993) and Feelders and Daniels (2001), a formal framework is presented for explanation and diagnosis of business performance with both qualitative and quantitative information. The essence of this framework is discussed in Section 4.2. In this chapter, we extend this framework on several points. In Section 4.2, we introduce the consistency constraint and explain the interpretation of the influence measure. In Section 4.3, we extend the framework in order to deal with the problem of cancelling-out effects. In Sections 4.4 and 4.4.2 we discuss how the framework can be used in explanation in multi-dimensional databases. In Section 4.6, we describe, next to the concept of parsimonious causes, several new methods to reduce information overload in explanation. Lastly, in Section 4.7, we discuss for different types of normative model how the consistency constraint is satisfied.

In Sarawagi (2001), an explanation operator is presented for multi-dimensional data that lets the analyst generate summarized reasons for drops or increases observed at an aggregated level. This operator partly eliminates the need to manually drill-down for such reasons. Sarawagi developed an information theoretic formulation for expressing these reasons and designed a greedy and dynamic programming algorithm for explaining differences. The operator also reduces information overload by conveying only key reasons to the user, similar as applied in RM_4. However, the operator is not based on a causal model of explanation, as described in this thesis, resulting in problems with defining causes and finding clear parameters for their algorithms. Moreover, norm values in the approach of Sarawagi are not pre-computed by a statistical model but are typically historical norm values. The approach taken in Cariou et al. (2008) is closely related to Sarawagi's. The authors developed a method, based on statistical associations, to discover interesting dimensions to expand.

A recent group of related work is found in methods that couple data mining techniques with OLAP databases to support various forms of discovery-driven analysis. In Giacometti et al. (2008) and Giacometti et al. (2011), the authors present a system for recommending OLAP database queries to the analyst. This system is based on the harvesting of OLAP server's log data with collaborative filtering techniques.

In comparison, our method of explanation does not make use of such data and does not induce a model from the data. The authors in Hsu and Li (2011), use clustering methods and multi-dimensional scaling to determine similarity knowledge in OLAP databases. They define similarity knowledge as hidden rules, similar reports, or trends. The objective of similarity reduction (RM_5) described in Section 4.6.5 is quite similar. An in-depth comparison between the methods might be an interesting topic for future research.

4.9 Conclusion

In this chapter, we first summarized the most important elements of the theory on automated explanation in the domain of business and finance, in Section 4.2. Additionally, it was shown that the generation of valid explanations is only possible if certain constraints are satisfied. Important in the theory on explanation is the computation of the influence measure, which embodies a form of ceteris paribus reasoning. Here it was shown that the interpretation of this measure is dependent on the functional form of the function considered for explanation generation. Elements from the theory are used in the development of three computerized methods for the explanation of an exceptional cell value $\partial y(c) = q$ in a cube C. The explanation methods discussed are look-ahead explanation, top-down explanation, and greedy explanation. Each method is used in a specific case.

In Section 4.3, an explanation method is discussed that can be used in the context of a business model. The existing explanation methodology is extended with a procedure to deal with cancelling-out effects in data sets. In this procedure hidden causes are made visible by the use of function substitution. A multi-level look-ahead algorithm, that applies function substitution, is proposed that visualizes hidden causes.

In Section 4.4, explanation generation in a system of solely drill-down equations is discussed. Here a general top-down explanation method and a specific greedy explanation method are developed. The top-down explanation method uses the method of maximal explanation in a system of drill-down equations and shows that the theory on automated explanation can indeed be applied in the OLAP context. The greedy explanation method considers systems of equations that are composed out of purely

additive drill-down equations, corresponding to application of the SUM() aggregation function. In this method the transitivity property of the influence measure is used, which simplifies explanation generation in such systems of equations. In the method the concept of an aggregated table is applied, that might contain the influence values for all elements in the exceptional cell's downset. In this table the causes for a symptom are determined greedily in (parts of) the symptom's downset, where first the largest is determined, then the second largest cause, and so on. Finally, greedy explanation in a system of average and maximum/minimum drill-down equations is treated and expressions for the influence measure are developed. These systems of drill-down do not exhibit the property of transitivity. However, an important result is that influence values from elements in the exceptional cell's downset can be sorted and a greedy method can be used.

In Section 4.5, we discussed explanation generation in a hybrid system of equations, that contains both drill-down and business model equations. In the OLAP context, computerized explanation is supported by these two internal structures. Therefore, we developed a compound explanation method for finding significant causes in these structures, based on the algorithms described in this chapter.

In addition, the explanation methods and algorithms use the concept of an explanation tree, in which the main causes for a symptom are presented to the analyst. To prevent an information overload, several reduction techniques are proposed in Section 4.6 to prune the tree. RM_1 constructs parsimonious sets of causes. RM_2 identifies specific explanations. RM_3 reduces the number of elements in the analysis path based on the application of a reduction heuristic. RM_4 is used in combination with greedy explanation to produce a tree with the n largest causes. Finally, RM_5 reduces sets of explanations to a *generic explanation* that hold for a number of exceptional cells.

To ensure the correct working of the explanation methods the consistency constraint has to be satisfied. Reference values are consistent if they satisfy the same equation as is given for the actual values. In Section 4.7, we discuss for each type of normative model R under what conditions the consistency constraint is satisfied. In particular, we describe a special class of additive ANOVA models that produce consistent reference values, as opposed to the general class of statistical models that do not produce such values.

Chapter 5

Sensitivity analysis

5.1 Introduction

In this chapter, we describe how sensitivity analysis can be implemented in a multi-dimensional database. Sensitivity analysis in multi-dimensional databases is related to the notion of comparative statics in economics. Where the central issue is to determine how changes in independent variables affect dependent variables in an economic model. Comparative statics is defined as the comparison of two different equilibrium states solutions, before and after a change in one of the independent variables, keeping the other variables unchanged (Samuelson 1941). It is one of the primary analytical methods used in economics, where it is commonly used, for example, in the study of changes in supply and demand when analyzing a market. Instead of repeating the phrase "keeping the other variables unchanged", economists use the more compact Latin equivalent *ceteris paribus* (c.p.). The underlying model for comparative statics is a set of equatiuons that define the vector of dependent variables $y_1, y_2, \ldots, y_m$ as functions of the vector of independent variables $x_1, x_2, \ldots, x_n$, i.e.

$$\mathbf{y} = f_l(\mathbf{x}), l = 1, 2, \ldots, m. \tag{5.1}$$

This corresponds to a system of business model equations (Equation (2.22)), where the function f might be non-linear, or a system of drill-down equations (Equation (2.11)), where the function f is linear. In the latter situation we use the terms non-base variables and base variables, as defined in Section 2.3, for dependent and independent variables, respectively. To implement sensitivity analysis in OLAP, we

define a new cube operator that supports the analyst in answering typical managerial what-if questions, while navigating the cube. We distinguish between two types of what-if questions:

- Questions related to a system of drill-down equations. For example, "How is the profit in the year 2010 affected when the profit for a certain product is changed with one percent in the first quarter in The Netherlands, c.p.?"
- Questions related to a system of business model equations. For example, "How is the profit in the year 2010 for a certain product affected when its unit price is changed with one additional unit in the sales model, c.p.?"

This chapter is structured as follows. In Section 5.2 we discuss sensitivity analysis in systems that consist of purely drill-down equations. In Section 5.3 we elaborate on sensitivity analysis in systems that consist of purely business model equations and mixed systems of equations. In Section 5.4 we discuss related work. Finally, in Section 5.5 we draw some conclusions.

5.2 Sensitivity analysis in a system of drill-down equations

In this section we investigate the influence of a change in a measure value of a cell in any cube, on a higher level value of the same measure in the aggregation lattice. Or in formal notation, what is the effect of changing $y(c')$ to $y(c') + \delta$ on a dependent variable $y(c)$ in the upset of c'. To solve this consider the lattice L' with top cube $C_p = [i_1, i_2, \ldots, i_n]$ and base cube $C_q = [j_1, j_2, \ldots, j_n]$. Notice that L' is a sublattice of L and $L' = \{\downarrow c\} \cap \{\uparrow c'\}$. The values of the measure y in the cube C_q are denoted by $x(c'_i)$, and are called the base variables where $i = 1, 2, \ldots, |C_q|$, and the values of the measure y in $\{\uparrow C_q\}$ are denoted by $y(c)$, and are called the non-base variables. We distinguish between the original values of a measure without change $x^r(C_q)$ and $y^r(C_p)$, and the values of the changed measure: $x^a(C_q)$ and $y^a(C_p)$, where $x^a(C_q) = x^r(C_q)$ except for one cell c'_i in the cube C_q, for which $x^a(c'_i) - x^r(c'_i) = \delta$.

The following theorem shows how the values of y change in the lattice L'.

Theorem 5.2.1. There is an unique additive drill-down measure $y^a(c)$ defined on all cube cells in the sublattice L' such that:

$$y^a(c) = y^r(c) + \beta(c) \cdot (x^a(c'_i) - x^r(c'_i)), \tag{5.2}$$

where:

$$\beta(c) = 1 \text{ if } c \in \{\uparrow c'_i\}, \text{ and}$$
$$\beta(c) = 0 \text{ if } c \notin \{\uparrow c'_i\}.$$

Proof. To show that $y^a(c)$ is additive it is sufficient to show that $\beta(c) \cdot (x^a(c'_i) - x^r(c'_i))$ is additive, because the sum of additive measures is also additive and $y^r(c)$ is additive by the consistency assumption. Hence, we must show that:

$$\beta(c) = \sum_q \beta(R_q^{-1}(c)), \tag{5.3}$$

where R_q^{-1} is the drill-down operation defined on a cell c in the lattice L. Now there are two cases:

1. $c \in \{\uparrow c'_i\}$, i.e. c is an ancestor of c'_i. In that case c'_i is also a descendant of *one* of the cells in $R_q^{-1}(c)$, $c'_i \in \{\downarrow R_q^{-1}(c)\}$, which is a child of c in dimension q. This property does not depend on dimension q. So both sides of Equation (5.3) are equal to 1.

2. $c \notin \{\uparrow c'_i\}$, i.e. c is not an ancestor of c'_i. In that case, c'_i is also not a descendant of *one* of the children of c. Hence, both sides of Equation (5.3) are zero. □

Notice that the drill-down measure $y^a(c)$ is unique. This follows from the general proposition that every additive measure with given values on the base cube is unique (Equation 2.14). This follows immediately from Theorem 2.3.1 (see Remark 2.3.1) and the fact that L' is a lattice of cubes.

In the case that $c \in \{\uparrow c'_i\}$, we can rewrite Equation (5.2) as follows

$$y^a(c) = y^r(c) + \inf(y^a(c'), y^a(c)). \tag{5.4}$$

If $y(c)$ is an additive drill-down measure then we use Equation (4.15) for the computation of $\inf(y^a(c'_i), y^a(c))$ in Equation (5.4) and if the variable $x^r(c')$ is changed with

δ in sensitivity analysis then $y^a(c)$ is computed as $y^a(c) = y^r(c) + (x^a(c') - x^r(c'))$. This result follows immediately from Theorem 5.2.1.

Moreover, in the case that $y^r(c)$ is an average drill-down measure we use Equation (4.18) for the computation of $\inf(y^a(c'_i), y^a(c))$ in Equation (5.4) and if the variable $x^r(c')$ is changed with δ in sensitivity analysis then $y^a(c)$ is computed as $y^a(c) = y^r(c) + \frac{1}{|C_q|}(x^a(c') - x^r(c'))$, where C_q is the context cube under consideration. This result is not proven here but the proof is similar to the proof of Theorem 5.2.1, with the difference that the RHS of the drill-down equation is divided by the number of cells in the context cube.

Example 5.2.1. Here we present a numeric example of a what-if analysis in the cube C = Store $\times$ Products for the measure sales, aggregated by the average function. The data of the cube is depicted in Table 5.1. We want to analyse a change δ in the cell (A, P_1) on its upset $\{\uparrow (A, P_1)\}$. The reference value of the cell is given by $\text{sales}^r(A, P_1) = 1$ and the actual value is given by $\text{sales}^a(A, P_1) = 1 + \delta$. By applying Equation (5.4) we compute the effect of this change on $\{\uparrow (A, P_1)\}$; these effects are given by

$$\begin{aligned}
\text{sales}^a(\text{All}, P_1) &= \text{sales}^r(\text{All}, P_1) + \tfrac{1}{3}\delta \text{ where } |R^{+1}_{\text{Stores}}(C)| = 3,\\
\text{sales}^a(A, \text{All}) &= \text{sales}^r(A, \text{All}) + \tfrac{1}{4}\delta \text{ where } |R^{+1}_{\text{Products}}(C)| = 4,\\
\text{sales}^a(\text{All}, \text{All}) &= \text{sales}^r(\text{All}, \text{All}) + \tfrac{1}{12}\delta \text{ where } |C| = 12.
\end{aligned}$$

Table 5.1: Sensitivity analysis in the example cube Store $\times$ Products for the average drill-down measure sales. Here the value of the cell (A, P_1) is changed with δ and this change is propagated in the cell's upset.

AVG(sales)		Stores			
		A	B	C	All
Products	P_1	$1 + \delta$	2	3	$2 + \frac{1}{3}\delta$
	P_2	4	5	6	5
	P_3	7	8	9	8
	P_4	10	11	12	11
	All	$5.5 + \frac{1}{4}\delta$	6.5	7.5	$6.5 + \frac{1}{12}\delta$

Remark 5.2.1. The subsystem of drill-down equations that corresponds with $\{\uparrow c'\}$ has an unique solution, after a change in $y(c')$ with some δ, as a result of Theorem 5.2.1. However, the complete system of equations becomes inconsistent because Equation 2.12 does not hold in that case:

$$y^{\max_1 \max_2 \ldots \max_n}(c) + \delta(c') \neq \sum_{c_n \in R_n^{-\max_n} \circ \ldots \circ R_2^{-\max_2} \circ R_1^{-\max_1}(c)} y^{00\ldots0}(c_n).$$

In other words, when the change in what-if analysis is not induced by a variable in the base cube, but by a (non-base) variable on some intermediate level in the lattice L, the complete system of equations will become inconsistent. For analysis restricted to $\{\uparrow c'\}$ this does not matter, however analysis in the complete system is obviously not useful anymore. The inconsistencies in the complete system of drill-down equations, can be corrected by a straightforward procedure, that repairs the OLAP database (Caron and Daniels 2008).

5.3 Sensitivity analysis in a system of business equations

In this section we discuss managerial what-if questions related to a system of business model equations and a mixed system of drill-down and business model equations. Multiple related measures in the business model and associated dimensions, result in a mixed, often non-linear, system of equations.

Example 5.3.1. For example, consider Table 5.2 with the equations of Example 2.1.1. The equations in Table 5.2 are isolated from a larger system of equations, depicted

Table 5.2: Subsystem of business model and drill-down equations derived from a multi-dimensional financial database.

1. Rev.(2005) = Rev.(2005.Q1) + Rev.(2005.Q2) + Rev.(2005.Q3) + Rev.(2005.Q4)
2. Rev.(2005) = Vol.(2005) × Unit Pr.(2005)
3. Rev.(2005.Q2) = Vol.(2005.Q2) × Unit Pr.(2005.Q2)
4. Vol.(2005) = Vol.(2005.Q1) + Vol.(2005.Q2) + Vol.(2005.Q3) + Vol.(2005.Q4)
5. Unit Pr.(2005) = ((Vol.(*.Q1) × Unit Pr.(*.Q1)) + (Vol.(*.Q2) × Unit Pr.(*.Q2)) + (Vol.(*.Q3) × Unit Pr.(*.Q3)) + (Vol.(*.Q4) × Unit Pr.(*.Q4))) / Unit Pr.(*)

in Figure D.1 in Appendix D, Section D.2. In shorthand notation

$$\begin{cases} -y_1 + x_1 + y_2 + x_2 + x_3 = 0 \\ -y_1 + y_3 \times y_4 = 0 \\ -y_2 + x_4 \times x_5 = 0 \\ -y_3 + x_6 + x_4 + x_7 + x_8 = 0 \\ -y_4 + ((x_6 \times x_9) + (x_4 \times x_5) + (x_7 \times x_{10}) + (x_8 \times x_{11}))/y_3 = 0, \end{cases} \tag{5.5}$$

where y_i with $i = 1, 2, 3, 4$ are the dependent variables and x_i with $i = 1, 2, \ldots, 11$ are the independent variables. The system of equations in (5.5) are represented as a business model graph (see Section 2.3.2) in Figure 5.1. In this system we want to

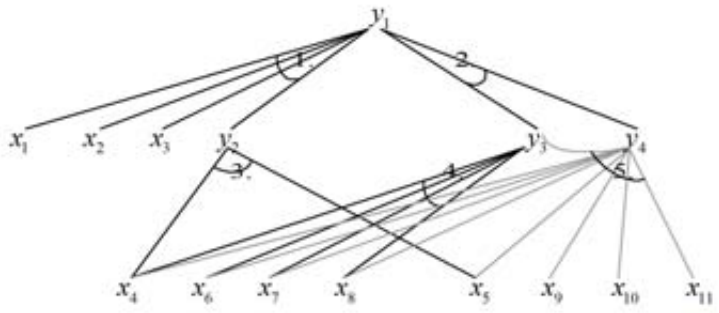


Figure 5.1: Graph representation of the implicit system of equations.

change an independent variable x_i, e.g. x_4 (= Volume(2005.Q2)) and/or x_5 (= Unit Price(2005.Q2)), and study the impact on its upset, in particular, the dependent root variable y_1 (= Revenues(2005)). Notice that (5.5) is overdetermined, because we have 4 independent variables and 5 equations.

In general, for a mixed system of equations

- the equations are linear and non-linear, and
- the system of equations is overdetermined.

A system of solely drill-down equations is also overdetermined in the case of multiple dimensions, as shown in Appendix D. Equation (5.5) can be written as

$$f_l(\mathbf{y}, \mathbf{x}) = \mathbf{0}. \tag{5.6}$$

The linearization of (5.6) in a neighborhood of a solution (y_0, x_0) reads:

$$A_1\mathbf{y} + A_2\mathbf{x} = \mathbf{0}. \tag{5.7}$$

The matrix A_1 is the $l \times m$ coefficient submatrix for dependent variables and A_2 is the $l \times n$ coefficient submatrix for independent variables. Here the matrix of the first derivatives of $\mathbf{f}$ with respect to $\mathbf{y}$ is represented by $A_1 = \mathbf{D_y f}(\mathbf{y}, \mathbf{x})$ and the matrix of first derivatives of $\mathbf{f}$ with respect to $\mathbf{x}$ is represented by $A_2 = \mathbf{D_x f}(\mathbf{y}, \mathbf{x})$.

With (5.7) we can examine the impact of a change in one or more independent variables c.p., given by $\Delta\mathbf{x}$, on the dependent variables, given by $\Delta\mathbf{y}$, where equation (5.6) has to be satisfied. In the next section, we investigate the conditions for consistency and solvability of (5.7), which is a necessary condition for solvability of (5.6).

5.3.1 Conditions for solvability

A necessary condition for solvability in a system of linear equations is the *rank criterium*. A system of linear equations (5.7), of $A_1\mathbf{y} + A_2\mathbf{x} = \mathbf{0}$, is solvable if and only if $\text{rank}(A_1| - A_2\mathbf{x}) = \text{rank}(A_1)$. The proof of this theorem is, for example, given in Schott (1997). In words, the rank criterium says that the vector $-A_2\mathbf{x}$ must be in the column space (range) of A_1 for the system to be solvable.

To investigate the solvability of (5.6), we assume that

$$(\mathbf{y}_0, \mathbf{x}_0) = (y_1^0, y_2^0, \ldots, y_m^0, x_1^0, x_2^0, \ldots, x_n^0)$$

is a solution of (5.6). We substitute this solution in the derivative matrices A_1 and A_2 to obtain the linearized matrix $[A_1 A_2]$ at the solution $(\mathbf{y}_0, \mathbf{x}_0)$. The linearized system of equations $A_1\Delta\mathbf{y} + A_2\Delta\mathbf{x} = \mathbf{0}$ is solvable if and only if $\text{rank}(A_1) = \text{rank}(A_1| - A_2\Delta\mathbf{x})$. Similarly, the linearized system of equations is solvable for an independent variable ∂x_i, if and only if, $\text{rank}(A_1) = \text{rank}(A_1|\text{column } x_i \text{ from } A_2)$. A column vector x_i of

the submatrix A_2 is represented by $a_2(i)$. Accordingly, the rank criterium can be used to determine whether an independent variable ∂x_i qualifies for what-if analysis in a system of business model equations. However, as we shall see in the next section, this criterium is a necessary but not sufficient condition for the solvability of a non-linear system of equations. (5.7), When the submatrix A_1 is nonsingular then the solution of $A_1\Delta\mathbf{y} + A_2\Delta\mathbf{x} = \mathbf{0}$ is unique and given by

$$\Delta\mathbf{y} = -A_1^{-1}A_2\Delta\mathbf{x}.$$

Notice that the rank criterium is a necessary but not sufficient condition for the solvability of a *non-linear system of equations*, as shown in the following example.

Example 5.3.2. Consider the system of equations

$$\begin{cases} -y + x^3 + x = 0 \\ -y + x^2 + x = 0. \end{cases} \tag{5.8}$$

Observe that the point $(0, 0)$ is a solution to this system of equations. Define $\mathbf{f}(y, x) = (-y + x^3 + x, -y + x^2 + x)$. The Jacobian of $\mathbf{f}$ is

$$A = \begin{pmatrix} -1 & 3x^2 + 1 \\ -1 & 2x + 1 \end{pmatrix}.$$

For $\mathbf{x} = \mathbf{0}$:

$$A = [A_1 \; A_2] = \left(\begin{array}{c|c} -1 & 1 \\ -1 & 1 \end{array}\right).$$

The system satisfies the rank criterium, because $\text{rank}(A_1|A_2) = \text{rank}(A_1) = 1$. However, the implicit function theorem cannot be applied because the submatrix A_1 is non-square. However, $A_1\Delta\mathbf{y} + A_2\Delta\mathbf{x} = \mathbf{0}$ is solvable for all $\Delta\mathbf{x}$ but the non-linear system represented by (5.8) is not solvable for $x \neq 0$.

Practically, this means that in such models the number of equations must be equal to the number of dependent variables to produce a square submatrix A_1 ($l = m$). For example, the business model in Table 1.1 (see Chapter 1) satisfies this condition, because it has 5 business equations and 5 dependent variables.

In the case that what-if analysis is performed in a mixed system of equations, the number of equations is larger than the number of dependent variables, thus $l >$

m, because it contains an OLAP subsystem. In such systems the implicit function theorem cannot be applied because the submatrix A_1 is non-square. However, in some cases it is still possible to derive unique solutions if certain independent variables are changed. This is shown in the example described in Section 5.3.2.

Now suppose that we are given an overdetermined system of equations as in (5.6) and a solution $(\mathbf{y}_0, \mathbf{x}_0)$ to this system such that all the equations are satisfied. The first derivatives of the equations can be written in matrix form as in (5.7). If the rank criterium for consistency holds for a certain independent variable x_i, considered for what-if analysis, then the solution $\mathbf{f}(\mathbf{y}_0, \mathbf{x}_0) = \mathbf{0}$ is filled in Equation (5.7). Subsequently,

$$\alpha_1 \cdot \text{eq. } 1 + \alpha_2 \cdot \text{eq. } 2 + \ldots + \alpha_l \cdot \text{eq. } l = 0, \tag{5.9}$$

holds if all the α_i's exist. If the α_i's exist we remove $(l-m)$ dependent equations from the system of equations and derive a $(m \times m)$ submatrix A_1. If the remaining system of equations in A_1 is nonsingular the implicit function theorem can be applied and the α_i's determined. In that case the removed equations are satisfied too, because Equation (5.9) holds and the general solution for x_i can be determined.

5.3.2 What-if analysis example

In this example we want to change an independent variable x_i and study the impact on elements in its upset. The Jacobian of the system of equations in (5.5) is

$$A = [A_1\ A_2] =$$

$$\left(\begin{array}{cccc|cccccccccc}
-1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & y_4 & y_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & x_5 & x_4 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 &)^* & -1 & 0 & 0 & 0 & \frac{x_5}{y_3} & \frac{x_4}{y_3} & \frac{x_9}{y_3} & \frac{x_{10}}{y_3} & \frac{x_{11}}{y_3} & \frac{x_6}{y_3} & \frac{x_7}{y_3} & \frac{x_8}{y_3}
\end{array}\right).$$

$$)^* = -\frac{x_6x_9 + x_4x_5 + x_7x_{10} + x_8x_{11}}{(y_3)^2}$$

Observe that the vector

$$(\mathbf{y_0}\ \mathbf{x_0}) = (48\ 16\ 15\ 3.2|13\ 12\ 7\ 4\ 4\ 6\ 3\ 2\ 2.75\ 3\ 3.25),$$

is a solution to the system of equations. The Jacobian at (x_0, y_0) is

$$A_0 = [A_1\ A_2] = \left(\begin{array}{cccc|cccccccccc} -1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 3.2 & 15 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 4 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -\frac{48}{225} & -1 & 0 & 0 & 0 & \frac{4}{15} & \frac{4}{15} & \frac{2.75}{15} & \frac{3}{15} & \frac{3.25}{15} & \frac{6}{15} & \frac{3}{15} & \frac{2}{15} \end{array}\right).$$

The rank criterium for solvability in this system is satisfied for the variables x_4 (= Volume(2005.Q2)) and x_5 (= Unit Price(2005.Q2)): $\text{rank}(A_1|a_{2(4)}) = \text{rank}(A_1) = 4$ and $\text{rank}(A_1|a_{2(5)}) = \text{rank}(A_1) = 4$. It can easily be verified that the rank criterium is not satisfied for the other independent variables. For example, for variable x_1 it can be concluded that $\text{rank}(A_1|a_{2(1)}) > \text{rank}(A_1)$. Therefore, the only candidate independent variables for what-if analysis in this example are x_4 and x_5.

As we saw, the rank criterium is a necessary but not sufficient condition for solvability. We cannot apply the implicit function theorem to verify solvability here, because the submatrix A_1 is non-square (5×4). But in this case we may eliminate one of the equations because we can find α_i such that:

$$\alpha_1 \cdot \text{ eq. } 1 + \alpha_2 \cdot \text{ eq. } 2 + \alpha_3 \cdot \text{ eq. } 3 + \alpha_4 \cdot \text{ eq. } 4 + \alpha_5 \cdot \text{ eq. } 5 = 0. \tag{5.10}$$

These α_i's are given by

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ -1 \\ 0 \\ y_3 \end{pmatrix}.$$

Now we proceed as follows. In the system of equations in (5.5) all independent variables are replaced by the solution $(\mathbf{y}_0, \mathbf{x}_0)$ except the independent variables x_4 and x_5, that are under consideration for what-if analysis. From the original system of equations, one dependent equation is removed and we derive a reduced system of equations, where the matrix A_1 is square. Removing eq. 2 yields

$$f(y_1, y_2, y_3, y_4, x_4, x_5) = \begin{cases} -y_1 + 32 + y_2 = 0 \\ -y_2 + x_4 x_5 = 0 \\ -y_3 + 11 + x_4 = 0 \\ -y_4 + (32 + x_4 x_5)/y_3 = 0. \end{cases} \tag{5.11}$$

$(\mathbf{y}_0, \mathbf{x}_0) = (48, 16, 15, 3.2, 4, 4)$ is a solution of (5.11). The 4×4 derivative submatrix A_1 of $\mathbf{f}$ with respect to $\mathbf{y}$ in $(48, 16, 15, 3.2, 4, 4)$ is

$$\mathbf{D_y f}(48, 16, 15, 3.2, 4, 4) = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & -\frac{48}{225} & -1 \end{pmatrix} = A_1.$$

It can easily be verified that

$$A_1 A_1^{-1} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & -\frac{48}{225} & -1 \end{pmatrix} \begin{pmatrix} -1 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & \frac{48}{225} & -1 \end{pmatrix} = I_4.$$

By the implicit function theorem we can find continuous differentiable functions $\varphi_i(x_4, x_5) : \mathcal{B} \to \mathbb{R}$, where $\mathcal{B} = \mathcal{B}_r(48, 16, 15, 3.2, 4, 4)$, such that

$$\begin{cases} y_1 = \varphi_1(x_4, x_5) \\ y_2 = \varphi_2(x_4, x_5) \\ y_3 = \varphi_3(x_4, x_5) \\ y_4 = \varphi_4(x_4, x_5), \end{cases}$$

is a solution of the system of equations (5.11). Moreover, also the removed equation $-y_1 + y_3 y_4 = 0$ (eq. 2) is satisfied because of (5.10). Computation gives:

$$\begin{cases} y_1 = 32 + x_4 x_5 \\ y_1 = (11 + x_4)(\frac{32 + x_4 x_5}{11 + x_4}) = 32 + x_4 x_5 \\ y_2 = x_4 x_5 \\ y_3 = 11 + x_4 \\ y_4 = \frac{32 + x_4 x_5}{11 + x_4}. \end{cases}$$

5.3.3 Alternative approach

In this section we propose an alternative approach for what-if analysis in a mixed system of equations as in (5.1). Suppose we have a system of equations as represented in (4.7) and (4.8). The indirect influence of the variable z_j on the root variable ∂y in M is defined as:

$$\inf(z_j, y) = y' - y^r, \tag{5.12}$$

where y' is derived by means of *value propagation* of z_j^a in $\{\uparrow z_j\}$ in the system of equations where all other variables are evaluated at their reference values. In the

value propagation process $x_i' = g_i(z_1^r, \ldots, z_j^a, \ldots, z_m^r)$ and $y' = f(x_1, \ldots, x_i', \ldots, x_n)$ are determined. In the case that the functions f and the g_i's in the system of equations are all smooth and the difference ∂y is small, the influence of a variable z_j on the root y can be approximated by:

$$\inf(z_j, y) = \left(\frac{\partial y}{\partial z_j}\right)_r \times \Delta z_j, \tag{5.13}$$

where $\frac{\partial y}{\partial z_j}$ is computed by applying the *chain rule* for partial differentiation, and is given by $\frac{\partial y}{\partial z_j} = \frac{\partial y}{\partial x_i}\frac{\partial x_i}{\partial z_j}$.

A necessary condition for solvability is

$$\inf(z_j, y)_{\text{path A}} = \inf(z_j, y)_{\text{path B}}, \tag{5.14}$$

where analysis path A and analysis path B are paths in the upset of the variable $\{\uparrow z_i\}$, and $y \in \{\uparrow z_i\}$. In the alternative approach this condition is used to single out systems of equations from sensitivity analysis that are not solvable. Besides if Equation (5.14) holds *for all paths* from z_j to y, i.e. it gives the same solution, then the system is uniquely solvable.

In the remainder of this section we present two typical examples, taken from the multi-dimensional sales database. In the first example the system is unsolvable and the second the system is uniquely solvable.

Example 5.3.3. In this example we illustrate what-if analysis in a mixed, non-linear, system of equations derived from a multi-dimensional sales database. Consider the following (partial) system of business model and drill-down equations in Table 5.3, derived from Figure D.2 in Appendix D, Section D.2.

Table 5.3: Partial system of sales model and drill-down equations.

1. Rev.(2005)= Rev.(2005.Q1) + Rev.(2005.Q2) + Rev.(2005.Q3) + Rev.(2005.Q4)
2. Rev.(2005) = Vol.(2005) × Unit Price(2005)
3. Rev.(2005.Q2) = Vol.(2005.Q2) × Unit Price(2005.Q2)
4. Vol.(2005) = Vol.(2005.Q1) + Vol.(2005.Q2) + Vol.(2005.Q3) + Vol.(2005.Q4)

In notation:

$$\begin{cases} y_1 = x_1 + y_2 + x_2 + x_3 \\ y_1 = y_3 \times x_4 \\ y_2 = x_5 \times x_6 \\ y_3 = x_7 + x_5 + x_8 + x_9. \end{cases} \tag{5.15}$$

In Figure 5.2, this system of equations is represented in a graph. A solution is

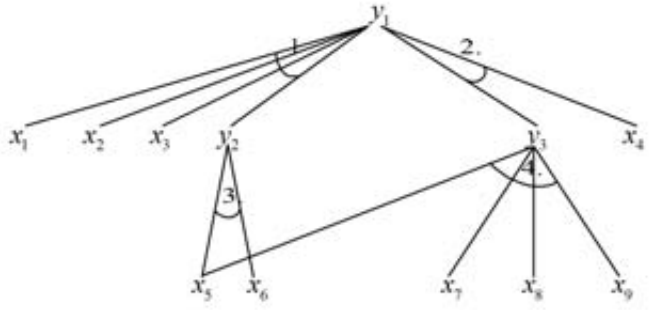


Figure 5.2: Graph of the system of non-linear equations.

$$(\mathbf{y}_0, \mathbf{x}_0) = (48\ 16\ 15|13\ 12\ 7\ 3.2\ 4\ 4\ 6\ 3\ 2).$$

We want to analyse the effect of a change in the variable x_5 on the root variable y_1 using Equation (5.13). There are two paths from x_5 to y_1, path A via the variable y_2 with equations 1 and 3, and path B via the variable y_3 with equations 2 and 4. We compute,

$$\begin{aligned}
\inf(x_5, y_1)_{\text{path A}} &= \tfrac{\partial y_1}{\partial x_5} \times \Delta x_5 \\
&= \tfrac{\partial y_1}{\partial y_2}\tfrac{\partial y_2}{\partial x_5} \times \Delta x_5 \\
&= 1 \cdot x_6 \times \Delta x_5 \\
&= x_6 \times \Delta x_5
\end{aligned}$$

and

$$\begin{aligned}
\inf(x_5, y_1)_{\text{path B}} &= \tfrac{\partial y_1}{\partial x_5} \times \Delta x_5 \\
&= \tfrac{\partial y_1}{\partial y_3}\tfrac{\partial y_3}{\partial x_5} \times \Delta x_5 \\
&= x_4 \cdot 1 \times \Delta x_5 \\
&= x_4 \times \Delta x_5.
\end{aligned}$$

From this we can conclude that $\inf(x_5, y_1)_{\text{path A}} \neq \inf(x_5, y_1)_{\text{path B}}$ and therefore Equation (5.14) does not hold. In other words, changing the variable x_5 is not allowed, because this will make the system insolvable.

Example 5.3.4. Consider again (5.5). The graph of the system of equations is given in Figure 5.1. In the system of equations we want to analyse the impact of a change in the variable x_4 on the root variable y_1 with Equation (5.13). There are two paths from x_4 to y_1, path A via the variable y_2 with equations 1 and 3 denoted by $y_1 = f_1(x_1, f_3(x_4, x_5), x_2, x_3)$, and path B via the variables y_3 and y_4 with equations 2, 4, and 5, denoted by $y_1 = f_2(f_4(x_4, x_6, x_7, x_8), f_5(x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}), f_4(x_4, x_6, x_7, x_8))$. First we compute,

$$\begin{aligned}
\inf(x_4, y_1)_{\text{path A}} &= \tfrac{\partial y_1}{\partial x_4} \times \Delta x_4 \\
&= \tfrac{\partial y_1}{\partial y_2}\tfrac{\partial y_2}{\partial x_4} \times \Delta x_4 \\
&= 1 \cdot x_5 \times \Delta x_4 \\
&= x_5 \times \Delta x_4
\end{aligned}$$

and then we compute

$$\begin{aligned}
\inf(x_4, y_1)_{\text{path B}} &= \tfrac{\partial y_1}{\partial x_4} \times \Delta x_4 \\
&= (\tfrac{\partial y_1}{\partial y_3}\tfrac{\partial y_3}{\partial x_4} + \tfrac{\partial y_1}{\partial y_4}\tfrac{\partial y_4}{\partial x_4} + \tfrac{\partial y_1}{\partial y_4}\tfrac{\partial y_4}{\partial y_3}\tfrac{\partial y_3}{\partial x_4}) \times \Delta x_4 \\
&= (1 \cdot y_4 + y_3 \cdot \tfrac{x_5}{y_3} + 1 \cdot y_3 \cdot -(\tfrac{x_6 x_9 + x_4 x_5 + x_7 x_{10} + x_8 x_{11}}{(y_3)^2})) \times \Delta x_4 \\
&= (y_4 + x_5 - y_4) \times \Delta x_4 \\
&= x_5 \times \Delta x_4.
\end{aligned}$$

From this we can conclude that $\inf(x_4, y_1)_{\text{path A}} = \inf(x_4, y_1)_{\text{path B}}$ and therefore Equation (5.14) holds. In addition, Equation (5.14) holds for all paths from x_4 to y_1, therefore the system remains solvable if x_4 is changed.

5.4 Related work

The variables, parameter values, and assumptions of any business or economic model are subject to change. Sensitivity analysis, generally defined, is the investigation of these potential changes and their impacts on conclusions to be drawn from the model (e.g. Baird (1990)). There are many possible applications of sensitivity analysis, described here within the categories of decision support, communication, increased understanding or quantification of the system, and model development (Pannell 1997). There is a very large literature on procedures and techniques for sensitivity analysis (Clemson et al. 1995). Two general classes of techniques for sensitivity analysis

are the *implicit function theorem* (Currier 2000; Heckman 2000) and *monotone comparative statics* (Milgrom and Shannon 1994). These are methods for characterizing whether an increase in a parameter causes the dependent variable to increase or decrease. Historically the implicit function theorem was used for this purpose and the implicit function theorem not only tells you whether the dependent variable increases or decreases but also the magnitude of change. In contrast, monotone comparative statics tells you only "up" or "down", i.e., it gives an ordinal rather than cardinal answer. In our research, we focused solely on quantitative what-if analysis within the multi-dimensional database.

To the best of our knowledge, Balmin et al. (2000) and Lakshmanan et al. (2007) are the only published research works that address sensitivity analysis in OLAP databases in a significant way. In Balmin et al. (2000), the authors have developed the SESAME system for the processing of hypothetical queries. For this system query algebra operators are proposed that are suitable for spreadsheet-style what-if computations. In the system hypothetical queries are modeled as a list of hypothetical modifications on the data in the fact table. A shortcoming of their approach is that it lacks a good mathematical underpinning, to decide whether a certain change is allowed in the model or not, as opposed to our approach. In Lakshmanan et al. (2007), a different perspective is taken on what-if analysis. They focus on what-if analysis related to changes in dimensions and their hierarchical structure. However, our focus is on data-driven what-if scenarios, as opposed to structural ones.

In many OLAP software products, sensitivity analysis is not possible at the moment. If one wants to do sensitivity analysis in these products one has to copy the data to a reporting environment, for example MS Excel, to compute manually the impact of changes in certain cells of the data cube. An exception is the software product Clickview (Cliqview Corporation 2010), where a fixed change in a base variable can be induced in a system of additive drill-down measures, to determine its impact on non-base variables. The difference with our approach is that we can induce variable changes in systems of additive and average drill-down measures and under certain conditions in non-linear systems of business equations.

5.5 Conclusion

In this chapter, we stated the theoretical underpinnings under which sensitivity analysis is allowed in multi-dimensional databases. We also discussed some theoretical issues and procedures related to sensitivity analysis in OLAP databases.

For sensitivity analysis in systems of additive drill-down measures we proved Theorem 5.2.1, and showed that there is an unique additive drill-down measure $y^a(c)$ defined on all cubes of the aggregation lattice. This theorem is the basis for sensitivity analysis here, where a change in some base cell in the lattice is propagated to all descendants in its upset. For the average drill-down measure a similar expression is determined. Moreover, sensitivity analysis might cause the multi-dimensional database to become corrupted, if the analysis is not carried out on cells in the base cube. To overcome this problem we proposed a correction procedure.

For sensitivity analysis in mixed systems of equations we introduced a matrix notation and we discussed the conditions for solvability. Because mixed systems are typically overdetermined the implicit function theorem cannot be applied. Therefore, we proposed a method to reduce the number of equations in the system and apply the implicit function theorem on a subsystem. Finally, an alternative method for what-if analysis in such systems is proposed.

Chapter 6

Case studies and Software implementation

6.1 Introduction

In this chapter, we present a number of case studies, to apply the methods/theory discussed in the previous chapters. In the case studies, typical business questions are addressed that emerge naturally when an analyst or decision-maker is analyzing a multi-dimensional business database. For example, business questions when analyzing a sales cube might be:

1. "Which products in the cube have good sales figures and which products have not?" (exception identification)

2. "What are the most important causes for the drop in profit in Spain 2008.Q1 compared to 2007.Q1?" (explanation)

3. "How is the profit on the aggregated year level affected when the revenues for product P1 are changed with 10% in the first quarter in Spain (c.p.)?" (sensitivity analysis)

Exception identification, explanation, and sensitivity analysis in the case studies is carried out mostly with a prototype software application.

The remainder of this chapter is organised as follows. In Section 6.2 (Case 1), the look-ahead explanation method is illustrated in a case study on interfirm comparison

with financial data about Dutch retail companies collected at Statistics Netherlands. This section is mainly based on the publications Daniels and Caron (2007) and Daniels and Caron (2009). Here explanation is based on solely business model equations and the reference values are obtained from an extra-organizational normative model and given by branch averages. In Sections 6.3 (Case 2a) and 6.4 (Case 2b), top-down and greedy explanation are illustrated in a case study analysing multi-dimensional sales and financial data. In these cases multiple analyses are carried out to show different aspects of our explanation methodology. Exceptional cell values in some cube C are identified first with statistical and managerial normative models. Subsequently, these exceptional cells are explained with top-down and greedy explanation. The generated explanation trees are pruned with various reduction methods. Parts of this section are published in Caron and Daniels (2012) and Caron and Daniels (2013). In Section 6.5 (Case 3), the top-down method for explanation is used in a case study on the analysis of real-life vehicle crime data. The research was executed for the PROTECT project (PROTECT 2006). In Section 6.6 (Case 4), sensitivity analysis is illustrated in a case study regarding supermarket sales data. What-if analysis is used in a system of drill-down and a system of business model equations. Parts of this section are published in Caron and Daniels (2010). If applicable, we globally describe the software applied in the sections. Finally, we draw conclusions in Section 6.7. The data used in the case studies is available in Appendix B, Appendix C, or is downloadable via the website http://www.emielcaron.nl/dissertation.html.

6.2 Case 1: Interfirm comparison at Statistics Netherlands

Interfirm comparison (IFC) is the standard way of measuring and comparing of a company's performance against its competitors or historic averages. By comparing the financial variables of a company with those of other companies in the same branch, the company can benchmark its performance against objective standards and gain insight into the weaknesses and strengths of the company. At present, the diagnostic process for IFC is mostly carried out manually by bankers, accountants and business consultants. The analyst has to explore large data sets in the domain of business

and finance to spot firms that expose exceptional behaviour compared to some norm behaviour. After abnormal behaviour is detected the analyst wants to find the causes, i.e. the set of financial variables responsible for the exceptional outcome. Traditional accounting methods's are variance decomposition and analysis of ratio's in a Du Pont model (Fridson and Alvarez 2002). Today's information systems for automated financial diagnosis and IFC have limited explanation or diagnostic capabilities. This functionality can be extended with the explanation formalism (see expression (4.1)), which supports the work of human analysts in diagnostic processes. In this case study the diagnostic process is largely automated and implemented in a computer program to support decision-makers. It is applicable to all kinds of underlying business models consisting of identities and behavioural equations. The Du Pont schema and the multi-dimensional business databases are special cases.

The following extensions are discussed. Firstly, a method for symptom detection is presented that takes into account the probability distribution of the business variable under consideration for diagnosis. Secondly, we apply the explanation methodology with look-ahead functionality (see Section 4.3) to deal with possible cancelling-out effects in the data set under consideration. These effects would be missed with the method of maximal explanation (see Section 4.2.6).

The method for diagnosis was originally implemented in PROLOG (Feelders 1993). This type of implementation has some advantages in terms of knowledge representation. However it also has some disadvantages in terms of applicability in an office environment and presentation of the output. Here we implemented the explanation method with look-ahead, as described in Algorithm 4, in MS Excel in combination with Visual Basic (VB).

6.2.1 Introduction

The business model M and data for IFC are obtained from Statistics Netherlands (2009)[1]. Statistics Netherlands is responsible for collecting, processing and publishing statistics to be used in practice, by policymakers and for scientific research. The business model M is derived from the production statistics for companies in the

[1]We thank Jeffrey Hoogland for his support at Statistics Netherlands.

Dutch *retail* and *wholesale trade* sectors. We use production statistics from the years 2001 and 2002. For both years, data sets with more than 5000 retail and wholesale companies are classified into branch sections. The model relations read:

1. $r_1 = r_2 + r_3 + r_4 + r_5$
2. $r_2 = r_6 - r_7$
3. $r_3 = r_8 - r_9$
4. $r_4 = r_{10} - r_{11}$
5. $r_5 = r_{12} - r_{13}$
6. $r_6 = r_{14} + r_{15}$
7. $r_7 = r_{23} + r_{24} + r_{25} + r_{26} + r_{27} + r_{28} + r_{29} + r_{30} + r_{31} + r_{32} + r_{33} + r_{34}$
8. $r_{14} = r_{16} + r_{17} + r_{18} + r_{19} + r_{20}$
9. $r_{15} = r_{21} + r_{22}$
10. $r_{23} = r_{35} + r_{36}$
11. $r_{24} = r_{37} + r_{38} + r_{39} + r_{40}$
12. $r_{25} = r_{41} + r_{42} + r_{43} + r_{44}$
13. $r_{26} = r_{45} + r_{46} + r_{47} + r_{48} + r_{49} + r_{50}$
14. $r_{27} = r_{51} + r_{52} + r_{53}$
15. $r_{28} = r_{54} + r_{55} + r_{56} + r_{57} + r_{58} + r_{59} + r_{60}$
16. $r_{29} = r_{61} + r_{62} + r_{63}$
17. $r_{30} = r_{64} + r_{65} + r_{66} + r_{67} + r_{68}$
18. $r_{32} = r_{69} + r_{70} + r_{71} + r_{72} + r_{73} + r_{74}$
19. $r_{33} = r_{75} + r_{76} + r_{77} + r_{78} + r_{79} + r_{80} + r_{81}$.

In short, three types of business equations are identified with depth $d = 4$ in M: *result* (eq. 1 through 5), *revenue* (eq. 6 through 8), and *cost* (eq. 9 through 19) *equations*. The variable (r_1) in the root equation gives the company's *total result before taxation*. This variable is split up into four types of results namely: total operating results (r_2), total financial results (r_3), total results allowances (r_4), and total extraordinary results (r_5).

A result variable is the difference between a revenues component and a cost component. Examples of revenues components are total operating revenues (r_6), financial

revenues (r_8), deductions from allowances (r_{10}) and extraordinary profits (r_{12}). Examples of cost components are total operating costs (r_7), financial expenses (r_9), additions to allowances (r_{11}) and extraordinary losses (r_{13}). Here the variable financial revenues is the sum of interests received, revenues from participations, payments of dividends, and profits from investments and other financial gains. Allowances (r_{11}) are the sum of internal provident funds, such as initial expenses, funds for business restructuring and maintenance. Furthermore, extraordinary profits are all gains that do not result from normal business management, like profits made on disposal of subsidiaries, fixed assets, and in foreign business units. Because Statistics Netherlands is interested in the structure of the operating revenues and costs, these variables are important in their statistics. Therefore, these variables are decomposed into lower level revenues and costs variables. In Appendix B the complete list of variables and their description is given. Here M consists purely of additive and difference relations. Our explanation methodology can also handle non-linear relations as shown in Example 4.2.1, if a consistent chain of reference objects is formed.

For the diagnosis of business performance we have to construct appropriate reference objects. Several factors that influence the business diagnosis results have to be taken into account, such as the Standard Industry Classification (SIC) for the retail and wholesale industries, and the size of the company. Therefore, computerized selections on the data set are made, such as supermarkets, liquor stores, do-it-yourself shops, etc. Within these subsets we make a further selection on the size class (small, medium, or large) of the companies. The company size classes are based on the number of employees of the firm in FTE's (full-time employees). The intervals for the different size classes are small ($1-9$ employees), medium ($10-99$ employees) and large (≥ 100 employees). In this way homogeneous subsets of the data for analysis are constructed. In addition, for the analysis *data is normalized* by dividing all variables in M by the total number of FTEs of each individual company.

The normative model R for IFC, the *industry average*, is computed by taking the mean value of all the companies in the selected normalized sample of a specific year for all variables (r_1 through r_{81}) in the business model. Industry averages are

computed as

$$r_p^r(\text{Industry average}, \text{Size class}, \text{Year}) = \frac{1}{N} \sum_{i=1}^{N} r_p^a(\text{Firm}_i, \text{Size class}, \text{Year}),$$

where N is the number of firms in the sample under consideration, and $p = 1, 2, \ldots, 81$. Here a consistent chain of industry averages is obtained because the equations in the business model are summarizations (See Section 4.7.2 for more detail). For example:

$$r_1^r(\text{Industry average}, \text{Size class}, \text{Year}) = \sum_{p=2}^{5} r_p^r(\text{Industry average}, \text{Size class}, \text{Year}).$$

Moreover, from the production statistics it is sometimes also possible to make historic comparisons, where R is selected to be a historical normative model. In that case, the reference objects for the business model variables are the values in one or more previous time periods.

6.2.2 Exception identification

Analysis is performed on a specific homogeneous sample selected out of the original data set with production statistics for the year 2001. The sample consists of 69 fashion shops with class size "medium". Exception identification in the data set starts with the variable total result before taxation (r_1) on the root level of the business model. This variable has a normal distribution. This was tested with the Shapiro-Wilks normality test with mean 11.30 (the industry average) and standard deviation 28.85. The population parameters of the distribution are estimated. The central question in the problem identification is: "*Which firms deviate significantly from their branch average in 2001?*" The symptom detection module of the diagnosis application identifies 9 firms that are higher or lower than the specified threshold value in the sample data set. Table 6.1 provides a full specification of the normative model. Here we select $\delta = 1.645$ corresponding to a probability of 95% in the standard normal distribution. With these test specifications we derive the following distribution of the number of firms over the three symptom types: 5 firms with symptom high, 60 firms with symptom normal and 4 firms with symptom low. For one of the fashion shops in the sample – the ABC-company – we present complete diagnostics. Moreover, the

Table 6.1: Specification of the normative model for the diagnostic example.

Slot name	Slot entry
Variable	Total result before taxation (r_1)
Norm object	Industry average (2001)
Industry	Fashion shops
Size class ($N = 69$ firms)	Medium
Distribution	Normal distribution $r_1 \sim N(11.30, 832.17)$
Threshold	$\alpha = .05$ (two one-tailed tests)

data is anonimized because Statistics Netherlands does not allow exposure of data on the micro level. The actual data for the company is r_1^a(ABC-company, Medium, 2001)= 61.75 and the reference data is r_1^r(Industry average(Fashion shops), Medium, 2001)= 11.30. For the ABC-company the detected symptom is "*high*" when comparing the actual result before taxation of the company with the branch average, because the one-tailed test $(61.75 - 11.30)/28.85 > 1.645$ is above the threshold value. Furthermore, the relative difference between the actual value and industry average for r_1 is $(61.75 - 11.30)/11.30 = 4.46$. Thus, the ABC-company is doing particularly well compared to its industry average, more than 4 times as good.

6.2.3 Explanation generation

We analyse the symptom

$$\langle \text{ABC-company}(2001),\ \partial r_1 = high,\ \text{branch average}(2001) \rangle$$

using the multi-level explanation method configured for *one-step look-ahead*, i.e. Algorithm 4 (see Section 4.3.2) is executed with $q = 1$. In other words, the following business question is addressed:

"*Why is total result before taxation (r_1) relatively high for the ABC-company compared with its branch average?*"

Here the selected reduction measure is RM_1 (see Section 4.6), where $T^+ = T^- = 0.85$ is taken. The explanation generation process starts with the root equation

in M. In Table 6.2 a comparison is made between the actual total result before taxation of the ABC-company and the branch average in the year 2001 for equation $M^{0;1}$. The equation $r_1 = r_2 + r_3 + r_4 + r_5$ holds for both actual and reference

Table 6.2: Actual and norm values for $M^{0;1}$.

	Actual	Norm	inf(x_i, y)	Difference (%)
r_1	61.75	11.30		446.46
r_2	60.42	14.79	45.62	308.52
r_3	1.33	-2.55	3.88	-152.16
r_4	0.00	-0.15	0.15	-100.00
r_5	0.00	-0.79	0.79	-100.00

values and we infer that $\text{Cb}_p = \{r_2\}$ and $\text{Ca}_p = \emptyset$. The variable r_2 (total operating results) explains 90.44% of the difference ∂r_1, and is therefore identified as the single parsimonious contributing cause because its value exceeds the fraction. Therefore, the result variables r_3, r_4 and r_5 are filtered out of the explanation because their influences are considered to be too small. However, instead of proceeding with purely explanation of the parsimonious contributing causes as in explanation without look-ahead, the extended method looks for potential cancelling-out effects in the analysis phase. The look-ahead procedure takes into account the effects of all variables on level 2 of M, i.e. the effects of the RHS-variables in equations 2, 3, 4 and 5 in M. This is illustrated graphically in the partial explanation tree depicted in Figure 6.1, where the curved black arrows "step over" the intermediate nodes on level 1, and point at the RHS variables of equation 2, 3, 4 and 5. In this figure, the straight black line indicates the identified parsimonious contributing cause, the straight grey lines indicate possible contributing causes and the dashed grey lines indicate possible counteracting causes.

In the analysis phase, function substitution is applied to find parsimonious causes, which were missed in the local explanation of differences. Equations 2 through 5 are substituted into the root equation and the following equation for explanation generation is derived:

$$M^{0;2}\text{: } r_1 = (r_6 - r_7) + (r_8 - r_9) + (r_{10} - r_{11}) + (r_{12} - r_{13}).$$

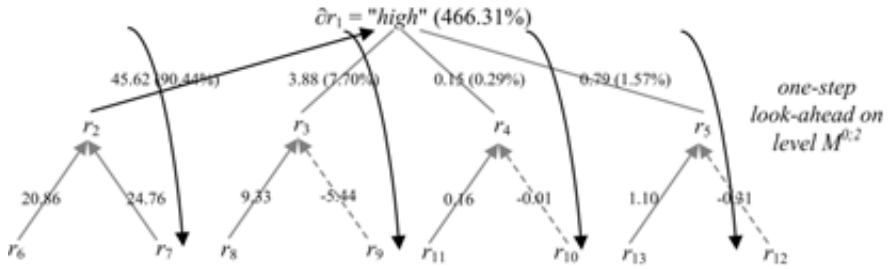


Figure 6.1: Illustration of one-step look-ahead in the analysis phase of the algorithm.

This equation is added to the set of business model equations. Notice that the specification of the event to explain ∂r_1 remains the same, but now equation $M^{0;2}$ is applied to explain the difference. Table 6.3 summarizes the results of our extended model of ABC-company's relatively high total result before taxation.

Table 6.3: Actual and norm values for $M^{0;2}$.

	Actual	Norm	inf(x_i, y)	Difference (%)
r_1	61.75	11.30		466.31
r_6	329.50	308.64	20.86	6.76
r_7	269.09	293.84	24.76	-8.42
r_8	11.17	1.84	9.33	507.07
r_9	9.83	4.39	-5.44	123.92
r_{10}	0.00	0.16	0.16	-100.00
r_{11}	0.00	0.01	-0.01	-100.00
r_{12}	0.00	0.31	-0.31	-100.00
r_{13}	0.00	1.10	1.10	-100.00

From the data in Table 6.3 it follows that $\text{Cb}_p = \{r_6, r_7, r_8\}$ and $\text{Ca}_p = \{r_9\}$. We now observe that the effects of causes r_8 and r_9 are significant at the specified fractions for parsimonious sets. These causes are identified as hidden causes, because $r_8 \in \text{Cb}_p(r_1)$ and $r_9 \in \text{Ca}_p(r_1)$. However, their parent variable $r_3 \notin \text{Cb}_p(r_1)$ according to Definitions 4.3 and 4.4 on page 94. These hidden causes would have been missed in an analysis without look-ahead, i.e. with maximal explanation. Figure 6.2.3 illustrates

the update process of the explanation tree in the reporting phase, where dashed black lines indicate counteracting causes. Notice that the variable total financial results (r_3) is not part of the explanation. This is indicated with a grey line.

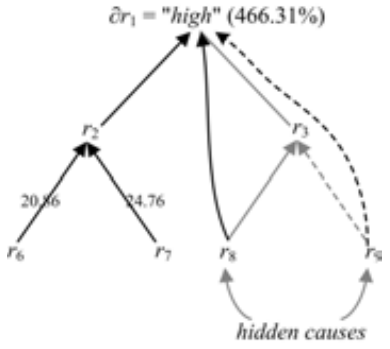


Figure 6.2: Explanation tree T^2 of hidden causes for $S = \{\partial r_1 = high\}$ on level $M^{0;2}$, generated in the reporting phase of the algorithm. In the tree, r_8 and r_9 are identified as hidden causes.

The diagnostic process is continued for all significant contributing causes. Thus, the next events to be explained are:

$$\langle \text{ABC-company}(2001),\ \partial r_6 = high, \text{branch average}(2001) \rangle$$

and

$$\langle \text{ABC-company}(2001),\ \partial r_7 = low,\ \text{branch average}(2001) \rangle.$$

The previous examples of different one-level explanations are now combined to a complete tree of causes. Figure 6.3 depicts the results of the explanation.

The following economic interpretation is given to the explanation tree in Figure 6.3. Recall the initial business question: Why are the ABC-company's total results before taxation relatively high compared to its branch average? Comparison of its results, revenues, and cost structures with those of the other companies show that the ABC-company's high results before taxation is due to a combination of comparatively high total operating results (r_2) and comparatively high financial revenues

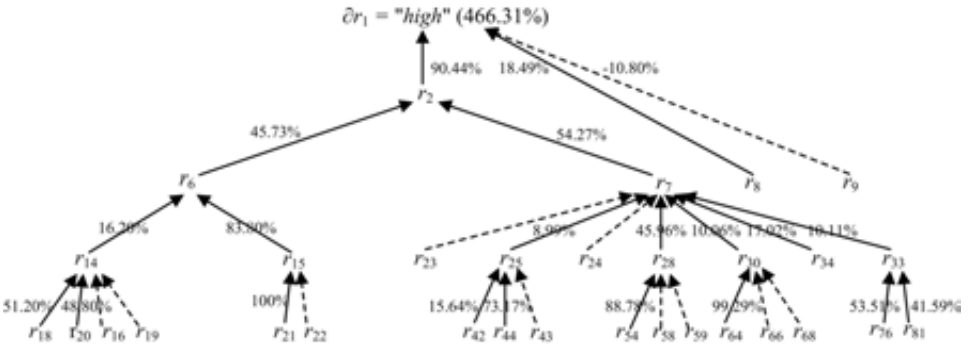


Figure 6.3: Diagnosis for $S = \{\partial r_1 = high\}$ at the ABC-company represented as the final explanation tree T^4.

(r_8), despite the fact of comparatively high financial expenses (r_9). Moreover, the ABC-company's high total operating results are explained by a combination of high total operating revenues (r_6) and low total operating costs (r_7). More specifically, the total operating revenues are high because of a combination of high total net sales (r_{15}) and additional revenues (r_{14}). The total operating costs of the company are low mainly because of low total housing costs (r_{28}), low total selling expenses (r_{30}), low total other operations costs (r_{33}) and low depreciations on tangible and intangible fixed assets (r_{34}), despite the fact that, costs of goods sold (r_{23}) and total costs of labour (r_{24}) are comparatively high, and so on. Notice that the explanation method, just as a human analyst, filters insignificant causes out of the explanation. In general, comparison of the result of our explanation method with human analysis shows clear similarities.

6.2.4 Software implementation

We present the most important concepts of the software for business diagnosis. The software is implemented in MS Excel in combination with Visual Basic. This application is initially developed to perform the experiments and analyses for the case study at Statistics Netherlands. However, the prototype software could also handle data

and business models from other domains. The software design is modelled and explained with a number of Unified Modeling Language (UML) *use cases*. In Figure 6.4, the main use case diagram is depicted. In this figure three actors are presented: the

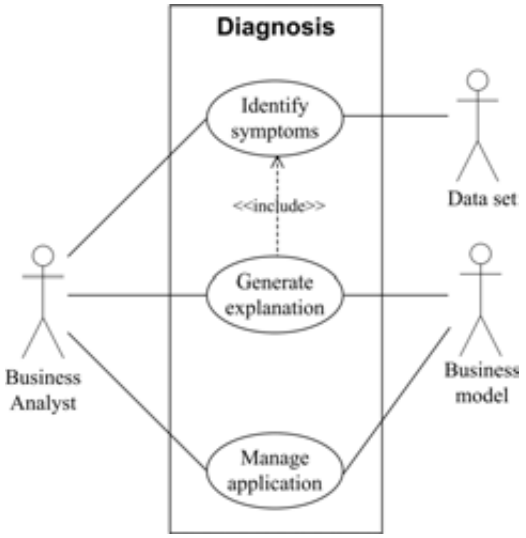


Figure 6.4: Main (UML) use case diagram for diagnosis.

human actor Business Analyst and two system actors labelled Data Set and Business Model respectively. The use case diagram represents the main functionality of the diagnostic application with the use cases: Identify Exceptions, Generate Explanation, and Manage Application. These use cases are explained in more detail in Appendix B, Section B.2.

With the use case Identify Exceptions the business analyst can detect symptoms in a data set. Here the analyst first has to start up the diagnostic application and load the data set. Subsequently, the analyst selects the appropriate normative model. Based on that, the reference values and various statistics are computed by the application. When the analyst specifies a threshold, exceptions can now be marked in the data set, for example, with a color. The analyst can select a certain exception for explanation generation, with the use case Generate Explanation. Here the

analyst has to specify the appropriate reduction measure and the method used for explanation, i.e. maximal explanation with or without look-ahead. The application computes the influence values for all variables in the business model with the selected method for explanation. Based on the influence values, the causes for the symptom are determined and with the reduction measure the set of causes is reduced to a set of significant causes. Subsequently, the analyst can view all the causes in the form of an explanation tree, which can be browsed in a tree viewer application. With the use case Manage Application, the analyst can maintain the business model that is associated with a certain data set. The analyst can add, change, and delete business equations in the business model.

Two screenshots of the main graphical user interfaces are depicted in Figure 6.5 and Figure 6.6. In Figure 6.5, the main user interface screen is depicted. This GUI controls the modules for symptom detection and explanation generation. In

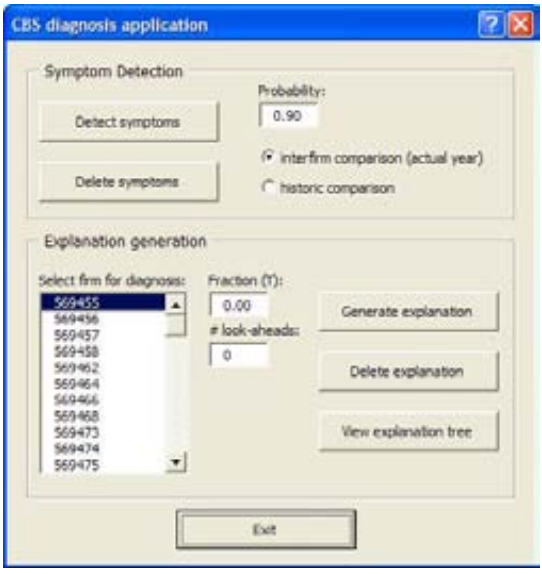


Figure 6.5: Main user interface of the CBS diagnosis application.

the upper part of the main user interface, symptoms are identified in the underlying

data set, based on interfirm or historic comparison. Here the analyst can select the desired threshold, expressed as a probability (e.g. .95 or .99) in the standard normal distribution. If the button "Detect symptoms" is selected, exceptional firms are highlighted in the sheet by a color scheme. The color red is used for a low exception and the color green for a high exception. If the button "Delete symptoms" is selected the exceptions are removed from the data sheet and the analyst might try a different threshold.

In the upper part of the main user interface, explanations can be generated by selecting a specific firm from the list. Before explanations can be generated, the analyst can specify the fraction T to construct parsimonious causes (RM_1) and the number of desired look-aheads in the business model. After that explanations can be generated by selecting the button "Generate explanation". Significant causes for the symptom are now computed and determined in the background. These causes can be represented as an explanatory tree by selecting the button "View explanation", then the procedure *tree-viewer* is invoked.

For the implementation of this procedure we applied tree programming to generate the tree of causes. The tree-viewer interface of the program is depicted in Figure 6.6. In the viewer the whole explanatory graph can be made visible by manipulating the tree. In addition, the tree of causes is projected on the explanatory graph by highlighting parsimonious causes with a color; green for a parsimonious contributing cause and red for a parsimonious counteracting cause. By clicking with the mouse on the cause under consideration, the details for the cause are made visible in the right panel of the screen, e.g. the influence value and the type of cause.

6.3 Case 2a: Financial OLAP database (Top-down explanation)

In this section (Case 2a) and the next (Case 2b), we study the GoSales financial OLAP database (IBM Cognos Software 2012). See also Example 2.1.1 in Chapter 2 for more information about this database. In the study, exceptional values are identified with both managerial and statistical normative models, and explained with both the top-down explanation method in Section 6.3, and the greedy explanation

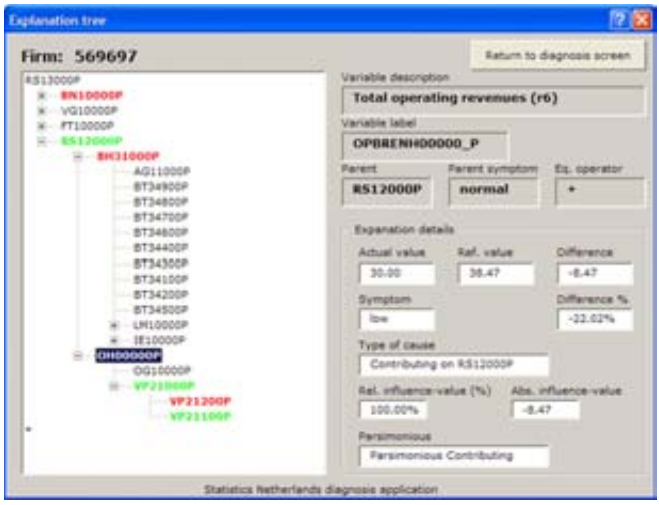


Figure 6.6: Tree viewer in the software for explanation. Here the results for the firm '569697' are depicted in the list box as an explanation tree on the screen's left panel. The variable 'Total operating revenues' (r_6) is selected in the tree and its explanation details are presented automatically on the screen's right panel.

method in Section 6.4.

6.3.1 Exception identification

The method for statistical exception identification (Section 3.6) is applied on the cube C = Year × Country × ProductLine, with slices

$$S^{\text{Year}=2001}(S^{\text{ProductLine}=\text{Personal Accessories}}(C)).$$

For the measure $y^{231}(C)$ = revenues$^{231}(C)$, an arbitrary context is taken from the financial databases to direct the business analyst to possible exceptional cells in this cube. The cube C is briefly denoted by C = Country × Personal Accessories. The cube's initial actual data is presented in Figure 6.7. It describes the revenue figures of the GoSales company in 20 countries, where the company is active for 5 types of product accessories in the year 2001.

Year	2001
ProductLine	Personal Accessories
Measure	Revenues

	Product Type				
Country	Watches	Eyewear	Knives	Binoculars	Navigation
Canada	€301,968.10	€162,196.78	€199,490.52	€194,169.58	€225,579.44
Germany	€349,894.36	€216,453.46	€291,174.56	€177,067.98	€283,151.24
France	€211,879.70	€147,412.66	€182,089.98	€201,389.98	€183,412.12
Mexico	€72,189.80	€38,760.80	€75,806.48	€62,400.96	€49,621.16
United States	€580,580.86	€289,605.28	€401,144.64	€81,882.00	€403,822.38
Japan	€359,831.80	€94,616.72	€190,541.36	€121,902.34	€175,577.84
Australia	€106,107.50	€75,941.36	€106,747.76	€107,962.58	€102,926.44
Austria	€151,922.54	€88,164.56	€122,915.96	€106,820.32	€98,012.42
China	€59,734.70	€29,288.36	€78,989.04	€60,992.20	€54,236.20
Italy	€176,139.28	€124,681.22	€169,902.54	€135,564.72	€131,588.50
Korea	€72,172.24	€57,824.34	€93,746.76	€86,651.74	€73,435.82
Netherlands	€257,168.24	€136,324.16	€203,095.58	€141,900.64	€148,383.72
Spain	€59,261.50	€30,115.34	€47,510.94	€44,858.60	€40,014.20
Sweden	€187,870.58	€96,728.36	€151,799.80	€148,417.24	€130,547.48
Switzerland	€120,966.00	€88,107.64	€144,843.66	€140,400.28	€90,412.46
Taiwan	€148,832.02	€85,623.98	€130,087.18	€114,723.04	€102,840.94
England	€282,944.54	€120,660.24	€174,573.94	€234,763.02	€174,347.28
Belgium	€80,282.90	€37,270.16	€53,600.56	€46,909.66	€39,468.28
Finland	€87,751.42	€44,794.44	€85,370.52	€67,914.16	€69,431.84
Brazil	€82,926.72	€56,625.68	€96,017.14	€82,870.78	€86,519.46

Pr	Color
0.99	
0.95	
0.90	
0.85	
0.01	
0.05	
0.10	
0.15	

Figure 6.7: Revenue figures, derived from the example financial database, organised per type of Personal Accessories (P^1) and Country (L^3) with a slice on the year 2001 (T^2). The colors indicate the level of exception. Notice that before exception identification the data is scaled by taking the natural logarithms of the data. Here the cell revenues(United States, Binoculars) is identified as a moderate "low exception" when the normative model R is a two-way ANOVA model.

The algorithm for exception identification is initially configured with a simple additive ANOVA model R, because the measure revenues is a continuous measure. The threshold for the scaled residuals is $\delta = 1.036$ for the high exceptions and $-\delta = -1.036$ for the low exceptions. In the algorithm the following steps are taken:

1. Data transformation. All the measure values in the cube are scaled by natural logarithms.
2. Statistical modeling. Here the simple additive two-way ANOVA model A_0

$$\hat{y}^{231}(\text{Country}, \text{Personal Accessories}) = \hat{\mu} + \hat{\lambda}_1(\text{Country}) + \hat{\lambda}_2(\text{Personal Accessories})$$

is applied initially.
3. Diagnostics.

(a) F-tests determine whether the main effects in the model are significant. The null hypotheses $H_{0;D_1^3}$ and $H_{0;D_2^1}$ say that there are no main effects for dimension level D_1^3 (Country) or dimension D_2^1 (Personal Accessories) respectively.

Table 6.4: Analysis of Variance Table for the additive model.

Response: log(Revenues)	df	Sum Sq	Mean Sq	f value	Pr(>F)	
Country	19	30.3627	1.5980	32.704	< 2.2e-16	***
Personal Accessories	4	3.9947	0.9987	20.438	1.835e-11	***
Residuals	76	3.7137	0.0489			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05'.'0.1 ' ' 1

In Table 6.4 we observe that the F-statistic (see Definition 3.13) for Country has the value 32.704 and for Personal Accessories the value 20.438. Both null hypotheses are clearly rejected because for Country $\Pr\{f > F_{(20-1);0.05}\} = 0.05$ and for Personal Accessories $\Pr\{f > F_{(5-1);0.05}\} = 0.05$. Therefore, it can be concluded that both the Country effect as well as the Personal Accessories effect should be included in the main-effects ANOVA model because these effects are significant.

In general, we use the simplest main-effects ANOVA model that meets the F-test. To single out exceptional cell values for explanation, the use of the full-effects ANOVA model, which includes (possible) interaction effects, is not strictly necessary, as long as the Gauss-Markov assumptions are not (too heavily) violated by the simplest main-effects model. The advantage of this model for the construction of explanations is that it produces consistent reference values (Theorem 4.7.1).

(b) Inspection of the two interaction plots in Figure 6.8 shows that the lines are fairly parallel. This suggest that interaction effects are negligible. Therefore, it is concluded that it is not necessary to consider the full-effects ANOVA model.

(c) The strict statistical normality tests, the Shapiro-Wilk test and the Kolmogorov-Smirnov test, reject the null hypothesis that the residuals come from a normal distribution. This is based on the statistics W = 0.834, p-value = 3.188e-09 and residuals D = 0.362, p-value = 8.253e-12 respectively, with significance level $\alpha = 0.05$. However, the normal Q-Q plot in Figure 6.9 does suggest that the model residuals are distributed normally to some extent, because most of the residual data points approximately lie on a 45° line. A number of outliers are clearly evident at both ends of the range.

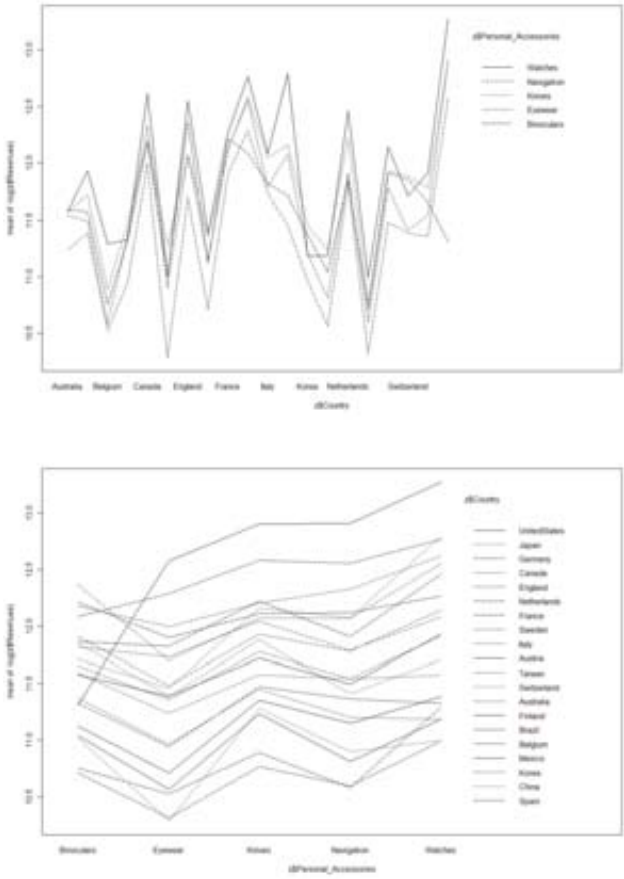


Figure 6.8: The interaction plot for dimension Location, level Country (upper figure) and interaction plot for dimension Product, level Personal Accessories (lower figure).

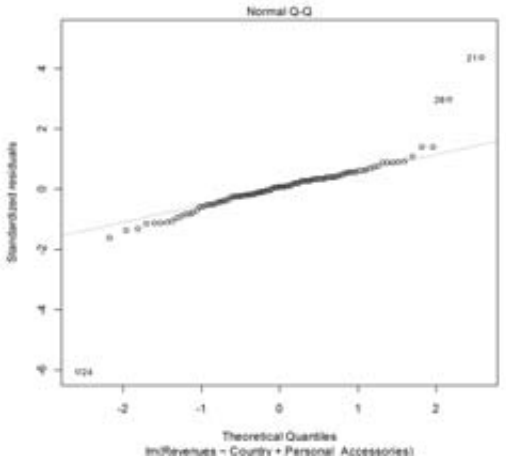


Figure 6.9: Normal Quantile-Quantile (Q-Q) plot of the standardized residuals.

(d) The homogeneity of variances is tested with the Bartlett test and the Fligner-Killeen test. Based on these tests we have to accept the null hypothesis of homoscedasticity if a significance level $\alpha = 0.05$ is used. The results for the Fligner-Killeen test are log(Revenues) by Personal Accessories by Country Fligner-Killeen: med Chi-squared = 3.2684, df = 4, p-value = 0.514 and log(Revenues) by Country by Personal Accessories Fligner-Killeen: med Chi-squared = 10.3277, df = 19, p-value = 0.9444. Since the probabilities are larger than 0.05, we conclude that the variances are the same for each cell in the cube.

4. Exception identification. Based on the above diagnostics the initial model is accepted for exception identification, because the effects in the model are all significant and there are no violations of the Gauss-Markov assumptions. Some additional model statistics are: $R^2 = 0.9025$ (see Definition 3.12), the model's F-statistic is 30.57 on 23 and 76 d.f., and the residual standard deviation is $s = 0.2211$ (see Definition 3.11). In Figure 6.10 the model residuals are plotted as a histogram. From this figure it can be concluded that there might be some low exceptions, indicated by the bar at the left-hand side of the histogram.

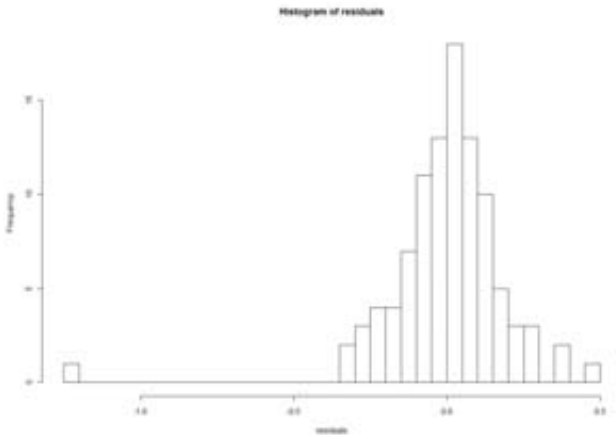


Figure 6.10: Histogram of model residuals.

All the scaled residuals (see Definition 3.4) in C are now compared with a range of threshold values given by the probability values 0.01, 0.05, 0.1, and 0.15. For the thresholds $\delta = 1.0364$ and $-\delta = 1.0364$, we find that the cell $c =$(United States, Binoculars) in the year 2001 is the only (low) exception with the scaled residual $\partial y(c)/s = -1.2120$, because $-1.2120 < -1.0364$. This exceptional cell is indicated with a yellow color in Figure 6.7. Then we (or the analyst) explore this deviating cell in more detail, to find the reasons for the deviation in the cell's downset.

A full specification of the event $\langle a, F, r\rangle$ to be explained is

$$\langle y^a(c), \partial y^{231}(c) = \text{“low”}, y^r(c) = \hat{y}(c)\rangle,$$

where $c =$ (2001, USA, Binoculars) (see expression 4.2). So, in words, we pose the following business question:

“Why are the revenues in the cell (2001, U.S.A., Binoculars) on level 231 *relatively low compared with the expected value for this cell, computed with the simple additive ANOVA model, in the cube C under consideration?”*

The exception is explained with top-down explanation (see Algorithm 5) in the downset $\{\downarrow c\}$. The algorithm is executed multiple times over various drill-down

analysis paths in the downset by applying appropriate additive ANOVA models. The following analyses are considered:

1. Explanation in the Time dimension along the path [231] → [131] → [031];
2. Explanation in the most specific dimension: Time along path [231] → [131], Location along path [231] → [221], and Product along path [231] → [230].

6.3.2 Explanation generation in analysis 1

In the first analysis, the business analyst wants to explain the event solely in the Time dimension over the drill-down path $p = [231] \rightarrow [131] \rightarrow [031]$, on the Quarter and Month level. This is an application of reduction method RM_{3b}. As an additional reduction method, RM_1 is applied here with fraction $T^+ = T^- = 0.9$, to remove the effect of marginal causes. For each cell on the path p in the downset $\{\downarrow c\}$, both the actual as well as the reference value are required for explanation of the event. Here y is the additive measure revenues, therefore the actual values are directly available by applying drill-down operators on the cell c. For example, the operation $R_T^{-1}(c)$ gives the actual values for cells on the Quarter level. This corresponds with the additive drill-down equation

$$y^{231}(c) = \sum_{i=1}^{4} y^{131}(2001.\mathrm{Q}_i, \mathrm{U.S.A.}, \mathrm{Binoculars}).$$

Moreover, the reference values for cells in p are computed by application of the same type of normative model R, as used for the computation of the reference value for the root cell c. Therefore, for each cell $c' = R_T^{-1}(c)$ its reference values are derived with the additive ANOVA model A_1

$$\hat{y}^{131}(c') = \hat{\mu} + \hat{\lambda}_1(2001.\mathrm{Quarter}) + \hat{\lambda}_2(\mathrm{Country}) + \hat{\lambda}_3(\mathrm{Personal\ Accessories}),$$

in the context cube 2001.Quarter × Country × Personal Accessories. See the data in Figure C.1 in Appendix C.2.

A_1 is a specialized ANOVA model for the quarters, which is a specialization of ANOVA model A_0 within an unfolded Time dimension (see page 121, Case 2). The model contains the effects of the ANOVA model that was used for the parent cell,

Table 6.5: Data for explanation of $\partial y^{231}(c)$ ="low" in the Time dimension, on the level Quarter in the context cube 2001.Quarter × Country × Personal Accessories.

	actual	reference	$\inf(y^{131}(c'), y^{231}(c))$	relative inf.
(2001,.,.)	81,822.00	331,445.52		
(Q1,.,.)	26,230.40	71,163.59	-44,933.19	0.18
(Q2,.,.)	18,738.80	84,500.17	-65,761.37	0.26
(Q3,.,.)	12,912.80	79,115.04	-66,202.24	0.27
(Q4,.,.)	24,000.00	96,666.71	-72,666.71	0.29

plus the 2001.Quarter-effects. The two conditions for Theorem 4.7.1 are fulfilled, and therefore the following equations holds:

$$\hat{y}^{231}(c) = \sum_{i=1}^{4} \hat{y}^{131}(2001.\mathrm{Q}_i, \mathrm{U.S.A.}, \mathrm{Binoculars}).$$

Hence, the drill-down equation for the actual quarter values holds also for the reference quarter values. Next in Table 6.5 a comparison is made between the actual and the reference values for the cell c to explanation the Time dimension at the level Quarter. In this table the influence values are computed using (4.12). Because the drill-down equation holds for both actual and reference values, Theorem 4.2.1 applies, and the inf-measure is correctly interpreted as a quantitative specification of the change in $y^{231}(c)$ that is explained by a change in $y^{131}(c')$. In the table relative influences are computed by $(y^a(c) - y^r(c))/\inf(y(c'), (c))$. From the data in the table it can be concluded that $\mathrm{Cb}_p = \{$(Q1,.,.), (Q2,.,.), (Q3,.,.), (Q4,.,.)$\}$, since all the contributing causes are needed to explain the desired fraction T^+. Because in this explanation step no parsimonious counteracting causes are identified, $\mathrm{Ca}_p = \emptyset$.

Because all causes on the Quarter level are significant, the top-down algorithm continues explanation for all quarters on their constituent months, i.e. the next level in the analysis path p. To determine the influences of these individual months, reference values have to be computed for each month. For each cell $c'' = R_T^{-1}(c')$ its reference value is computed by ANOVA model A_2

$$\hat{y}^{031}(c'') = \hat{\mu} + \hat{\lambda}_1(2001.\mathrm{Month}) + \hat{\lambda}_2(\mathrm{Country}) + \hat{\lambda}_3(\mathrm{Personal\ Accessories}),$$

in the context cube 2001.Month × Country × Personal Accessories. The model A_2

Table 6.6: Data for explanation of ∂y^{131}(2001.Q4, U.S.A, Binoculars)="low" in the Time dimension, on the level Month in the context cube 2001.Quarter.Month × Country × Personal Accessories.

	actual	reference	inf($y^{031}(c'')$, $y^{131}(c')$)	relative inf.
(2001.Q4,.,.)	24,000.00	96,666.71		
(Oct,.,.)	18,560.00	50,220.16	-31,660.16	0.44
(Nov,.,.)	0.00	22,417.20	-22,417.20	0.31
(Dec,.,.)	5,440.00	24,029.35	-18,589.35	0.26

is a specialization of model A_1 within the Time dimension, from the Quarter to the Month level (see page 120, Case 1b).

In this way, consistent reference values are obtained for each quarter Q_i, given by

$$\hat{y}^{131}(2001.\mathrm{Q}_i, \mathrm{U.S.A.}, \mathrm{Binoculars}) = \sum_{j=1}^{3} \hat{y}^{031}(2001.\mathrm{Q}_i.\mathrm{Month}_j, \mathrm{U.S.A.}, \mathrm{Binoculars}),$$

where $i = 1, 2, 3, 4$ and $j = 1, 2, 3$. The reference objects are consistent because the ANOVA model applied at the Month level. It is a specialization of the ANOVA model applied on the Quarter level, and therefore the conditions for Theorem 4.7.1 are met. As an example, a comparison is made in Table 6.6 between the actual and the reference values for the cell (2001.Q4, U.S.A, Binoculars) and its children on the Month level. From the data in the table, it can be concluded that Cb_p = {(Q4.Oct,.,.),(Q4.Nov,.,.),(Q4.Dec,.,)}, since all the contributing causes are needed to explain the desired fraction T^+. Obviously, $\mathrm{Ca}_p = \emptyset$. All the months of the last quarter show the same pattern: in each month the realized revenues are relatively low in the U.S.A for the ProductType Binoculars. In particular, the month October stands out as a large contributing cause. It explains 44% of ∂y^{131}(2001.Q4, U.S.A, Binoculars) and 13% of ∂y^{231}(2001, U.S.A, Binoculars). In addition, the previous examples of one-level, top-down, explanations for the symptom, are combined to a complete diagnosis in the Time dimension. The explanation tree in the lower part of Figure 6.11 summarizes the results. In this figure, the straight lines indicate parsimonious contributing causes and dotted lines indicate counteracting causes, the numbers on the lines indicate the relative values for the influence measures, and the ratios indicate the specificity measure (S) value of the explanation step.

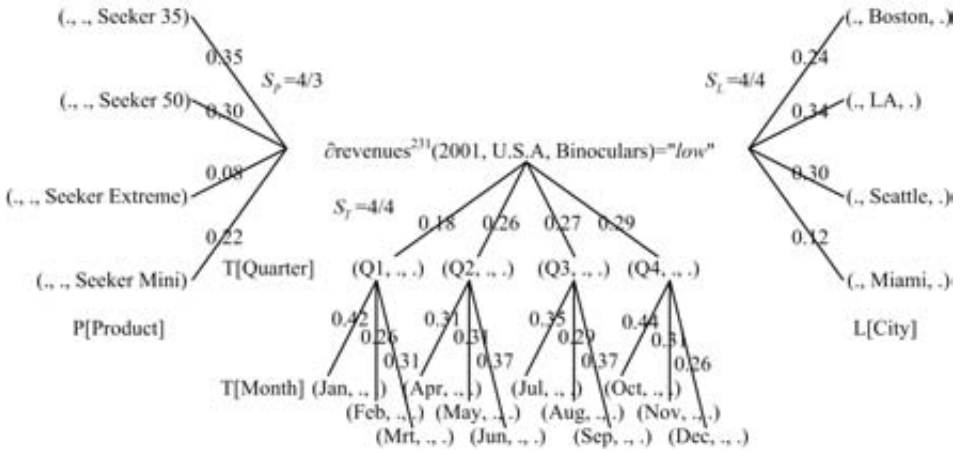


Figure 6.11: Explanation trees that partially explain the exceptional cell revenues(2001, U.S.A, Binoculars) = "low" in the Product (P), Time (T), and Location (L) dimension.

In addition, we give a business interpretation of the complete explanation tree in the Time dimension. From its inspection it can be concluded that the revenues in the cell c declined because the revenues decreased in all quarters and all months, they basically all show the same pattern. However, the largest part of the decrease, 56%, occurred in the last two quarters on the year. Especially, the months July, September, and October are relatively large causes and are sure candidates for further inspection.

Notice that the root symptom of the explanation tree is always an exceptional cell value. Other parts of the explanation tree, i.e. significant causes that explain the symptom, can be exceptional values in some context cube, but this is not required for the explanation of the initial symptom. Moreover, the structure of the final explanation tree depends on the class of normative models R that is used, because the reference object r in the explanation formalism (see expression 4.1) changes. The same reasoning holds for different normative models within one class R. For example, when in the analysis under consideration a different ANOVA model, e.g. the main-effects ANOVA model with only the Country effect included, would have been used

in explanation of the same symptom, the resulting explanation tree would very likely have a different structure.

6.3.3 Explanation generation in analysis 2

In the second analysis we apply reduction method RM_2. The reference values for the dimensions are computed by main-effects ANOVA models. For explanation in the Location dimension along the path [231] → [221] the reference values are computed with the ANOVA model

$$\hat{y}^{131}(c') = \hat{\mu} + \hat{\lambda}_1(\text{Country.City}) + \hat{\lambda}_2(\text{Personal Accessories}),$$

in the context cube Country.City × Personal Accessories, and for explanation in the Product dimension on the path [231] → [230] the reference values are computed with

$$\hat{y}^{131}(c') = \hat{\mu} + \hat{\lambda}_1(\text{Country}) + \hat{\lambda}_2(\text{Personal Accessories.Product}),$$

in the context cube Country × Personal Accessories.Product. Notice that reference values for the Time dimension on the level Quarter were already computed in the previous case.

Explanatory details for the explanation in the Product dimension are given in Table 6.7. From the data in this table, it can be concluded that Cb_p = {(.,.,Seeker 35), (.,.,Seeker 50), (.,.,Seeker Mini)}, since these three contributing causes explain the desired fraction T^+, as shown by $(0.35+0.3+0.22)/1 \geq 0.9$ (see expression (4.5)), and $\text{Ca}_p = \emptyset$. Remarkable here is that the actual values for the parsimonious causes

Table 6.7: Data for explanation of $\partial y^{231}(c)$ ="low" in the Product dimension, on the level Product.

	actual	reference	$\inf(y^{230}(c'), y^{231}(c))$	relative inf.
(.,.,Binoculars)	81,822.00	331,445.52		
(.,.,Seeker 35)	0.00	86,255.75	-86,255.75	0.35
(.,.,Seeker 50)	0.00	75,055.49	-75,055.49	0.30
(.,.,Seeker Extreme)	81,822.00	101,933.53	-20,111.53	0.08
(.,.,Seeker Mini)	0.00	68,200.75	-68,200.75	0.22

are 0, i.e. no revenues were generated for these products. Next the specificity value for this explanation step is computed as $S_P = 4/3$, because the number of possible causes is $|R_T^{-1}(c)| = 4$, and the number of identified actual causes is $|\mathrm{Cb}_p| + |\mathrm{Ca}_p| = 3 + 0 = 3$ (Equation (4.23)).

In summary, three partial explanation trees are depicted in Figure 6.11, from west, south, to east, corresponding with the explanation trees for the Product, Time, and Location dimension, respectively. For the root level in each of the trees we have computed the measure of specificity S with Equation (4.22) for each dimension. For all the explanation steps in the downset of the exceptional cell c that are possible, the specificity value range is $S_P \geq S_T \geq S_L$ ($4/3 \geq 4/4 \geq 4/4$). With RM_2 the most specific explanation step is taken, in this case in the direction of the Product dimension. Top-down algorithm now proceeds the explanation process with the cells (.,., Seeker 35), (.,., Seeker 50), and (.,., Seeker Mini). For each of these cells the measure of specificity is applied again and the explanation step is selected with the highest specificity value, and so on. This mechanism can be continued until the base cube is reached.

6.4 Case 2b: Financial OLAP database (Greedy explanation)

In this section, we illustrate the greedy algorithm for explanation (Algorithm 6).

6.4.1 Exception identification

We identify exceptions in the GoSales cube $C = 2001 \times$ Country for the measure profit (y). The cube's data is presented in Figure 6.12. In this case, the profit figures on the previous year, represented by the cube $C' = 2000 \times$ Country, are used as a historical normative model (see Section 3.2.3). The cube of reference data is depicted in Appendix C, Figure C.6. Algorithm 1 is applied to identify exceptional values in C:

1. Reference values are taken from C';

Year	2001
Product	All-Products
Measure	Profit

Country	Year 2001
Canada	€141,777.29
Germany	€139,862.55
France	€292,408.31
Mexico	€106,229.07
United States	€660,789.73
Japan	€396,443.34
Australia	€134,915.71
Austria	€234,421.81
China	€325,409.44
Italy	€343,646.06
Korea	€226,896.86
Netherlands	€199,690.65
Spain	€86,248.94
Sweden	€369,004.16
Switzerland	€136,396.48
Taiwan	€263,271.48
England	€414,341.77
Belgium	€88,679.70
Finland	€130,403.65
Brazil	-€134,054.69

Pr	Color
0.99	
0.95	
0.90	
0.85	
0.01	
0.05	
0.10	
0.15	

Figure 6.12: Profit figures of the year 2001, organised per Country (L^3) and All-Products (P^3). The colors indicate the level of exception. Here the normative model is based on the profit figures of the previous year, see Figure C.6 for the data.

2. Cell residuals are computed by $\partial y(C) = y(C) - y(C')$;

3. Scaled cell residuals (see Definition 3.4) are computed by $\partial y(c)/\sigma$, where $\sigma = 77,409.62$, and $c \in C$;

4. The scaled residuals are compared with a range of threshold probability values taken from the standard normal distribution: 0.99, 0.95, 0.9, and 0.85 for high exceptions and 0.01, 0.05, 0.1, and 0.15 for low exceptions.

5. The following cells in C are marked as exceptions with a color scheme: the cell (2001, China) is a moderate high exception and the cells (2001, Canada), (2001, The Netherlands), (2001, Spain), (2001, Sweden), and (2001, Belgium) are moderate low exceptions. The largest low exception is found in the cell c =(2001, The Netherlands), where $\partial y(c) = y^a(c) - y^r(c') = 199,690.65 - 378,324.70 = -178,634.05$, $\partial y(c)/\sigma = -2.31$, and $-\delta = -1.64$ (= Pr. 0.05).

Subsequently, we identify possible causes in $\{\downarrow c\}$. A full specification of the event to be explained is

$$\langle y^a(c), \partial y^{233}(c) = \text{“low”}, y^r(c')\rangle,$$

where $c =$ (2001, The Netherlands) and $c' =$ (2000, The Netherlands). In words:

“*Why is the measure profit in the cell (2001, The Netherlands) on level* 233 *relatively low compared with the reference value for this cell, the profit in the previous year in The Netherlands on the aggregated product level ‘ALL-Products’, represented by the cell (2000, The Netherlands), in the cube C under consideration?*”

This event is explained with greedy explanation (Algorithm 6) in the downset $\{\downarrow c\}$. Subsequently, we explain multiple exceptional cell values (events) in the cube C to formulate a generic explanation (Section 4.6.5 with RM_5).

6.4.2 Greedy explanation generation

Here the exceptional cell $\partial y(c)$ is explained with greedy explanation in:

1. the Product dimension;
2. the Time dimension;
3. a combination of the Product and Time dimension.

Table 6.8 shows the aggregated table which is the basis for greedy explanation in the Product dimension, for the city of Amsterdam. The complete table is composed out of 143 records. From the data in the table we can conclude that y^{222}(., ., Camping Equipment) is the largest contributing cause in the Product dimension and y^{220}(., ., Golf Equipment.Irons. Hailstorm Titanium Irons) is the largest counteracting cause. In Figure 6.13, the results are depicted in an explanation tree, which reports only the 10 largest contributing causes for the symptom in the Product dimension (see RM_4). Table 6.9 shows the aggregated table which is the basis for greedy explanation in the Time dimension for the city of Amsterdam. The complete table has 16 records. Observe that y^{123}(Quarter 2, ., .) is the largest contributing cause in the Time dimension, which explains 81% of the symptom, and y^{123}(Quarter 1, ., .) is the

Table 6.8: Aggregated table for the Product dimension where the actual object is profit(2001, Netherlands), the norm object is profit(2000, Netherlands), and the influence values for instances within the Product dimension are related to the exceptional cell value profit$^{223}(c)$.

Nr.	ProductLine P^2	ProductType P^1	Product P^0	Actual (2001)	Norm (2000)	Rel. Inf.
	All	All	All	199,690.65	378,324.70	
1	Camp. Equip.	All	All	-67,075.17	16,796.14	0.47
2	Mount. Equip.	All	All	49,098.42	86,611.58	0.21
3	Camp. Equip.	Tents	All	-121,318.02	-93,058.71	0.16
4	Golf Equip.	All	All	106,474.92	131,752.22	0.14
5	Pers. Acces.	All	All	105,043.91	130,653.60	0.14
6	Golf Equip.	Woods	All	55,612.59	76,180.27	0.12
7	Camp. Equip.	Packs	All	18,250.44	37,208.57	0.11
8	Camp. Equip.	Lanterns	All	20,309.57	37,713.44	0.10
9	Mount. Equip.	Rope	All	6,602.70	23,717.68	0.09
10	Camp. Equip.	Tents	Star Dome	-50,067.72	-33,682.12	0.09
...	...	...	...	...	...	...
143	Golf Equip.	Irons	Hail. Tit. Ir.	14,780.06	5,468.46	-0.05

Table 6.9: Aggregated table for the Time dimension where the actual object is profit(2001, Netherlands), the norm object is profit(2000, Netherlands), and the influence values for instances within the Time dimension are related to the exceptional cell profit$^{223}(c)$.

Nr.	Quarter T^1	Month T^0	Actual (2001)	Norm (2000)	Rel. Inf.
	All	All	199,690.65	378,324.70	
1	Quarter 2	All	49,683.14	194,707.50	0.81
2	Quarter 2	April	24,531.46	86,596.44	0.35
3	Quarter 2	June	29,253.24	74,822.76	0.26
...	...	...	...	...	...
16	Quarter 1	All	24,520.99	13,446.28	-0.06

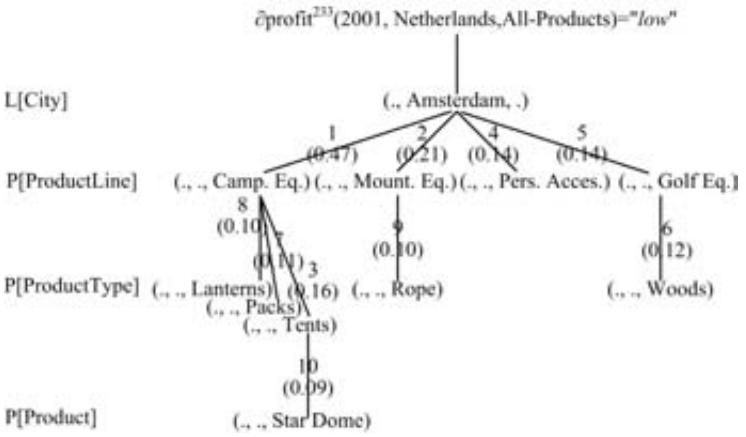


Figure 6.13: Greedy explanation in the Product dimension reporting the 10 largest contributing causes.

largest counteracting cause. The explanation trees with the 10 largest contributing causes (see RM_4) in the Time dimension are depicted in Figure 6.14. Table 6.10 shows the aggregated table which is the basis for greedy explanation in the Product and Time dimension, for the city of Amsterdam. The complete table has 2413 records. From this table we can find the cell, on the lowest level in the exceptional cell's downset, with the largest positive influence, i.e. the largest contributing cause. The corresponding drill-down path is (see record 32 in Table 6.10):

y^{223}(2001, Amsterdam, All-Products) $\rightarrow$ y^{123}(Q2, ., .) $\rightarrow$ y^{023}(May, ., .) $\rightarrow$
y^{022}(., ., Golf Equipment) $\rightarrow$ y^{021}(., ., Woods) $\rightarrow$
y^{020}(., ., Hailstorm Titanium Woods Set).

The cell with the largest negative influence, is corresponding to the following drill-down path (see record 2413 in Table 6.10):

y^{223}(2001, Amsterdam, All-Products) $\rightarrow$ y^{123}(Q3, ., .) $\rightarrow$
y^{122}(., ., Camping Equipment) $\rightarrow$ y^{121}(., ., Tents) $\rightarrow$ y^{120}(., ., Star Gazer 3).

Hidden contributing causes in an aggregated table are records that correspond with cells with an influence value larger than, or equal to, T^+, but that have an ancestor with an influence value that is smaller than T^+. We now identify possible hidden

Table 6.10: Aggregated table for both the Product and Time dimension where the actual object is profit(2001, Netherlands), the norm object is profit(2000, Netherlands), and the influence values for instances within the Product and Time dimension are related to the exceptional cell profit$^{223}(c)$.

Nr.	Quar. T^1	Mon. T^0	P.Line P^2	P.Type P^1	Prod. P^0	Actual (2001)	Norm (2000)	Rel. Inf.
	All	All	All	All	All	199,690.65	378,324.70	
1	Q2	All	All	All	All	49,683.14	194,707.50	0.81
2	All	All	Camp. Eq.	All	All	-67,075.17	16,796.14	0.47
3	Q2	Apr.	All	All	All	24,531.46	86,596.44	0.35
...	...	...	...	...	...	...	...	...
32	Q2	May	Golf Eq.	Woods	H.Tit.Set	5,192.82	20,337.40	0.08
...	...	...	...	...	...	...	...	...
59	Q3	All	Pers. Acc.	Knives	Surv.Edge	594.00	9,368.75	0.05
...	...	...	...	...	...	...	...	...
73	Q3	All	All	All	All	60,119.14	67,569.40	0.04
...	...	...	...	...	...	...	...	...
2413	Q3	All	Golf Eq.	Tents	St.Gaz.3	-5,693.20	-23,481.20	-0.10

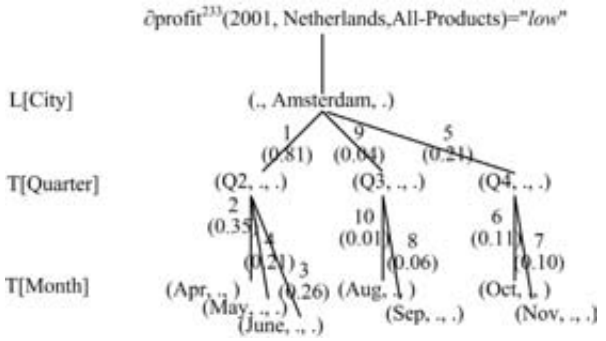


Figure 6.14: Greedy explanation in the Time dimension reporting the 10 largest contributing causes.

causes. From Table 6.10, we observe that the cell (2001.Q3, Amsterdam, Personal Accessories Knives.Bear Survival Edge) is a parsimonious hidden contributing cause (see record 59), because its influence on the symptom is larger than T^+ ($0.05 \geq T^+$), despite the fact that its ancestor cell (2001.Q3, Amsterdam, All-Products) has an influence on the symptom that is smaller than T^+ ($0.04 < T^+$) (see record 73). The identification of the hidden cause corresponds with the following drill-down analysis path:

y^{223}(2001, Amsterdam, All-Products) $\rightarrow y^{123}$(Q3, ., .) $\rightarrow y^{122}$(., ., Pers. Accessories) $\rightarrow y^{121}$(., ., Knives) $\rightarrow y^{120}$(., ., Bear Survival Edge).

Notice that record 2413 in the table corresponds with a hidden counteracting cause ($-0.10 \leq T^-$).

The explanation tree with the top-15 contributing causes in both the Time and Product dimension, with a slice on Quarter 2, is depicted in Figure 6.15. The causes in the tree are the 15 largest contributing causes in the total set of 2413 causes for Quarter 2. The business interpretation of the explanation tree is that the exceptional cell value is explained mainly by lower profits made in Quarter 2 (81%), in all its constituent months, over all product lines, except the line Outdoor Protection. An interesting, specific explanation of the symptom is the relative low profits over the path (, ., Camping Equipment) $\rightarrow$ (., ., Tents) $\rightarrow$ (., ., Star Gazer 3).

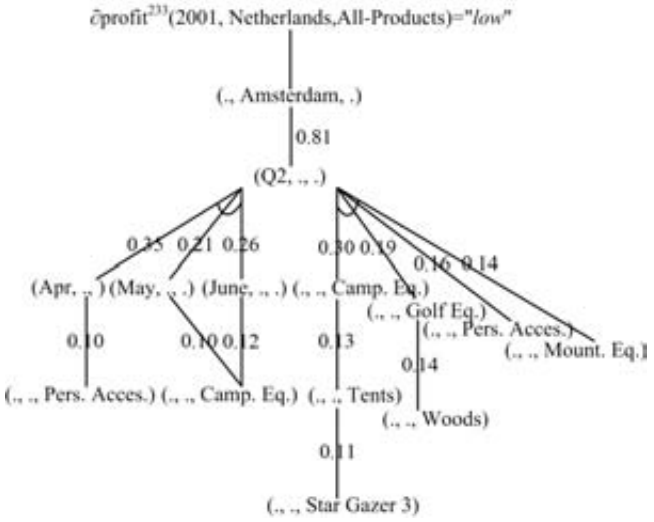


Figure 6.15: Greedy explanation in the Time and Product dimension with a slice on Quarter 2, reporting the 15 largest contributing causes in this Quarter.

Remark 6.4.1. Notice that in the above example analyses, hidden causes for the event are identified (automatically) by the greedy explanation method, because all significant causes at some level in the lattice are reported. In contrast, if the same event was explained top-down in $\{\downarrow c\}$ with Algorithm 5, without configuring the look-ahead functionality, possible hidden causes might be missed in the explanation. This is shown in the following example. We now explain the event with top-down explanation solely in the Product dimension, in the context cube Year.Q3 $\times$ Amsterdam. The algorithm is configured with $T^+ = 0.95$ for RM_1, i.e. in each step at least 95% of the difference is explained. From the data in Table C.4 in Appendix C, it can be concluded that, (., ., Personal Accessories.Knives) and (., ., Personal Accessories.Knives.Bear Survival Edge) are elements of Cb_p because its influences are $\geq T^+$, and (., ., Camping Equipment.Tents) and (., ., Camping Equipment.Tents.Star Gazer 3) are elements of Ca_p because its influence are $\leq T^-$. Their parent, the cell (2001.Q3, Amsterdam, All-Products), is not included in the set of parsimonious causes, due to the neutralization

of positive by negative influence values. Those causes are clearly contributing and counteracting hidden causes (see Definitions 4.3 and 4.4). The problem of cancelling-out effects is illustrated graphically in Figure 6.16. The neutralized cause Quarter 3 is represented with a grey line in the figure.

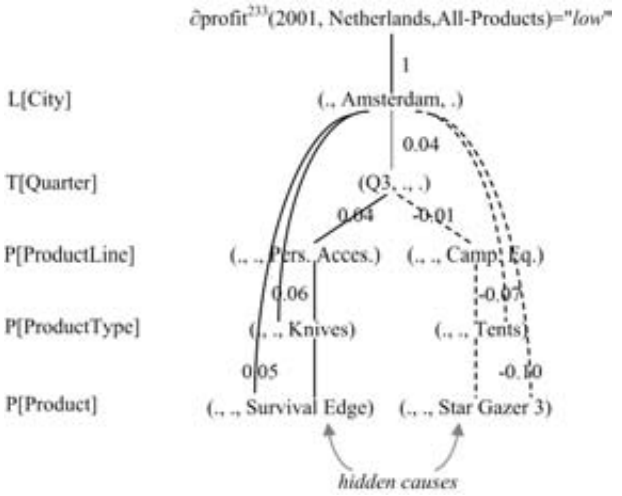


Figure 6.16: The presence of hidden causes in the downset of Quarter 3. This quarter itself is not part of the explanation of the identified symptom due to neutralization. However, cells in the downset, (Accessories.Knives.Bear Survival Edge) and (Camping Equipment.Tents.Star Gazer 3) are hidden parsimonious causes.

6.4.3 Generic explanation generation

In this section a generic explanation is formulated for all the identified low symptoms depicted in the direction of the Product dimension of Figure 6.12 (See Section 4.6.5 for RM_5). We try to find a pattern explaining the declining profit level in the 5 countries. They are represented by the range of cells in the cube C: (2001, Canada), (2001, The Netherlands), (2001, Spain), (2001, Sweden), and (2001, Belgium). For each of these symptoms a top-15 report of the largest contributing causes was generated. Table 6.8 presents these causes for The Netherlands. Tables C.5, C.6, C.7, and C.8 in Appendix

C.3 represent the causes for the other countries. It follows that $\text{sCb}_{\text{similarity}} = \{(2001, X, \text{Camping Equipment}), (2001, X, \text{Camping Equipment.Tents}), (2001, X, \text{Camping Equipment.Lanterns})\}$ (see Equation 4.24), where $X \in \{\text{Canada, The Netherlands, Spain, Sweden, Belgium}\}$. The contributing causes in $\text{Cb}_{\text{similarity}}$ are all parsimonious, because they all exceed T^+. These results are summarized in Figure 6.17. In this

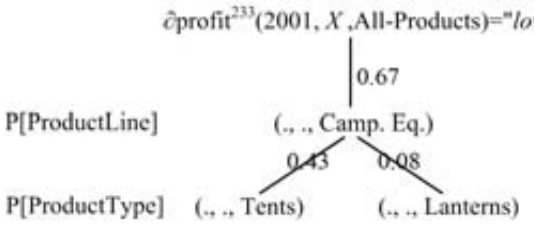


Figure 6.17: Generic explanation for the cell $\partial\text{profit}^{233}(2001, X, \text{All-Products}) =$ "low", where $X \in \{\text{Canada, The Netherlands, Spain, Sweden, Belgium}\}$, in the Product dimension.

figure the average influence values over the five countries are depicted next to lines that connect the causes. An obvious recommendation could be, based on a inspection of this explanation tree, to focus on extra marketing activities for products in the Camping Equipment product line, especially for the product types Tents and Lanterns.

6.4.4 Software implementation

The software for the explanation of exceptional values in OLAP databases and the scalability of the approach is presented briefly in this section. The software design has large similarities with the software described in Section 6.2.4, but also has some important differences. Different actors compared to Figure 6.4 are used. The system actors are:

- the OLAP cube. The diagnostic application needs to connect with an OLAP data cube, represented in MS Access or MS Excel, instead of a flat data file.

- the R statistical software package. For the identification of exceptions, the software implementation uses functions from the software package "R for statistical computing" (The R Foundation for statistical computing 2011).

- the Multi-dimensional/Business model. For explanation of an exceptional cell in an OLAP cube an explicit model of the multi-dimensional and/or business model is required. This model can, in theory, automatically be obtained from the OLAP cube. In our prototype software, this model needs to be modelled/defined by hand. The software environment to define this model is depicted on the left hand side of Figure 6.19.

In the R package the following functions are used in the algorithm for exception identification (see Section 3.6):

- lm() fits a linear model. In addition, the functions summary() is used to produce summary statistics for the fitted model;

- anova() computes the analysis of variance table with F-tests for the fitted model;

- interaction.plot() generates an interaction plot;

- shapiro.test() and ks.test() are used to test for normality;

- qqplot() generates a Q-Q plot;

- bartlett.test() and fligner.test() are used to test for homogeneity.

Moreover, the graphical capabilities of R are used to produce statistical figures to illustrate various statistical tests (see Section 6.3.1). Figure 6.18 depicts the UML class diagram of the application for diagnosis. This figure gives a more detailed outline of the diagnostic application's design and shows the most important classes and their attributes, operations, and relations. In this diagram the class OLAP cube is composed out of (relations between) measures, dimensions, and dimension hierarchies. In this OLAP cube both the actual and (computed) reference data is present. The actor Business Analyst navigates to a certain Context Cube in the OLAP cube. In this symptom cube the Business Analyst can identify possible exceptional cells based on some Normative Model. An Exceptional Cell can be explained by

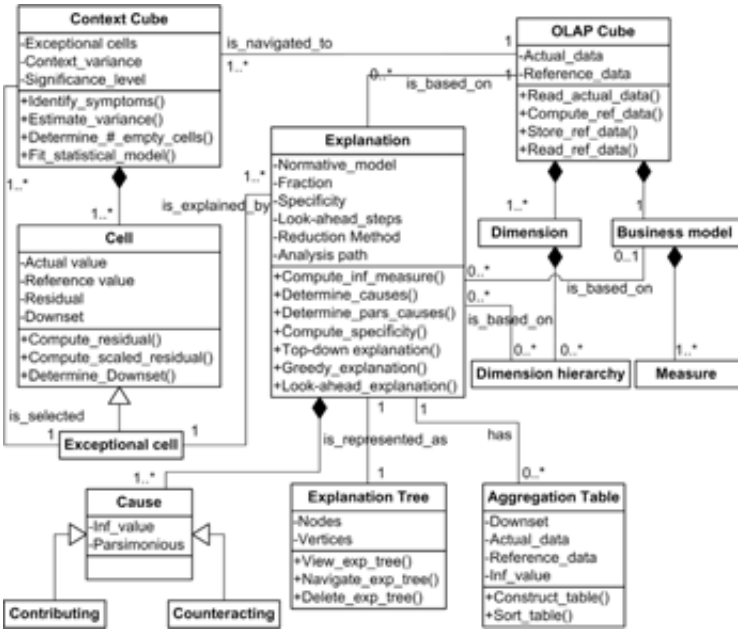


Figure 6.18: UML class diagram of the application for diagnosis.

the class Explanation. This class is composed out of Causes and based on actual and reference data from the OLAP Cube. The class Explanation contains the actual algorithms for explanation, i.e. implementation of Algorithms 4, 5, 7, and 6, described in Chapter 4. The class Explanation uses the class Aggregated Table, necessary for greedy explanation. The Explanation can be represented as an Explanation Tree to the Business Analyst.

The software itself is implemented in MS Excel, with pivot tables [2], and MS Access in combination with Visual Basic. The system architecture of the application has the following structure. The back-end of the application is the MS Access database, where the multi-dimensional model and OLAP source data is stored. In the MS Excel front-end, that connects with the database via an Open DataBase Connectivity

[2] A pivot table is a two-dimensional spreadsheet with associated subtotals and totals that supports viewing more complex data by nesting several dimensions and dimension levels on the x- or y-axis.

(ODBC) connection, the OLAP data can be represented as an OLAP cube with pivot tables. In MS Excel the diagnostic application can be configured for both exception identification and explanation via a number of graphical user interfaces. In Figures 6.19 and 6.20, a number of screenshots from the prototype software [3] are depicted for illustration. The tree-viewer interface of the application is depicted in Figure

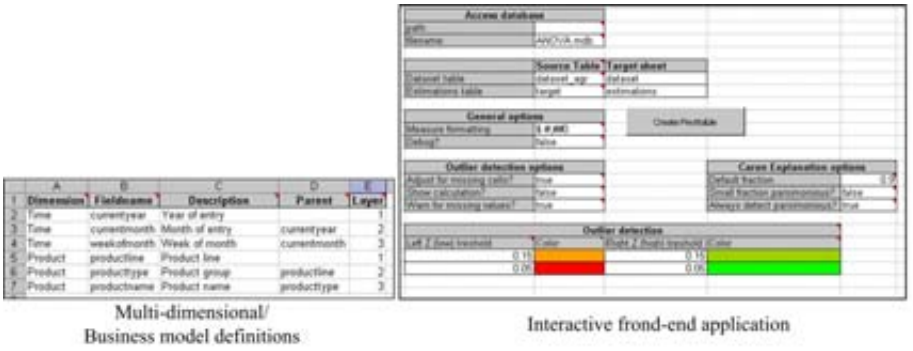


Figure 6.19: On the left hand side of the figure, the software to define the multi-dimensional/business model of the OLAP database is represented. On the right hand side of the figure, the software that enables customization of the diagnostic application is depicted.

6.21. In the viewer the whole explanatory graph can be made visible my manipulating the tree. In addition, the tree of causes for an exceptional cell is projected on the explanatory graph by highlighting parsimonious causes with a colour. By clicking on the cause under consideration, the details for the cause become visible in the right panel of the screen.

[3] We would like to thank Arjen Gideonse from SAP for his contribution to the implementation of the software for explanation of exceptional values in multi-dimensional databases.

Figure 6.20: On the left hand side of the figure, the GUI is shown that is used to select a specific cell for explanation in a context cube. On the right hand side of the figure, it is shown how the diagnostic module is integrated in the existing software, in this case MS Excel, to analyse OLAP data.

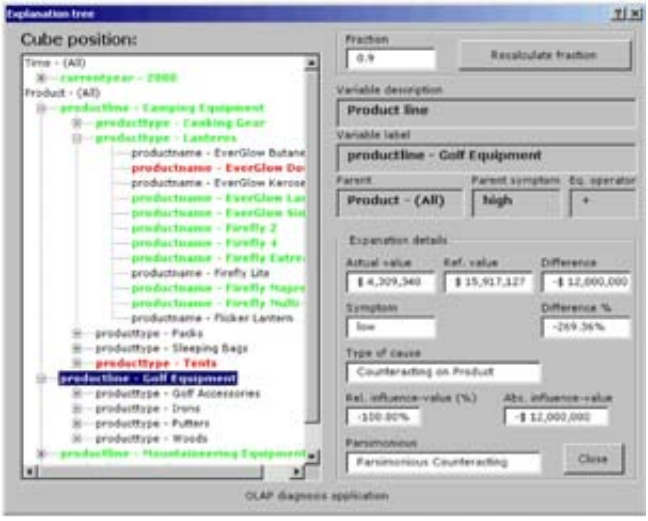


Figure 6.21: Tree viewer GUI of the OLAP diagnostic application. In the tree viewer the Time and Product dimension are visualized. The hierarchy of the Product dimension is unfolded. Parsimonious causes for a symptom are depicted in the hierarchy with a green or red color.

Although transaction databases can be very large, the kind of analysis discussed in this thesis is mostly performed on aggregated data, as in Sections 6.3 and 6.4. The method for explanations as described in this thesis is scalable in the software since all operations are linear in the number of records in an OLAP data cube. Notice that here the ANOVA models for computation of the reference values also have linear complexity. If other more complex statistical models are applied, such as complex time series models or neural networks with many parameters, the computational complexity may increase drastically. Another point of concern is the huge number of drill-down paths in OLAP if the number of dimensions and their depth increases. The full tree of explanations can have P paths (Equation 2.10). In this case the complexity is still linear in the size of the data set, but exponential in the number and depth of the dimensions. However, this can be resolved by applying the specificity heuristic (see RM_2) such that in each step only the most specific dimension is selected for explanation.

6.5 Case 3: Vehicle crime OLAP data

Here the methodology for explanation is used in a practical case study on vehicle crime data. The analyses for this study are performed with the prototypical diagnostic software (Section 6.4.4). The results of the study determine good threshold values for significance levels in problem identification and fractions in explanation generation.

The research for this case study was carried out as part of the project PROTECT (PROTECT 2006) and results were published in Caron and Veenstra (2007). This project aimed to contribute to the knowledge and insight that improve the performance of global supply chains in terms of their reliability. The study is performed on multi-dimensional vehicle criminality data obtained from the Dutch Foundation for Tackling Vehicle Crime (AVc 2006). The goal of the foundation is the reduction of vehicle crime (e.g. theft and fraud) in the Netherlands by means of prevention and by supporting public partners (e.g. police and insurance companies) in investigations. An important way to support the tackling of vehicle crimes is to perform analyses on vehicle crime data. At present, data analyses, mostly in the form of summary reports, statistics and trends, are used by the foundation for the detection and prevention of

vehicle crimes. If it is known, for example, how many cars and trucks are stolen in certain locations, police can patrol more often in these areas. The vehicle theft data set consists of records describing vehicle thefts in the period $1995 - 2004$. In the Netherlands approximately $30,000$ vehicles are stolen every year. The normalized data set consists of $295,291$ records in the fact table and five dimension tables describing hierarchies organized in a star schema, shown in Figure 6.22. In the dimension tables the numbers within brackets denote the cardinality of that level in the dimension hierarchy.

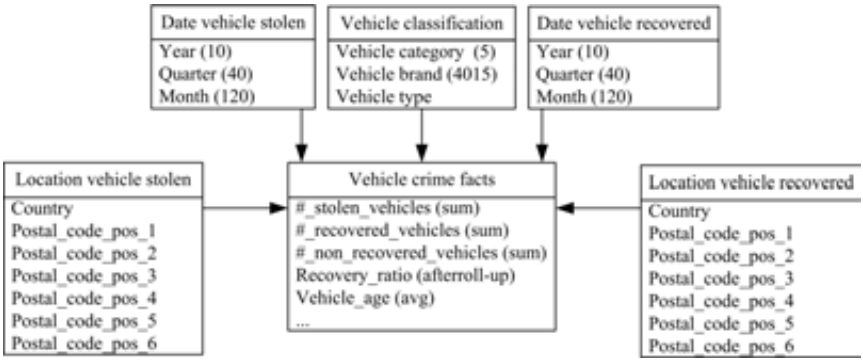


Figure 6.22: Star model with five dimension tables and a central fact table.

6.5.1 Exception identification

In this data set, we detect and explain exceptional values, such as a region with a relatively high number of vehicle thefts. Here we present an example of multi-level explanation of a symptom in the dimension "Location vehicle stolen" of the data cube under consideration. Suppose that an analyst starts exploring the context cube Year $\times$ Postal_code_pos_1 for the measure #_stolen_vehicles. A postal code in The Netherlands is composed of 4 digits and two alphabetic letters, for example, 1234 AB is a postal code. Where the first digit represents the most global location and the last letter character the most local location. An "?" indicates an aggregate at that

position of the postal code. The expected values for this context cube are computed with a main-effect ANOVA model, using Equation (3.3):

$$\hat{y}(\text{Year,Postal\_code\_pos\_1}) = \bar{y}(\text{Year}, +) + \bar{y}(+, \text{Postal\_code\_pos\_1}) - \bar{y}(+, +).$$

We write y for the measure #_stolen_vehicles. Here $\delta = 1.645$ is determined as the proper threshold value corresponding to a probability of .95 in the standard normal distribution. The cells in Figure 6.23 represent the scaled residuals, expressed as a percentage. The colored cells represent the identified symptoms for this context. For example, the program singles out the cell (2004,3???), because the standardized residual for the first position of the postal code "3???" in the year 2004 (= 2.7726) is larger than the threshold. Therefore, problem identification labels this cell as symptom ∂y^{12}(2004,3???)="high".

		First digit of post costal								
		1???	2???	3???	4???	5???	6???	7???	8???	9???
Years	2004	7.89%	0.78%	277.26%	-39.00%	-19.14%	32.81%	-82.57%	-84.73%	-93.31%
	2003	-29.17%	10.62%	247.31%	-21.64%	-39.85%	-6.09%	-55.68%	-52.01%	-53.49%
	2002	41.18%	1.11%	-132.62%	-7.84%	80.66%	39.39%	-2.95%	-8.80%	-10.14%
	2001	-7.58%	3.78%	-195.60%	21.12%	12.85%	-32.25%	68.35%	59.23%	70.11%
	2000	-12.32%	-16.30%	-196.35%	47.36%	-34.51%	-33.86%	72.85%	86.32%	86.83%

Figure 6.23: Computed exceptional values (in %) and identified symptoms in the context cube Year × Postal_code_pos_1.

6.5.2 Explanation generation

A full specification of the event to be explained is: $\langle y(2004,3???), \partial$#_stolen_vehicles ="high", $\hat{y}(2004,3???)\rangle$. Accordingly, the following business question is addressed:

"*Why are the number of stolen vehicles in cell* $(2004, 3???)$ *relatively high compared with the expected value for this cell in the context cube under consideration?*"

The algorithm is configured for one-step look-ahead. In addition, we omit insignificant influences from the explanations to prevent the human analyst from an information overload. After experimentation $T^+ = 0.80$ and $T^- = -0.80$ were determined as

appropriate fraction for the data set. We explain in the "Location vehicle stolen" dimension on the level "Postal_code_position_1". The increase in the number of stolen vehicles in the first digit of postal code region "3???" in The Netherlands is examined on the second digit level of the postal code. Hence the first corresponding equation used for explanation generation is (see Equation (2.12)):

$$y^{12}(2001, \text{Postal\_code\_pos\_1}) = \sum_{j=1}^{10} y^{13}(2001, \text{Postal\_code\_pos\_2}_j).$$

The reference values for this equation are determined in the context cube Year × Postal_code_pos_1 with the two-way ANOVA model:

$$\hat{y}(\text{Year}, \text{Postal\_code\_pos\_2}_j) = \bar{y}(\text{Year}, +) + \bar{y}(+, \text{Postal\_code\_pos\_2}_j) - \bar{y}(+, +).$$

Therefore, y^{12}(2001,3???) is the root of the explanation tree. The norm values

Table 6.11: Data for ∂y^{12}(2004,3???)="high".

	actual	reference	$\inf(y^{13}, y^{12})$
y(2004,3???)	7362	5411	
y(2004,30??)	3154	1465	1689
y(2004,31??)	585	349	236
y(2004,32??)	254	219	35
y(2004,33??)	799	674	125
y(2004,34??)	418	411	7
y(2004,35??)	937	950	-13
y(2004,36??)	187	208	-21
y(2004,37??)	396	416	-20
y(2004,38??)	376	425	-49
y(2004,39??)	256	294	-38

for explanation generation are based on the expected values for the entries of the dimension level Postal_code_pos_2 in the context Year × Postal_code_pos_2. Computation of the influences of the individual variables for the additive equation above with (4.4.1) yields the results in Table 6.11. From the data in this table it can be concluded that $\text{Cb}_p = \{y(2004,30??), y(2004,31??)\}$, since these two relatively large causes explain the desired fraction of $\inf(C^+, y(2004,3???))$. The set of parsimonious counteracting causes is given by $\text{Ca}_p = \{y(2000,36??), y(2000,37??), y(2000,38??),$

y(2000,39??)}. The parsimonious contributing causes are explained further on the levels Postal_code_pos_3, Postal_code_pos_4, etc. One-level explanations are com-

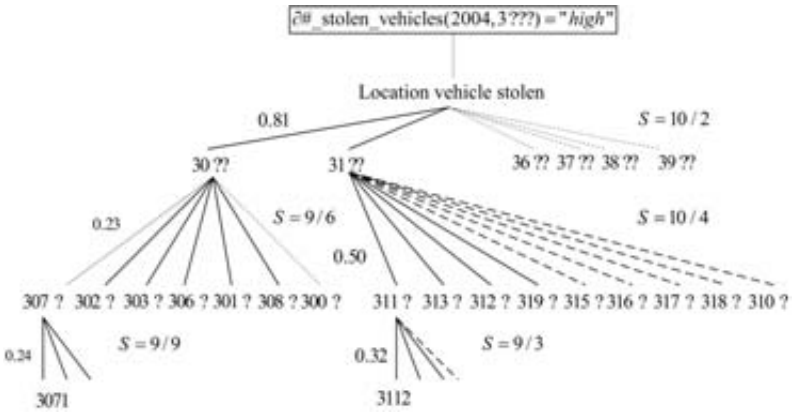


Figure 6.24: Diagnosis for ∂#_stolen_vehicles(2004,3???) ="high" for dimension "Location vehicle stolen".

bined to a complete explanation for the dimension "Location vehicle stolen". Figure 6.24 summarizes the results of the multi-level diagnosis in the form of an explanation tree. The lines indicate parsimonious contributing causes, the numbers on the lines indicate the relative values for the influence measures, and the ratios indicate the specificity value of the explanation step. The specificity values are determined using (4.22). For example, the explanation step on the Postal_code_pos_2 level is very specific for postal codes "30??" and "31??", because only 2 of the 10 possible causes are required here to explain the desired fraction. In summary, the explanation tree depicted in Figure 6.24 shows the analyst the set of regions, districts, streets, and part of streets, that are identified as the largest causes in the dimension "Location vehicle stolen". Moreover, similar explanation trees can be constructed automatically by the analyst for the hierarchies in the other dimensions. A comparison with model results and human analyses showed a large correspondence. In this way the analyst is assisted in processing and analysing large amounts of data.

6.6 Case 4: Supermarket OLAP sales data

6.6.1 Sensitivity analysis in a system of drill-down equations

In this section, we apply what-if analysis software on an artificial supermarket sales data in MS Access. The Foodmart data warehouse has 164, 558 records in the sales fact table for the years 1997 and 1998 for supermarkets in North America, with measures as sales, costs, revenues, units ordered, units shipped, total supply time, average supply time. Typical dimensions are:

- Time with the hierarchy: Month $\prec$ Quarter $\prec$ Year;
- Store Region with the hierarchy: Store Name $\prec$ Store City $\prec$ Store Region $\prec$ Store Country $\prec$ Store Type;
- Product with the hierarchy: Product Name $\prec$ Brand Name $\prec$ Product Sub-Category $\prec$ Product Category $\prec$ Product Department $\prec$ Product Family;
- Warehouse with the hierarchy: Name $\prec$ City $\prec$ State $\prec$ Country $\prec$ Type of warehouse;
- etc.

Here we describe a what-if analysis on the cube $C = 1998 \times$ All-Stores $\times$ Product Department $\times$ Type of Warehouse for the measure supply time (in days) denoted by y, aggregated with the average function. The data of the cube is depicted in Figure 6.25. In this case we analyse a change with some δ in the cell $c =$(1998, All-Stores, Alcohol Beverages, Large Independent) on its upset $\{\uparrow c\}$. The reference value of the cell is given by $y^r(c) = 76$ and the actual value is given by $y^a(c) = 76 + \delta$. The changes in $\{\uparrow c\}$ are computed by Equation (5.4) and given by:

- y^a(1998, All-Stores, Drink, Large Independent)= y^r(1998, All-Stores, Drink, Large Independent) + $\frac{1}{3}\delta$;
- y^a(1998, All-Stores, Alcohol Beverages, Warehouse)= y^r(1998, All-Stores, Alcohol Beverages, Warehouse) + $\frac{1}{6}\delta$;

AVG(Supply Time): in days	Large Independent	Large Owned	Medium Independent	Medium Owned	Small Indepdendent	Small Owned	Warehouse
Alcoholic Beverages	76+δ	81	90	53	82	94	476+1/6δ
Beverages	120	153	220	99	167	300	1059
Dairy	63	63	38	24	42	68	298
Drink	259+1/3δ	297	348	176	291	462	1833+1/18δ
Baked Goods	82	83	95	48	96	73	477
Baking Goods	251	276	221	138	235	313	1434
Breakfast Foods	41	30	59	16	30	74	250
Canned Foods	259	206	216	139	233	336	1389
Canned Products	22	34	15	15	22	21	129
Dairy	122	119	145	79	155	231	851
Deli	122	150	150	96	129	198	845
Eggs	66	52	54	35	45	82	334
Frozen Foods	250	341	343	168	318	491	1911
Meat	12	21	25	15	22	23	118
Produce	412	468	442	256	429	680	2687
Seafood	8	26	14	10	24	24	106
Snack Foods	339	393	339	235	335	457	2098
Snacks	53	115	86	28	87	128	497
Starchy Foods	60	62	66	32	65	82	369
Food	2099	2376	2272	1310	2225	3213	13495
Carousel	10	6	13	7	11	31	78
Checkout	11	9	27	21	17	30	115
Health and Hygiene	165	203	182	117	166	292	1125
Household	323	410	357	157	277	468	1992
Periodicals	53	46	41	17	74	73	304
Non-Consumable	562	674	620	319	545	894	3614
Product	2920+1/23δ	3347	3240	1805	3061	4569	18942+1/138δ

Figure 6.25: Sensitivity analysis in the cube 1998 × All-Stores × Product Department × Type of Warehouse for the average drill-down measure supply time (in days). Here the value of the cell 1998 × All-Stores × Alcohol Beverages × Large Independent is changed with δ and this change is propagated in the cell's upset. The changed cells are given a grey color in the cube.

- y^a(1998, All-Stores, Drink, Warehouse)= y^r(1998, All-Stores, Drink, Warehouse) + $\frac{1}{18}\delta$;
- y^a(1998, All-Stores, All-Products, Large Independent)= y^r(1998, All-Stores, All-Products, Large Independent) + $\frac{1}{23}\delta$;
- y^a(1998, All-Stores, All-Products, Warehouse)= y^r(1998, All-Stores, All-Products, Warehouse) + $\frac{1}{138}\delta$.

Now consider the following case. Suppose that we want to decrease y^r(1998, All-Stores, Drink, Large Independent) with one day by inducing a change in y^r(1998, All-Stores, Alcohol Beverages, Large Independent). This is done by inducing $\delta = -3$

then y^a(1998, All-Stores, Alcohol Beverages, Large Independent)= 73 and y^a(1998, All-Stores, Drink, Large Independent)= 258.

Sensitivity analysis in a system of business model equations is illustrated already in Section 5.3. In Section 5.3.2 a typical example is given related to what-if analysis in a system of business model and drill-down equations derived from a multi-dimensional financial database (see Example 2.1.1). In Section 5.3.3, the alternative method for what-if analysis is illustrated on the same database.

6.6.2 Software implementation

In this section, the most important concepts of the prototype software [4] for sensitivity analysis in a multi-dimensional database are discussed. This section is largely based on Caron and Daniels (2009) and Caron and Daniels (2010).

The prototype software is implemented in MS Excel and MS Access in combination with Visual Basic. The software connects with an OLAP database in MS Access with ODBC. In MS Excel a cube can be constructed from this database and inspected via pivot tables. In a pivot table, the analyst can do what-if analysis on a specific cell, by selecting the cell and pushing the analysis button.

If the analysis is started, the analyst can decide to change a cell in the pivot table with some percentage or absolute value, see the screenshots of the GUI in Figure 6.26. In the figure, the cell c =(2000.Q1, Mexico.Acapulco, Food) with the value $10,820.89$ for the measure store sales, in the context cube (2000.Quarter, Country.City, Food), is changed with, for example, 10% by the analyst. The result will be that the original cell value and its upset $\uparrow c$ are changed with that percentage. In the pivot table all changed cells in the upset are automatically indicated with a color (Figure 6.27). In the figure, for example, the parent of the cell c, the cell (2000.Q1, Mexico, Food), is changed from $85,520.91$ to $86,603,00$ in the what-if analysis.

After some actions the analyst can always return to the original situation because all operations are executed on a virtual copy of the multi-dimensional database. Obviously, in the software only the modified cell, in some cube in the lattice, and its

[4] We would like to thank the Wim Zuiderwijk and Arno van den Berg for their contributions to the implementation of the software for sensitivity analysis in the OLAP context.

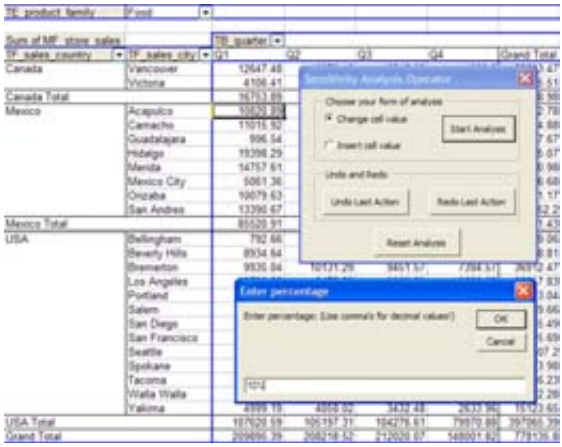


Figure 6.26: GUI for sensitivity analysis in MS Excel. A 10% cell increase in store sales(2000.Q1, Mexico.Acapulco, Food)= $10,820.89$ is analysed in the cube 2000.Quarter $\times$ Country.City $\times$. This change is automatically propagated in the cell's upset.

changed upset need to be stored for a single analysis.

6.7 Conclusion

In this chapter, we applied the concepts and techniques introduced in the previous chapters in a number of case studies. The case studies showed that diagnosis in the OLAP context has indeed useful business applications. We believe that the methodology put forward and applied here, can be effectively employed in a wide range of BI systems. Example applications are interfirm comparison Daniels and Caron (2009), sales analysis Caron and Daniels (2007), crime analysis Caron and Veenstra (2007), analysis of variance in accounting, and the generation of fish bone diagrams. The method can also be used in a continuous auditing framework, the expected values can be used as a benchmark and are compared with the actual values as described in this thesis. Larger deviations serve as a trigger for audit activities in which case the explanation method automatically generates important dimensions that can be

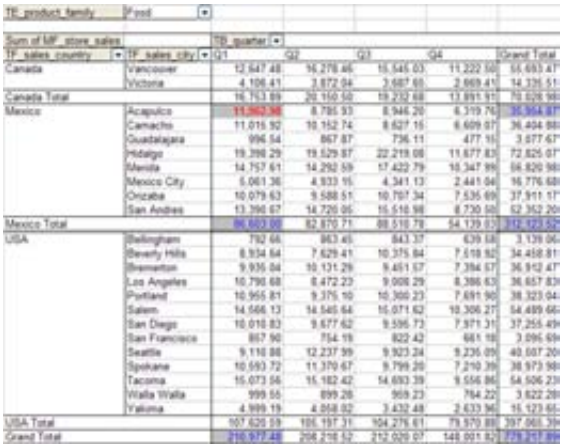


Figure 6.27: Result of the what-if analysis in the cube 2000.Quarter × Country.City × Food. Colors indicate the changed cells in the upset.

explored in further detail.

In Section 6.2 (Case 1), the look-ahead method for explanation is illustrated in a case study on interfirm comparison. In the case study with cancelling-out effects it is shown that the explanation method with the look-ahead procedure makes significant hidden causes visible that would have been missed by the explanation methodology of maximal explanation. In the implementation, special attention is given to presentation of the program output, where symptoms and causes are presented graphically as a tree of causes in GUI. In this manner, an analyst can view and access the results of the explanation process for diagnosis of company performance as a compact tree.

In Section 6.3 (Case 2a) and Section 6.4 (Case 2b), the top-down, the greedy, and the generic explanation method, are illustrated in a case study on the analysis of multidimensional financial data. In Section 6.3 (Case 2a), it is shown that exceptional cell values can be identified meaningfully in an OLAP cube with a statistical normative model. Subsequently, an exceptional cell is explained with top-down explanations over various paths in the cell's downset by applying a number of suitable ANOVA models.

In Section 6.4 (Case 2b), it is shown that exceptional cell values can be identified in an OLAP cube with a historic normative model. Subsequently, one of the exceptional cell values is explained with greedy explanation and we show that in this way hidden causes are identified, that might be missed in top-down explanation (without look-ahead). Finally, a generic explanation is given with greedy explanation for a range of exceptional cells in a context cube. It is also demonstrated with our software that our method is capable of assisting analysts in generating explanations for exceptional values in OLAP data. The results suggest that our method (semi-)automates the current manual discovery process of problem diagnosis in OLAP databases. The results also suggest that the explanation methodology could lead to improved managerial decision-making based on OLAP business data, because it can make causes visible, e.g. hidden causes, that might be missed in purely human analysis. Additionally, the results of this research can be used to develop professional diagnostic software that can be integrated in existing OLAP systems.

In Section 6.5 (Case 3), the explanation methodology is demonstrated in a case study describing the analysis of a real OLAP data set with vehicle criminality figures in The Netherlands. In this study it is shown that our method is capable of assisting analysts in generating explanations for exceptional values in multi-dimensional vehicle criminality data.

In Section 6.6 (Case 4), an extension of the OLAP framework for sensitivity analysis is illustrated in a compact case study on the analysis of supermarket sales data. By means of Theorem 5.2.1, we showed that there is an unique additive measure for each cube in the lattice. This is the basis for what-if analysis, where a change in some base cell in the lattice is propagated to all elements in its upset. We showed its working on a cube for the supermarket data set for an additive and an average measure. Sensitivity analysis in a system of business or mixed equations is discussed in Section 5.3. In such systems what-if analysis is only possible when specific conditions are satisfied. Finally a prototype software application for what-if analysis is described. This application is an additional tool for business analysts wanting to analyse their company data interactively. With this tool, they are able to 'play' with the OLAP data by doing sensitivity analyses.

Chapter 7

Summary of the main results

In this chapter, we give a summary of the main findings for each chapter. Multi-dimensional databases or OnLine Analytical Processing (OLAP) databases are a popular business intelligence technique in the field of business information systems for analytics and decision support. Multi-dimensional databases are capable of capturing the structure of business data in the form of multi-dimensional tables which are also known as data cubes. Manipulation and presentation of information through interactive multi-dimensional tables and graphical displays provide important support for the business decision-maker. The main goal of this dissertation is *"to extend the functionality of multi-dimensional business databases with diagnostic capabilities to support managerial decision-making"*. In this dissertation, the OLAP database is indeed extended with novel functionality for the detection of exceptional values, explanation generation, and sensitivity analysis. The purpose of the methods and algorithms presented here, is to provide OLAP databases with more powerful explanatory analytics and reporting functions.

In Chapter 1, a general introduction to the business intelligence framework and the position of the OLAP database in this framework is provided. This is followed by a description of diagnostic problem solving. After that the concepts of diagnosis are introduced within the OLAP domain and its potential use is illustrated.

In Chapter 2, important concepts related to the OLAP database and model are introduced and formalized. These concepts lay the foundation for our research objectives. We introduced a formal notation to express the internal structures of the OLAP database: dimensions, dimension hierarchies, full cubes, subcubes, base cube,

top cube, cells, and measures. The notation is coupled with navigational operators as roll-up and drill-down. The strength of the notation is that it can express both OLAP components as basic mathematical relations. Furthermore, we defined the lattice structure of all aggregation levels in the OLAP, obtained by aggregating a certain measure y over all its dimensions and hierarchies. In the lattice the concept of an analysis path is described. A path resembles the way a business analyst drills down or rolls up cubes in an analysis. It is shown mathematically that OLAP databases are often too large in practise to be analysed effectively, because of the large number of cell contexts and lattice analysis paths. They both grow exponentially fast when the number of dimensions and hierarchies increase. Finally, we use the notation to discuss drill-down equations for a single measure and relations between multiple measures. Drill-down equations are formed by the application of an aggregation function on a measure in the lattice from the base to the top. Relations between measures are called business model equations. The result of this application is a system of drill-down and/or business model equations. We can express both additive and non-additive measures in our notation. For additive measures we show that the system of drill-down equations is uniquely solvable.

In Chapter 3, a framework for the identification of exceptional values in OLAP databases is developed. This provides the OLAP analyst the possibility to identify regions of exceptions in an OLAP data cube during navigation. In this dissertation, an exceptional value is defined as a value that is surprisingly high or low in relation to the other values, and therefore of potential interest to the business analyst regardless of its cause. We use our notation to describe exception identification process in OLAP databases. Moreover, it is shown that both managerial and statistical normative models can be applied in OLAP databases as suitable reference classes. Appropriate managerial models are: planning and budget models, historical models, and extra/intra-organizational models. Two classes of appropriate statistical models are described, multi-way ANOVA models for continuous OLAP data and contingency table models for discrete OLAP data. It is found that for full-effects ANOVA models the mean-based estimates are directly available in the cubes of the lattice. This is explained by the concept of a complement cube. Finally, a general algorithm for exception identification in OLAP databases is proposed. In the case that this algorithm

is configured to be used with a multi-way ANOVA model, a specific algorithm for statistical model fitting can be applied to compute estimates for model coefficients.

Chapter 4 can be considered the core of this dissertation. In this part the OLAP database is extended with the functionality to give computerized explanations for surprising cell values. A general method that gives the OLAP analyst explanations for significant decreases or increases in business measures, identified at an aggregated level, is presented. This method for automated diagnosis is based on a generic explanation formalism. Explanation generation is supported by the two internal structures of the OLAP database: the dimension hierarchies and the business model. Therefore, explanation methods are developed for finding significant contributing and counteracting causes in these structures. The methods described are based on an influence measure, which can be considered to be a form of ceteris paribus reasoning. It is shown that a valid interpretation of the influence measure is only possible if the consistency and conjunctiveness constraint are satisfied. Moreover, it is shown that additive functions and non-additive differentiable functions (where the identified difference is relatively small) satisfy these constraints. The following methods for automated explanation are presented: look-ahead, top-down, and greedy explanation. Look-ahead explanation deals with the problem of potential cancelling-out effects. The method enhances the original method of maximal explanation with look-ahead functionality to detect hidden causes. The method is based on function substitution. Explanation generation in this method is continued until a contributing cause cannot be explained further. The result of the method is an explanation tree, where the main causes for a surprising value are presented to the analyst. In addition, a top-down approach for explanation in systems with both OLAP drill-down and business model equations, and a greedy approach for explanation in systems that consist purely of drill-down equations, are demonstrated. The greedy explanation method uses the transitivity property which simplifies the explanation generation process. Furthermore, to prevent an information overload to the analyst, several techniques are created to prune the explanation tree. Appropriate pruning methods are: the identification of parsimonious causes, the selection of specific causes over general causes, the application of heuristics that reduce the number of equations considered, the selection of only large causes, and the reporting of generic explanations. Finally, to guarantee the correct

working of the explanation methods the consistency constraint has to be satisfied. Basically, reference values are consistent if they satisfy the same equation as is given for the actual values. For the various normative models we show under what conditions the consistency constraint is satisfied. In particular, we proof that a special class of additive ANOVA models produces consistent reference values, as opposed to the general class of statistical models that do not produce such values.

In Chapter 5, the theoretical underpinnings under which sensitivity analysis is valid in OLAP databases are dealt with. In this dissertation, sensitivity analysis is considered to be the reverse of explanation generation in diagnostic reasoning. Our exposition differentiates between sensitivity analysis in systems of purely drill-down equation and mixed systems of equations with also business model equations. It is proven that there is an unique additive drill-down measure defined on all cubes of the aggregation lattice. This proof is the basis for sensitivity analysis in OLAP databases, where a change in some base cell in the lattice is propagated to all descendants in its upset. For sensitivity analysis in mixed systems of equations a matrix notation is presented and the conditions for solvability are discussed. Due to the fact that such systems are typically overdetermined in OLAP databases, the implicit function theorem cannot be applied. Therefore, we proposed a method to reduce the number of equations in the system and apply the implicit function theorem on a subsystem of the original system. We conclude with an alternative method for what-if analysis in mixed systems of equations.

In Chapter 6, it is shown that our methodology has a wide range of business applications, such as variance analysis in accounting, competition benchmarking, analysis of sales and financial data, and the analysis of any other data that possess a multi-dimensional hierarchical structure. The methodology is demonstrated in several case studies. In Case 1, the applicability of the look-ahead method for explanation is illustrated in a study on interfirm comparison. In this study it is shown that the explanation method with the look-ahead procedure makes significant hidden causes visible that would have been missed by the maximal explanation methodology In the software implementation of the method, special attention is given to presentation of the program output, where symptoms and causes are presented graphically as a tree of causes in the GUI. In this manner, an analyst can view and access the results

of the explanation process for diagnosis of company performance as a compact tree. The top-down, the greedy, and the generic explanation method, are illustrated in a case study on the multi-dimensional analysis of financial data. In the study, first exceptional cells are identified with various normative models and after that these cells are explained with the explanation methods. In Case 2a, it is shown that exceptional cell values can be identified meaningfully in an OLAP cube with a statistical normative model. Subsequently, an exceptional cell is explained with top-down explanations over various analysis paths in the cell's downset by use of a number of suitable ANOVA models. In Case 2b, it is demonstrated that exceptional cell values can be identified in an OLAP cube with a historic normative model. Subsequently, one of the exceptional cell values is explained with greedy explanation and it is shown that in this way hidden causes are identified, that might be missed in top-down explanation (without look-ahead). The results from the case studies suggest that our method (partly) automates the current manual discovery process of problem diagnosis in OLAP databases. This is clearly an advantage because the human analysis of OLAP data can get tedious and error-prone, especially when the data set is large. Our explanation methodology could lead to improved managerial decision-making based on OLAP business data. Moreover, in Case 3, it is shown that our method is capable of assisting analysts in generating explanations for exceptional values in multi-dimensional vehicle criminality data. Finally, in Case 4, an extension of the OLAP framework for sensitivity analysis is illustrated in a compact case study on the analysis of supermarket sales data. We demonstrate what-if analysis on a cube for the supermarket data set for an additive and an average measure. Sensitivity analysis in a system of business or mixed equations is discussed in Chapter 5, where also an illustative example is presented. In such systems what-if analysis is only possible when specific conditions are satisfied.

The dissertation concludes with a number of appendices that give background information. In Appendix A, an overview is given of computer-based diagnosis in various application domains. In Appendix D, the mathematics in matrix notation are given for a system of implicit drill-down equations. As far as we know, this has never been pointed out in the existing literature. Subsequently, the notation is used systems of drill-down equations to prove solvability and uniqueness of solutions.

Nederlandse Samenvatting (Summary in Dutch)

Dit proefschrift gaat over het verklaren van exceptionele waarden in multi-dimensionele of OnLine Analytical Processing (OLAP) bedrijfsdatabanken. OLAP-databases zijn een populaire business intelligence techniek op het gebied van bedrijfsinformatiesystemen. De parapluterm 'business intelligence' staat voor combinaties van methoden, processen, technieken en toepassingen, welke nodig zijn om ruwe bedrijfsdata om te vormen tot bruikbare informatie en kennis in bedrijven en organisaties. OLAP-databases representeren data in de vorm van datakubussen. Het exploreren van multi-dimensionele gegevens in deze kubussen door een analist is relatief eenvoudig omdat de software gebruik maakt van interactieve operatoren en grafische displays. Het doel van dit proefschrift is om diagnostische functionaliteit voor het maken van verklarende analyses in OLAP-databases te brengen en om zo te komen tot betere bestuurlijke beslissingsondersteuning. De functionaliteit wordt uitgebreid met automatische methoden voor de detectie en het verklaren van exceptionele waarden en gevoeligheidsanalyse.

In hoofdstuk 1 wordt een algemene introductie tot het business intelligence raamwerk gegeven en de positie van OLAP binnen dit raamwerk wordt uitgelegd. Daarna volgt een uitleg van de belangrijkste concepten van automatische diagnose. Vervolgens worden deze concepten geïntroduceerd in het domein van OLAP-databases, en het mogelijke nut daarvan voor de bedrijfsanalist wordt geïllustreerd.

In hoofdstuk 2 worden de belangrijkste concepten van het OLAP-database model besproken en geformaliseerd. Deze concepten leggen de basis voor onze onderzoeksdoelstelling. Een formele notatie om de interne structuren van een OLAP-database uit te drukken wordt voorgesteld. In deze notatie kunnen OLAP concepten zoals:

dimensies, dimensiehiërarchieën, volledige kubussen, deelkubussen, basiskubus, topkubus, cellen, en meetwaarden eenvoudig worden beschreven. Daarnaast kunnen ook navigatie-operatoren binnen OLAP, zoals 'roll-up' en 'drill-down' in de notatie beschreven worden. De kracht van de notatie is gelegen in het feit dat de notatie zowel overweg kan met OLAP concepten als met wiskundige relaties. Vervolgens wordt er een roosterstructuur voor alle aggregatieniveaus in een OLAP-database gedefinieerd. Dit rooster wordt verkregen door een bepaalde meetwaarde y over alle mogelijke dimensies en hiërarchieën te aggregeren. In het rooster wordt het idee van een analyse pad uitgelegd. Een analyse pad geeft aan hoe een analist de operatoren 'drill-down' en 'roll-up' kan gebruiken in een analyse. Wiskundig wordt uitgelegd dat OLAP-databases in de praktijk vaak te groot zijn om goed te worden geanalyseerd. Dit komt omdat een cel in een kubus in veel contexten kan worden bekeken en dat er vanuit een cel veel mogelijke analysepaden zijn. Beide groeien exponentieel als het aantal dimensies en hiërarchieën toeneemt. Als laatste gebruiken we de notatie om 'drill-down' vergelijkingen voor een enkele meetwaarde en relaties tussen meetwaarden te bespreken. De 'drill-down' vergelijkingen worden gevormd door het toepassen van een aggregatiefunctie op een meetwaarde vanaf de basis van het rooster tot de top. Het resultaat van deze toepassing is een systeem van 'drill-down' vergelijkingen. Relaties tussen meetwaarden worden bedrijfsmodelvergelijkingen genoemd. Zowel additieve als niet-additieve meetwaarden kunnen worden beschreven. Voor additieve meetwaarden tonen we aan dat een systeem van drill-down meetwaarden uniek oplosbaar is.

In hoofdstuk 3 wordt een raamwerk voor het bepalen van exceptionele waarden in een OLAP-database ontwikkeld. Dit raamwerk geeft de OLAP analist de mogelijkheid om regio's van opvallende waarden op te sporen tijdens een analyse. Deze opvallende waarden kunnen mogelijk wijzen op nieuwe bedrijfskansen of naar specifieke problemen. Een exceptionele waarde wordt gezien als een verrassend hoge of lage waarde voor een cel ten opzichte van andere cellen in de kubus. Er wordt aangenomen dat de verrassende celwaarde interessant is voor de analist onafhankelijk van de mogelijke oorzaak. In dit hoofdstuk gebruiken we notatie van het vorige hoofdstuk om het proces van opsporen van exceptionele cellen in detail te beschrijven. Vervolgens wordt aangetoond dat zowel bestuurlijke- als statistische modellen

toegepast kunnen worden als geschikte referentieklassen binnen de OLAP context. Geschikte bestuurlijk modellen zijn: planning- en budgetmodellen, historische modellen en extra/inter-organisatiemodellen. Twee klassen van geschikte statistische modellen worden in detail beschreven, dit zijn 'multi-way ANOVA' modellen voor continue OLAP data en 'contingency table' modellen voor discrete OLAP data. Het blijkt dat voor 'full-effects' ANOVA modellen de schatters, welke op gemiddelden zijn gebaseerd, direct beschikbaar zijn in de kubussen van het rooster. In een voorbeeld wordt aangetoond dat verschillende ANOVA modellen verschillende sets van uitzonderlijke waarden kunnen geven. Tot slot wordt een algemeen algoritme voor exceptie identificatie in OLAP-databases voorgesteld. In het geval dat dit algoritme is geconfigureerd voor gebruik met een multi-way ANOVA model, kan een specifiek algoritme worden toegepast om de modelcoëfficiënten berekenen.

Hoofdstuk 4 kan worden beschouwd als de kern van dit proefschrift. In dit deel wordt de OLAP-database daadwerkelijk uitgebreid met de functionaliteit om geautomatiseerd verklaringen voor opvallende celwaarden te geven. Een algemene methode, die de OLAP-analist verklaringen geeft voor significante dalingen of stijgingen in meetwaarden op een geaggregeerd niveau, wordt voorgesteld. Deze methode voor automatische diagnose is gebaseerd op een algemeen verklaringsformalisme. Het geven van verklaringen wordt ondersteund door twee interne structuren van de OLAP-database: de dimensiehiërarchieën en het bedrijfsmodel. Er worden specifieke verklaringsmethoden voorgesteld voor het vinden van bijdragende en tegengestelde oorzaken in deze structuren. Al deze methoden zijn gebaseerd op een 'maat van invloed'. Deze maatstaf kan worden beschouwd als een vorm van ceteris paribus redeneren. In het proefschrift wordt aangetoond dat een geldige interpretatie van deze maat van invloed alleen mogelijk is als aan restricties voor consistentheid en conjunctie is voldaan. Zowel additieve als niet-additieve differentieerbare functies (onder de voorwaarde dat het geïdentificeerde verschil relatief klein is) voldoen aan deze restricties. De volgende methoden voor het genereren van verklaringen en hun eigenschappen worden gepresenteerd: 'look-ahead', 'top-down' en 'greedy'. De look-ahead methode behandelt het probleem van elkaar opheffende effecten. Deze methode verbetert de oorspronkelijke methode van maximale verklaringen door dieper te kijken in het bedrijfsmodel om zo mogelijke verborgen oorzaken te vinden. De 'look-ahead' methode is gebaseerd op

functie-substitutie. Het geven van verklaringen door deze methode wordt voortgezet totdat een bijdragende oorzaak niet verder kan worden ontwikkeld. Het resultaat van de verklaringsmethode is een verklaringsboom, waarin de belangrijkste oorzaken voor een exceptionele waarde worden gepresenteerd aan de analist. De 'top-down' verklaringsmethode kan gebruikt worden in systemen met zowel 'drill-down' als bedrijfsvergelijkingen. De 'greedy' verklaringsmethode werkt in systemen met alleen 'drill-down' vergelijkingen. Deze methode maakt gebruik van een transitiviteitseigenschap, welke het genereren van verklaringen eenvoudiger maakt. Om de analist te beschermen tegen teveel informatie afkomstig uit de verklaringsmethode, wordt een aantal technieken voorgesteld om de verklaringsboom te snoeien, om zo alleen de belangrijkste oorzaken aan de analist aan te bieden. Geschikte snoeimethoden zijn: het vaststellen van de significante oorzaken, de selectie van specifieke oorzaken boven algemene oorzaken, het toepassen van heuristieken welke het aantal vergelijkingen in de analyse vermindert, de selectie van alleen grote oorzaken en het rapporteren van generieke verklaringen. Om de correcte werking van de verklaringsmethoden te garanderen moet aan de consistentie-eis voldaan zijn. Referentie waarden zijn consistent als ze aan de dezelfde vergelijking voldoen als gegeven voor de actuele waarden. Voor de besproken normatieve modellen laten we zien onder welke condities ze voldoen aan deze eis. In het bijzonder tonen we aan dat onder bepaalde voorwaarden additieve ANOVA modellen consistente ketens referentie waarden geven. In het algemeen produceren statistische modellen geen consistente ketens van referentiewaarden.

In hoofdstuk 5 wordt de theoretische basis voor gevoeligheidsanalyse in OLAP-databases besproken. Hier wordt gevoeligheidsanalyse beschouwd als het omgekeerde van het genereren van een verklaring in de context van diagnostisch redeneren. Gevoeligheidsanalyse wordt besproken in stelsels met alleen drill-down vergelijkingen en in stelsels met zowel drill-down als bedrijfsmodelvergelijkingen. In dit hoofdstuk wordt bewezen dat er een unieke additieve drill-down measure kan worden gedefinieerd op alle kubussen van het rooster. Dit bewijs is de basis voor gevoeligheidsanalyse in OLAP-databases, waar een verandering in een basiscel wordt gepropageerd naar al zijn afstammelingen in de bovenliggende structuur. Voor gevoeligheidsanalyse in gemengde stelsels van vergelijkingen wordt een matrixnotatie voorgesteld. In deze notatie worden de voorwaarden voor oplosbaarheid van deze stelsels besproken. Deze

stelsels zijn doorgaans overgedetermineerd zodat de impliciete functie stelling niet kan worden toegepast. Daarom wordt een methode voorgesteld om het aantal vergelijkingen in het systeem te verminderen en de impliciete functie stelling toe te passen op een subsysteem van het oorspronkelijke systeem. Het hoofdstuk wordt afgesloten met een voorstel voor een alternatieve methode voor gevoeligheidsanalyse in gemengde stelsels.

In hoofdstuk 6 wordt getoond dat onze methodologie een breed scala van toepassingen heeft, zoals variantie-analyse in accountancy, het vergelijken van de prestaties van ondernemingen, de analyse van verkoopdata en financiële gegevens, en de analyse van alle andere gegevens die beschikken een multi-dimensionele hiërarchische structuur. De methodologie wordt toegepast in een aantal casestudy's. In case study 1 wordt de 'look-ahead' methode gebruikt in een studie waarin bedrijven met elkaar worden vergeleken om inzicht te geven in hun prestaties. De gegevens zijn afkomstig uit de productiestatistieken van het Centaal Bureau voor de Statistiek. In deze studie wordt aangetoond dat de 'look-ahead' methode, significante verborgen oorzaken zichtbaar kan maken, welke gemist zouden zijn door de klassieke methode van maximaal verklaren. In de software-implementatie van de methode, wordt speciale aandacht besteed aan de presentatie van het programma-uitvoer, waar de oorzaken van opvallende waarden worden voorgesteld als een boom van oorzaken in de grafische gebruikersinterface. Op deze wijze, kan een analist de resultaten van het verklaringsproces, ten behoeve van het maken van bedrijfsvergelijkingen, eenvoudig bekijken in de vorm van een compacte boom. De 'top-down', de 'greedy' en de generieke verklaringsmethode, worden geïllustreerd in een case study over de multi-dimensionele analyse van financiële gegevens. In de studie, worden eerst opvallende cellen geïdentificeerd met gebruik van normatieve modellen, daarna worden deze opvallende cellen verklaard met de verschillende verklaringsmethoden. In case study 2a wordt aangetoond dat met een statistisch model op een zinvolle manier opvallende cellen kunnen worden opgespoord in een datakubus. Vervolgens worden verklaringen gegeven voor een opvallende celwaarde met de 'top-down' methode, configureerd met ANOVA modellen als normatief model, toegepast op verschillende analysepaden van nakomelingen van de betreffende cel. In case study 2b wordt gedemonstreerd hoe exceptionele celwaarden kunnen worden bepaald in een OLAP kubus met een historisch normatief model.

Daarna wordt één van de opvallende celwaarden verklaard met de 'greedy' methode en wordt getoond dat op deze manier verborgen oorzaken zichtbaar gemaakt kunnen worden, welke gemist zouden zijn door het puur toepassen van de 'top-down' methode (zonder 'look-ahead'). De resultaten van de case studies suggereren dat onze methode het huidige handmatige 'verklarende analyseproces' van datakubussen deels kan automatiseren. Dit is duidelijk een voordeel, omdat de handmatige analyse van OLAP-gegevens met het oog veel tijd kost en foutgevoelig is, vooral als de datakubus erg groot is. Het praktische gebruik van de verklaringsmethode zou kunnen leiden tot betere bestuurlijke besluitvorming op basis van OLAP gegevens.

Het proefschrift sluit af met een aantal bijlagen die achtergrondinformatie geven. In appendix A wordt een overzicht gegeven van computergebaseerde diagnose in verschillende toepassingsdomeinen. In appendix D, wordt de wiskunde in matrixnotatie gegeven voor een systeem van impliciete 'drill-down' vergelijkingen. Voor zover wij weten, is zo'n notatie nooit eerder naar voren gebracht in de literatuur. Vervolgens wordt de notatie gebruikt om te bewijzen dat systemen van zulke vergelijkingen een unieke oplossing hebben.

Appendices

Appendix A

Overview of computer-based diagnosis

The field of AI research has paid much attention to the formalisation and automation of diagnostic reasoning for decision support in the past (Console and Torasso 1989; de Kleer and Williams 1992; de Kleer et al. 1992; Lucas 1997; Reiter 1987). The main aspects in computer-based diagnosis are (Verkooijen 1993):

- How is the "understanding" of the system formalized and represented? (the *knowledge representation formalism*);
- What diagnostic reasoning methods are applied to explain discrepancies given a specified domain formalisation? (the *diagnostic reasoning method*).

Obviously, these two aspects are connected; a diagnostic reasoning method cannot be chosen independently of the knowledge representation formalism.

The type of knowledge representation formalisms for the underlying system determine the main classification of the kinds of reasoning methods applied. A classification that is often made is the distinction between *rule-based diagnostic systems* and *model-based diagnostic systems*. Many researchers have discussed the differences between these systems (Console and Torasso 1989; Davis and Hamscher 1988). For diagnosis in the domain of multi-dimensional databases only model-based diagnosis is relevant because of the type of knowledge that is available in the database structures.

Model-based diagnosis became a research topic in AI research around 1980 as an attempt to address shortcomings (e.g. the knowledge elicitation problem, unsatisfactory explanation capabilities, and brittle problem solving) of the contemporary generation of rule-based diagnostic systems. Therefore, a new type of diagnosis system evolved that did not include heuristics about symptoms and diagnosis, but relied on basic knowledge of the domain. A diagnosis system of this type is called model-based, because it has an explicit model of the diagnosis system, which describes the system (de Kleer and Williams 1992). The term *deep knowledge* is used for knowledge that describes characteristic aspects of the system at a certain level of abstraction. Among them are models of structural, topographical, functional, or behaviourial system features. An important assumption in model-based diagnosis is that "shallow" expert rules usually turn out to be specialized pre-compiled statements that are, in fact, derivable from the underlying theory (Feelders 1993). In model-based diagnosis, however, the underlying theory is explicitly modelled in the program (Apte and Hong 1986; Chandrasekaran and Mittal 1984). In general, this approach leads to better explanation capabilities and more robust problem solving behaviour. A model-based diagnosis system usually applies a general algorithm to find a diagnosis.

Two important properties tend to characterize model-based reasoning according to Hamscher (1992): an emphasis on categorical knowledge and separation of domain knowledge from problem-solving knowledge. The content of the knowledge base in a model-based system tends to concern categorical causes and effects rather than probabilistic associations among problem features. Model-based systems aspire to use general-purpose models. They achieve an even stricter separation between *what* is known from *how* the knowledge will be used, than the comparable situation found in typical rule-based systems. These properties mean that most research in this area is grounded in physics and medicine, and thereby inherits three common attributes. First, systems that consist of decomposable structures and constrained interactions between the elements of those structures. Second, the diagnosis task to be performed reduces to making and evaluating predictions about the evolution of the aggregate states of such decomposable structures. Third, predictions are made for a virtually closed system whose initial state and all relevant subsequent exogenous influences are

knowledgeable. Although the literature contains exceptions, these properties characterize the majority of the research (Weld and de Kleer 1990; Williams and de Kleer 1991).

The ability to model is a condition for the application of model-based diagnosis methods. Typically, these methods makes use of the following types of models representing structural or causal information of the system (Feelders 1993):

- a *structural model* contains information about the structure and correct functional behaviour of the system's components and its interactions;
- a *causal model* consists of cause-effect relationships between elements that are important for the description of the system's behaviour.

The formal theories of computer-based diagnosis and explanation, described in the previous parts, have no inherent relationships to certain application domains. AI research on diagnostic reasoning has almost exclusively been concerned with the medical and the physical domain, and scarcely to the domain of business and management. An extensive comparison is available regarding these three traditional application domains (Courtney et al. 1987; Feelders 1993; Verkooijen 1993). The objective of this thesis is to provide a formalisation of diagnostic problem-solving in the relatively new application domain of multi-dimensional business databases. For this purpose the main characteristics of diagnostic problem-solving in the traditional application domains are discussed briefly in order to create points of comparison with diagnosis in multi-dimensional databases. Diagnosis in the domain of business and management is discussed in Section 4.8. The results from the comparison serve as a basis for knowledge representation and diagnostics in multi-dimensional databases.

A.1 Diagnosis in the physical domain

Consistent with the general diagnosis task in figure 1.2, the model-based approach to diagnosis of a technical device (e.g. electronic circuit, car, DVD-player, etc.) is based on a comparison of an incorrect device and a representation of an ideal (correct) device. The actual behaviour of the device is typically observed by means

of input/output measurement. In the domain of diagnosis of technical devices, there are logical theories of diagnosis. These theories assume the availability of a logical description (model) of the structure of the system to be diagnosed, and of the normal behaviour and interactions of its components. The model of the device can make predictions about its intended behaviour. Discrepancies between observation and prediction are due to a defect in the device. The diagnostic objective is to locate defective components in a way that explains the discrepancies. The usual therapy is to replace the defective component. Three well-know diagnostic systems for diagnosis in the physical domain are Sherlock and GDE by de Kleer and Williams (1989) and de Kleer and Williams (1992), and DART by Genesereth (1984).

A.2 Diagnosis in the medical domain

Discovering what is wrong in a patient with particular symptoms and signs, i.e. diagnosis, is the usual starting point in a medical decision process. It is, therefore, not surprising that automated medical diagnosis was one of the first research fields of AI (Lucas 1997). In general, the impact of rule-based diagnostic systems on AI has been large. Examples are MYCIN (Shortliffe 1976) for the diagnosis and treatment of bacterial infections, and INTERNIST-1 (Miller et al. 1982) as an expert consultant program for diagnosis in general internal medicine.

In line with the general diagnosis task, in medical diagnosis the underlying system is the human body or some specific part of it. Medical knowledge about this system is incomplete, although particular subsystems may be well understood. Normal behaviour is often not precisely defined. In general, one makes use of the behaviour which is observed most frequently in practice, a kind of average behaviour. The presence of a particular disease usually serves as an explanation for the set of observed symptoms. Directly after the diagnosis hypothesis the therapy is started, based on this hypothesis, and the appropriate medical treatment is given.

A.3 Comparison and evaluation

Managerial problem diagnosis (Bouwman 1983) differs from diagnosis in other domains because the managerial problem domain is not as structured as many other problem domains. As opposed to problem domains such as engineering, mathematics, or electrical circuit design, the managerial problem domain is not governed by well formulated relationships. Although specific areas of management may be well structured, such as, sales, accountancy, and financial models. Furthermore, in automated business diagnosis the system, in contrast to the previous systems, is not tangible, in the sense of physical components (Verkooijen 1993). For example, the financial statements are an abstraction of the underlying financial process. The financial items do not have a prescribed functional behaviour as, for example, the components of a technical device, or the organs of the human body. They represent the input they receive from other financial items or from the financial environment.

In the domain of business the application of diagnosis from a structural model is nearly impossible. For example, in the financial domain, it would be far too complex to describe the structure and behaviour of the system in the form required by first principle approaches. Therefore, as in Feelders and Daniels (2001), a causal view of explanation is taken that is able to deal with quantitative and qualitative phenomena that pervade the domain of business, finance, and management. This causal model should capture the underlying cause-effect relations of the managerial problem domain. In fact a similar approach is taken as Courtney et al. (1987). It describes a managerial diagnosis system based on a causal model in terms of economic variables and their influence relations.

In the previous two sections and Section 4.8, we have reviewed the problem of automated diagnosis and explanation in several domains. Although in all domains the global idea of diagnosis is the same, i.e. explaining unexpected behaviour of a system, it has become clear that each domain has its own characteristics. A characteristic that all systems have in common is that they are model-based. From a reasoning viewpoint Feelders notices two major differences (Feelders 1993). Firstly, a situation of incomplete information is presupposed in technical and medical diagnosis, whereas this is generally not the case in diagnosis in the business domain. In the latter domain

there is usually complete information about actual and norm values of variables in the business model. And the problem of diagnosis is reduced to *selecting* relevant influences from the available information. This problem is addressed in the work of Kosy and Wise (1984), Kosy (1989), Courtney et al. (1987), Mohammed et al. (1988), Feelders (1993), and Feelders and Daniels (2001). If not all actual and norm values are known, the problem of diagnosis is one of finding consistent hypotheses. Secondly, the difference between the business domain on the one hand and the medical and technical domain on the other hand is that their objects of comparison do not change over time, whereas in the business domain the proper object of comparison constantly changes. It entails that a company's performance may be considered satisfactory this year, whereas the same performance is considered mediocre for the next year, simply because the object of comparison has changed, due to macro-economic developments (Feelders 1993). In addition, in the business domain often multiple objects of comparison are applied at the same time because they are all important for managerial decision-making. For example, a company's performance of the current year is compared to its performance in the previous year and at the same time the company's performance is benchmarked against its competitors.

The objective for describing the diagnostic process in such different domains was two-fold. Firstly, to position and compare diagnosis in the domain of business and management between the more traditional fields of diagnosis. We agree with Feelders (1993) that it difficult to transfer reasoning and knowledge representation from the traditional fields of diagnosis to the domain of DSS's for business and management. Secondly, to introduce the concept of diagnosis in a special class of DSS's, namely OLAP databases. In this thesis we show that OLAP databases are an appropriate domain for model-based diagnosis. Furthermore, it is shown that diagnosis in these databases requires a knowledge representation formalism and diagnostic reasoning methods that resembles diagnosis in quantitative financial models. In conclusion, we summarize in Table A.1 the main characteristics of model-based diagnosis in the different application domains.

Table A.1: Model-based diagnosis in four application domains

	Technical	Medical	Business and management	Multi-dimensional databases
Diagnosis objective	Locate minimal set of faulty components	Find minimal set of causes (the disease)	Explain deviating behaviour of financial indicators	Explain deviating cells in data cube
Actual system	Description of the internal structure and behaviour of the device	Causal model the human body with medical knowledge	Financial statements or processes of the firm	Multidimensional database with measures, dimensions, and hierarchies
Example system	DVD-player, car, computer, robot, etc.	Model of the heart, lungs, etc.	Sales model, income statements, balance sheets, etc.	Multidimensional sales data, socio-economic data, etc.
System understanding (normative model)	Representation of an ideal (correct) device	The healthy human organism	Historical, planning, inter- and intra-organizational models	Managerial normative models and various statistical models
Example symptom	Defective component	Patient has high temperature and fever	Net sales of firm have gone down compared with last year	Exceptional cell value in context of sales data cube
Knowledge representation	Structural	Causal	Causal	Structural and Causal
Diagnostic systems	DART (Genesereth 1984), Sherlock (de Kleer and Williams 1989), and GDE (de Kleer and Williams 1992)	MYCIN (rule-based) (Shortliffe 1976) and INKBLOT (Citro et al. 1997)	DSS's for business diagnosis (Courtney et al. 1987; Hamscher 1994; Feelders and Daniels 2001; Daniels and Caron 2009)	iCube and iDiff (Sarawagi et al. 1998; Sarawagi 2001), and OLAP explanatory analytics (Caron and Daniels 2007; Caron and Daniels 2013)

Appendix B

Model and data for case study 1

B.1 Data for interfirm comparison

The meaning of the variables for interfirm comparison at Statistics Netherlands are described in this Appendix in detail. The variable descriptions in English have been translated from the original Dutch surveys. In addition, in Table B.1 the complete data set for interfirm comparison of the ABC-company is given.

Result variables:
r_1: total result before taxation
r_2: total operating results
r_3: total financial results
r_4: total results allowances
r_5: total extraordinary results
r_6: total operating revenues
r_7: total operating costs
r_8: financial revenues
r_9: financial expenses
r_{10}: additions to allowances
r_{11}: deductions from allowances and provisions released
r_{12}: extraordinary profits
r_{13}: extraordinary losses

Revenue variables:
r_{14}: total additional revenues
r_{15}: total net sales
r_{16}: allowances for secondment
r_{17}: activated production for own company

r_{18}: subsidies and restitutions
r_{19}: received payments of damages
r_{20}: other additional revenues
r_{21}: net sales main activity of company
r_{22}: net sales other activities

Cost variables:
r_{23}: cost of goods sold
r_{24}: total costs of labour
r_{25}: total additional personnel expenses
r_{26}: total costs of transportation
r_{27}: total costs of energy
r_{28}: total housing costs
r_{29}: total cost of production machines, equipment, and office equipment
r_{30}: total selling expenses
r_{31}: total costs of communication
r_{32}: total cost of third party professional services
r_{33}: total other operations costs
r_{34}: depreciations on tangible and intangible fixed assets

r_{35}: costs of commodity goods sold
r_{36}: other costs of goods sold

r_{37}: gross wages and salaries
r_{38}: employer's part of social security insurance
r_{39}: pensions
r_{40}: other social security contributions

r_{41}: payments to temporary workers
r_{42}: payments to other temporary workers
r_{43}: training costs
r_{44}: other personnel expenses

r_{45}: costs of leasing/renting means of transportation
r_{46}: costs of maintenance for means of conveyance
r_{47}: costs of fuel
r_{48}: ownership tax
r_{49}: insurance premiums for means of conveyance
r_{50}: other costs of transportation

r_{51}: costs of natural gas
r_{52}: costs of electricity
r_{53}: other costs of energy (excluding fuels)

r_{54}: costs of leasing/renting land and buildings

r_{55}: maintenance/repairs land and buildings
r_{56}: costs of cleaning land and buildings
r_{57}: environment tax
r_{58}: property tax
r_{59}: insurance premium for building and contents assurances
r_{60}: other housing costs

r_{61}: renting/leasing machines, equipment, installations, and office equipment
r_{62}: maintenance of machines, equipment, installations, and office equipment
r_{63}: other costs machines, equipment, installations, and office equipment
r_{64}: advertising and promotion expenses
r_{65}: commissions for agents
r_{66}: travelling, accommodation and representation costs
r_{67}: research and development costs
r_{68}: other selling expenses

r_{69}: banking business
r_{70}: other insurance premiums
r_{71}: accountancy, juridical, economical, tax advice
r_{72}: third-party services for automation and computerization
r_{73}: refuse and waste processing
r_{74}: other third-party costs for professional services

r_{75}: licenses, royalties, copyright
r_{76}: intra concern/administrative costs
r_{77}: stationary, contributions, subscriptions, specialist literature
r_{78}: other costs for renting/leasing (not mentioned elsewhere)
r_{79}: other maintenance/reparation costs (not mentioned elsewhere)
r_{80}: other cost price increasing taxes (not mentioned elsewhere)
r_{81}: other general costs (not mentioned elsewhere)

Table B.1: Actual, norm, influence, and difference values for the ABC-company.

	actual	norm	inf(x_i, y)	diff. %		actual	norm	inf(x_i, y)	diff. %
r_1	61.75	11.30		446.46	r_{42}	0.00	0.51	-0.51	-100.00
r_2	60.42	14.79	45.62	308.52	r_{43}	0.17	0.12	0.05	44.29
r_3	1.33	-2.55	3.88	-152.16	r_{44}	0.25	2.62	-2.37	-90.47
r_4	0.00	-0.15	0.15	-100.00	r_{45}	0.00	0.62	-0.62	-100.00
r_5	0.00	-0.79	0.79	-100.00	r_{46}	0.00	0.16	-0.16	-100.00
r_6	329.50	308.64	20.86	6.76	r_{47}	0.00	0.33	-0.33	-100.00
r_7	269.09	293.84	24.76	-8.42	r_{48}	0.00	0.06	-0.06	-100.00
r_8	11.17	1.84	9.33	507.07	r_{49}	0.00	0.12	-0.12	-100.00
r_9	9.83	4.39	-5.44	123.92	r_{50}	0.50	0.42	0.08	19.40
r_{10}	0.00	0.16	0.16	-100.00	r_{51}	0.67	0.51	0.15	29.74
r_{11}	0.00	0.01	-0.01	-100.00	r_{52}	1.17	1.38	-0.21	-15.45
r_{12}	0.00	0.31	-0.31	-100.00	r_{53}	0.08	0.38	-0.29	-77.97
r_{13}	0.00	1.10	1.10	-100.00	r_{54}	0.00	15.26	-15.26	-100.00
r_{14}	4.92	1.54	3.38	220.06	r_{55}	0.50	0.86	-0.36	-41.73
r_{15}	324.58	307.10	17.48	5.69	r_{56}	0.00	0.21	-0.21	-100.00
r_{16}	0.00	0.22	-0.22	-100.00	r_{57}	0.08	0.05	0.03	51.78
r_{17}	0.00	0.00	0.00	0.00	r_{58}	0.58	0.24	0.35	147.92
r_{18}	2.33	0.35	1.98	559.19	r_{59}	1.00	0.49	0.51	104.18
r_{19}	0.00	0.26	-0.26	-100.00	r_{60}	0.00	1.36	-1.36	-100.00
r_{20}	2.58	0.70	1.89	270.91	r_{61}	0.00	0.20	-0.20	-100.00
r_{21}	324.58	304.42	20.16	6.62	r_{62}	0.17	0.37	-0.20	-54.92
r_{22}	0.00	2.68	-2.68	-100.00	r_{63}	0.17	0.10	0.07	69.54
r_{23}	181.42	178.30	3.12	1.75	r_{64}	1.83	6.76	-4.93	-72.89
r_{24}	64.00	56.42	7.58	13.43	r_{65}	0.00	0.03	-0.03	-100.00
r_{25}	0.42	3.61	-3.19	-88.37	r_{66}	1.00	0.48	0.52	106.96
r_{26}	0.50	1.71	-1.21	-70.76	r_{67}	0.00	0.01	-0.01	-100.00
r_{27}	1.92	2.27	-0.36	-15.42	r_{68}	5.58	4.71	0.88	18.63
r_{28}	2.17	18.47	-16.31	-88.25	r_{69}	1.17	0.64	0.53	82.85
r_{29}	0.33	0.67	-0.34	-50.75	r_{70}	0.67	0.54	0.12	22.87
r_{30}	8.42	11.99	-3.57	-29.77	r_{71}	1.33	1.81	-0.47	-26.24
r_{31}	1.00	0.98	0.02	2.04	r_{72}	0.33	0.43	-0.09	-21.67
r_{32}	3.50	4.39	-0.89	-20.27	r_{73}	0.00	0.04	-0.04	-100.00
r_{33}	1.42	5.00	-3.59	-71.60	r_{74}	0.00	0.93	-0.93	-100.00
r_{34}	4.00	10.04	-6.04	-60.16	r_{75}	0.00	0.00	0.00	0.00
r_{35}	181.42	177.69	3.73	2.10	r_{76}	0.00	1.92	-1.92	-100.00
r_{36}	0.00	0.61	-0.61	-100.00	r_{77}	0.67	0.69	-0.03	-3.91
r_{37}	53.50	45.93	7.57	16.49	r_{78}	0.00	0.01	-0.01	-100.00
r_{38}	6.83	6.17	0.66	10.76	r_{79}	0.00	0.14	-0.14	-100.00
r_{39}	3.50	2.95	0.55	18.78	r_{80}	0.00	0.00	0.00	0.00
r_{40}	0.17	2.95	-1.21	-87.93	r_{81}	0.75	2.24	-1.49	-66.53
r_{41}	0.00	0.36	-0.36	-100.00					

B.2 UML use case of diagnostic application

Figure B.1: UML use case sub-diagrams that describe the main use cases in more detail.

Appendix C

Statistics and data for case study 2

C.1 Statistics for OLAP exception identification

Table C.1: Analysis of variance table with response log(profit).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Country	19	69.916	3.680	16.663	< 2.2e-16	***
Personal Accessories	4	41.775	10.444	47.292	< 2.2e-16	***
Residuals	76	16.783	0.221			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4699 on 76 degrees of freedom

Multiple R-squared: 0.8694, Adjusted R-squared: 0.8298

F-statistic: 21.99 on 23 and 76 Df, p-value: < 2.2e-16

Table C.2: Analysis of variance table with response log(revenues).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Country	19	30.3627	1.5980	32.704	< 2.2e-16	***
Personal Accessories	4	3.9947	0.9987	20.438	1.835e-11	***
Residuals	76	3.7137	0.0489			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2211 on 76 degrees of freedom

Multiple R-squared: 0.9025, Adjusted R-squared: 0.8729

F-statistic: 30.57 on 23 and 76 DF, p-value: < 2.2e-16

Table C.3: Tests for homogeneity of variances.

Bartlett test data: log(Revenues) by Personal Accessories by Country Bartlett's K-squared = 2.8585, df = 4, p-value = 0.5818
Bartlett test data: log(Revenues) by Country by Personal Accessories Bartlett's K-squared = 28.4864, df = 19, p-value = 0.0745
Fligner-Killeen test data: log(Revenues) by Personal Accessories by Country Fligner-Killeen:med chi-squared = 3.2684, df = 4, p-value = 0.514
Fligner-Killeen test data: log(Revenues) by Country by Personal Accessories Fligner-Killeen:med chi-squared = 10.3277, df = 19, p-value = 0.9444

C.2 Revenues figures

Year	2001.Quarter
ProductLine	Personal Accessories
Measure	Revenues

	Product Type				
Country.2001.Quarter	Watches	Eyewear	Knives	Binoculars	Navigation
Canada.2001.Q.1	€16,920.00	€26,717.96	€39,202.96	€44,196.96	€52,165.32
Canada.2001.Q.2	€100,872.58	€58,628.26	€29,342.50	€33,229.62	€71,109.64
Canada.2001.Q.3	€104,074.98	€22,268.86	€49,169.04	€54,954.00	€32,962.20
Canada.2001.Q.4	€80,100.54	€54,581.68	€81,776.00	€61,789.00	€69,342.28
Germany.2001.Q.1	€73,829.92	€73,652.52	€77,828.26	€45,088.00	€85,805.36
Germany.2001.Q.2	€71,761.70	€20,672.12	€82,516.76	€13,019.24	€29,416.12
Germany.2001.Q.3	€99,322.14	€49,542.48	€27,108.98	€59,399.24	€57,064.62
Germany.2001.Q.4	€104,980.60	€72,586.34	€103,730.56	€59,561.50	€130,865.24
France.2001.Q.1	€28,471.44	€20,198.00	€50,561.46	€40,047.60	€16,350.40
France.2001.Q.2	€33,533.32	€66,037.08	€57,090.64	€48,860.58	€40,124.62
France.2001.Q.3	€76,428.38	€41,665.32	€39,174.36	€26,161.92	€43,830.42
France.2001.Q.4	€73,446.56	€19,522.26	€35,263.32	€58,319.90	€83,106.66
Mexico.2001.Q.1	€7,209.02	€8,420.54	€4,566.16	€8,728.00	€19,382.66
Mexico.2001.Q.2	€27,902.56	€5,623.38	€17,302.66	€16,973.44	€2,020.00
Mexico.2001.Q.3	€6,909.00	€8,719.58	€14,812.76	€5,760.00	€2,448.50
Mexico.2001.Q.4	€30,169.22	€17,997.30	€39,824.90	€28,939.52	€25,772.00
United States.2001.Q.1	€106,965.40	€56,759.80	€71,670.96	€26,230.40	€78,144.42
United States.2001.Q.2	€93,578.42	€72,817.96	€91,756.74	€19,738.80	€97,568.06
United States.2001.Q.3	€167,364.40	€92,551.64	€97,334.38	€12,912.80	€84,616.16
United States.2001.Q.4	€212,672.64	€67,475.88	€140,382.56	€24,000.00	€143,483.72
Japan.2001.Q.1	€54,596.36	€20,799.54	€48,899.86	€17,682.56	€91,397.50
Japan.2001.Q.2	€98,189.16	€30,331.00	€56,487.42	€44,073.68	€53,590.62
Japan.2001.Q.3	€124,852.92	€25,869.68	€38,715.18	€17,413.84	€25,938.30
Japan.2001.Q.4	€82,193.36	€17,616.50	€46,438.90	€42,732.26	€34,651.42
Australia.2001.Q.1	€21,693.00	€27,242.88	€33,373.06	€58,438.28	€29,665.22
Australia.2001.Q.2	€31,092.20	€11,068.76	€39,060.80	€14,088.96	€22,364.88
Australia.2001.Q.3	€0.00	€7,099.80	€4,084.00	€0.00	€0.00
Australia.2001.Q.4	€53,322.30	€30,509.92	€30,249.88	€35,435.34	€50,896.34
Austria.2001.Q.1	€5,550.00	€4,364.52	€3,130.50	€0.00	€5,920.20
Austria.2001.Q.2	€41,975.96	€31,498.38	€37,196.16	€43,777.72	€47,362.96
Austria.2001.Q.3	€53,640.26	€22,107.52	€32,932.98	€25,648.48	€17,846.82
Austria.2001.Q.4	€50,756.32	€30,194.16	€49,656.32	€37,396.14	€27,082.72
China.2001.Q.1	€2,900.00	€8,599.64	€5,078.40	€13,217.64	€5,239.44
China.2001.Q.2	€18,491.76	€10,479.58	€28,495.76	€18,995.28	€11,459.40
China.2001.Q.3	€11,333.74	€10,049.14	€14,896.12	€9,839.68	€5,039.64
China.2001.Q.4	€29,009.20	€2,160.00	€30,719.76	€18,839.60	€32,497.72
Italy.2001.Q.1	€4,080.00	€8,060.00	€0.00	€0.00	€0.00
Italy.2001.Q.2	€49,246.30	€47,003.62	€70,735.54	€40,871.70	€77,483.30
Italy.2001.Q.3	€65,265.44	€34,273.48	€62,894.86	€37,048.90	€33,341.20
Italy.2001.Q.4	€57,547.54	€37,344.12	€36,472.14	€57,644.12	€20,764.00
Korea.2001.Q.1	€14,050.30	€1,800.00	€18,106.80	€12,238.46	€20,139.16
Korea.2001.Q.2	€18,895.62	€11,038.90	€21,316.96	€32,082.64	€24,318.12
Korea.2001.Q.3	€12,819.78	€16,999.34	€22,027.04	€24,603.28	€13,139.20
Korea.2001.Q.4	€26,406.54	€27,986.10	€32,295.96	€17,717.36	€15,839.34
Netherlands.2001.Q.1	€38,846.52	€16,368.02	€24,550.02	€15,943.74	€29,951.60
Netherlands.2001.Q.2	€64,938.12	€47,117.58	€66,384.08	€54,678.42	€40,600.72
Netherlands.2001.Q.3	€75,351.82	€37,702.76	€33,728.82	€22,807.64	€25,794.04
Netherlands.2001.Q.4	€79,231.76	€35,135.80	€78,432.66	€48,470.84	€52,037.36
Spain.2001.Q.1	€5,312.00	€0.00	€800.00	€0.00	€960.00
Spain.2001.Q.2	€13,973.32	€7,369.92	€15,180.48	€8,147.44	€7,555.14
Spain.2001.Q.3	€18,464.20	€10,365.58	€12,523.56	€19,529.64	€17,823.06
Spain.2001.Q.4	€21,511.96	€12,379.84	€19,006.90	€19,161.52	€13,676.00
Sweden.2001.Q.1	€1,880.00	€15,878.00	€44,128.12	€36,660.58	€21,126.96
Sweden.2001.Q.2	€41,937.30	€28,317.12	€29,936.36	€49,778.48	€46,942.74
Sweden.2001.Q.3	€58,737.82	€18,921.90	€31,370.88	€35,564.20	€29,469.26
Sweden.2001.Q.4	€85,305.46	€35,611.34	€46,364.44	€26,413.98	€33,008.52
Switzerland.2001.Q.1	€22,451.20	€7,895.72	€9,262.10	€11,119.68	€26,579.06
Switzerland.2001.Q.2	€23,183.40	€24,335.84	€40,337.44	€44,979.56	€10,671.00
Switzerland.2001.Q.3	€34,239.22	€26,604.86	€34,106.88	€11,514.00	€8,731.72
Switzerland.2001.Q.4	€41,092.18	€29,471.42	€61,137.24	€72,787.04	€46,430.68
Taiwan.2001.Q.1	€24,571.92	€30,667.22	€46,730.78	€20,319.62	€12,979.88
Taiwan.2001.Q.2	€26,064.00	€24,718.48	€53,262.36	€50,473.88	€10,438.92
Taiwan.2001.Q.3	€8,180.00	€4,559.80	€1,383.40	€6,732.00	€7,019.84
Taiwan.2001.Q.4	€90,016.10	€25,678.68	€28,710.64	€37,197.54	€72,402.50
England.2001.Q.1	€19,043.04	€25,856.42	€12,305.40	€42,846.80	€26,408.76
England.2001.Q.2	€105,501.38	€40,067.28	€58,319.00	€69,106.10	€40,844.88
England.2001.Q.3	€83,267.10	€26,362.44	€23,472.78	€33,237.00	€43,681.52
England.2001.Q.4	€95,133.02	€28,374.10	€80,476.76	€89,573.12	€83,612.12
Belgium.2001.Q.1	€8,473.50	€5,499.72	€9,674.08	€3,410.08	€935.00
Belgium.2001.Q.2	€29,408.86	€10,456.20	€11,318.38	€10,513.44	€5,652.86
Belgium.2001.Q.3	€11,377.80	€6,886.50	€9,974.40	€9,801.36	€10,648.78
Belgium.2001.Q.4	€31,022.74	€14,325.74	€22,633.70	€23,384.78	€22,231.94
Finland.2001.Q.1	€5,396.30	€0.00	€8,901.16	€0.00	€0.00
Finland.2001.Q.2	€24,730.34	€12,881.62	€34,780.64	€30,391.50	€18,672.64
Finland.2001.Q.3	€20,288.30	€11,608.58	€15,384.28	€13,829.48	€20,023.86
Finland.2001.Q.4	€37,334.48	€20,304.24	€26,304.24	€23,593.20	€29,735.54
Brazil.2001.Q.1	€29,246.60	€12,968.80	€16,508.84	€17,711.20	€10,054.02
Brazil.2001.Q.2	€0.00	€0.00	€0.00	€0.00	€0.00
Brazil.2001.Q.3	€9,000.00	€17,727.92	€41,335.98	€21,323.70	€29,755.92
Brazil.2001.Q.4	€44,680.12	€25,928.96	€38,172.32	€43,835.88	€46,709.52

Figure C.1: Revenues figures in the cube 2001.Quarters × Country × Personal Accessories.ProductType

Year	2001.Q1
ProductLine	Personal Accessories
Measure	Revenues

	Product Type				
Country	Watches	Eyewear	Knives	Binoculars	Navigation
Canada	€16,920.00	€26,717.98	€39,202.98	€44,196.96	€52,165.32
Germany	€73,829.92	€73,652.52	€77,828.26	€45,068.00	€65,805.36
France	€28,471.44	€20,198.00	€50,561.46	€40,047.60	€16,350.40
Mexico	€7,209.02	€6,420.54	€4,066.16	€8,728.00	€19,382.66
United States	€106,965.40	€56,759.80	€71,670.96	€26,230.40	€78,144.42
Japan	€54,596.36	€20,799.54	€48,899.86	€17,682.56	€61,397.50
Australia	€21,693.00	€27,242.88	€33,373.08	€58,438.28	€29,665.22
Austria	€5,550.00	€4,364.52	€3,130.50	€0.00	€5,920.20
China	€2,900.00	€6,599.64	€5,078.40	€13,217.64	€5,239.44
Italy	€4,080.00	€6,060.00	€0.00	€0.00	€0.00
Korea	€14,050.30	€1,800.00	€18,106.80	€12,238.46	€20,139.16
Netherlands	€38,646.52	€16,368.02	€24,550.02	€15,943.74	€29,951.60
Spain	€5,312.00	€0.00	€800.00	€0.00	€960.00
Sweden	€1,890.00	€15,878.00	€44,128.12	€36,660.58	€21,126.96
Switzerland	€22,451.20	€7,695.72	€9,262.10	€11,119.68	€26,579.06
Taiwan	€24,571.92	€30,667.22	€46,730.78	€20,319.62	€12,979.88
England	€19,043.04	€25,856.42	€12,305.40	€42,846.80	€26,408.76
Belgium	€8,473.50	€5,499.72	€9,674.08	€3,410.08	€935.00
Finland	€5,398.30	€0.00	€8,901.16	€0.00	€0.00
Brazil	€29,246.60	€12,968.80	€16,508.84	€17,711.20	€10,054.02

Figure C.2: Revenues figures in the cube 2001.Q1 × Country × Personal Accessories.ProductType

Year	2001.Q2
ProductLine	Personal Accessories
Measure	Revenues

	Product Type				
Country	Watches	Eyewear	Knives	Binoculars	Navigation
Canada	€100,872.58	€58,628.26	€29,342.50	€33,229.62	€71,109.64
Germany	€71,761.70	€20,672.12	€82,516.76	€13,019.24	€29,416.12
France	€33,533.32	€66,037.08	€57,090.84	€46,860.56	€40,124.62
Mexico	€27,902.56	€5,623.38	€17,302.66	€18,973.44	€2,020.00
United States	€93,578.42	€72,817.96	€91,756.74	€18,738.80	€97,568.08
Japan	€98,189.16	€30,331.00	€56,487.42	€44,073.68	€53,590.62
Australia	€31,092.20	€11,088.76	€39,060.80	€14,088.96	€22,364.88
Austria	€41,975.96	€31,498.36	€37,196.16	€43,777.72	€47,362.98
China	€16,491.76	€10,479.58	€28,495.76	€18,995.28	€11,459.40
Italy	€49,246.30	€47,003.62	€70,735.54	€40,871.70	€77,483.30
Korea	€18,895.62	€11,038.90	€21,316.96	€32,092.64	€24,318.12
Netherlands	€64,938.12	€47,117.58	€66,384.08	€54,678.42	€40,600.72
Spain	€13,973.32	€7,369.92	€15,180.48	€6,147.44	€7,555.14
Sweden	€41,937.30	€26,317.12	€29,936.36	€49,778.48	€46,942.74
Switzerland	€23,183.40	€24,335.84	€40,337.44	€44,979.56	€10,671.00
Taiwan	€26,064.00	€24,718.48	€53,262.36	€50,473.88	€10,438.92
England	€105,501.38	€40,067.28	€58,319.00	€69,106.10	€40,644.88
Belgium	€29,408.86	€10,458.20	€11,318.38	€10,513.44	€5,652.86
Finland	€24,730.34	€12,881.62	€34,780.84	€30,391.50	€19,672.64
Brazil	€0.00	€0.00	€0.00	€0.00	€0.00

Figure C.3: Revenues figures in the cube 2001.Q2 × Country × Personal Accessories.ProductType

Year	2001.Q3
ProductLine	Personal Accessories
Measure	Revenues

	Product Type				
Country	Watches	Eyewear	Knives	Binoculars	Navigation
Canada	€104,074.98	€22,268.86	€49,169.04	€54,954.00	€32,962.20
Germany	€99,322.14	€49,542.48	€27,108.98	€59,399.24	€57,064.52
France	€76,428.38	€41,655.32	€39,174.36	€26,161.92	€43,830.42
Mexico	€6,909.00	€8,719.58	€14,812.76	€5,760.00	€2,446.50
United States	€167,364.40	€92,551.64	€97,334.38	€12,912.80	€84,616.16
Japan	€124,852.92	€25,869.68	€38,715.18	€17,413.84	€25,938.30
Australia	€0.00	€7,099.80	€4,064.00	€0.00	€0.00
Austria	€53,640.26	€22,107.52	€32,932.98	€25,646.46	€17,646.52
China	€11,333.74	€10,049.14	€14,696.12	€9,839.68	€5,039.64
Italy	€65,265.44	€34,273.48	€62,694.86	€37,048.90	€33,341.20
Korea	€12,819.78	€16,999.34	€22,027.04	€24,603.28	€13,139.20
Netherlands	€75,351.82	€37,702.76	€33,728.82	€22,807.64	€25,794.04
Spain	€18,464.20	€10,365.58	€12,523.56	€19,529.64	€17,823.06
Sweden	€58,737.82	€18,921.90	€31,370.88	€35,564.20	€29,469.26
Switzerland	€34,239.22	€26,604.66	€34,106.88	€11,514.00	€6,731.72
Taiwan	€8,180.00	€4,559.60	€1,383.40	€6,732.00	€7,019.64
England	€63,267.10	€26,362.44	€23,472.78	€33,237.00	€43,681.52
Belgium	€11,377.80	€6,986.50	€9,974.40	€9,601.36	€10,648.78
Finland	€20,288.30	€11,608.58	€15,384.28	€13,929.46	€20,023.66
Brazil	€9,000.00	€17,727.92	€41,335.98	€21,323.70	€29,755.92

Figure C.4: Revenues figures in the cube 2001.Q3 × Country × Personal Accessories.ProductType

Year	2001.Q4
ProductLine	Personal Accessories
Measure	Revenues

	Product Type				
Country	Watches	Eyewear	Knives	Binoculars	Navigation
Canada	€80,100.54	€54,581.68	€81,776.00	€61,789.00	€69,342.28
Germany	€104,980.60	€72,586.34	€103,720.56	€59,561.50	€130,865.24
France	€73,446.56	€19,522.26	€35,263.32	€88,319.90	€83,106.68
Mexico	€30,169.22	€17,997.30	€39,624.90	€28,939.52	€25,772.00
United States	€212,672.64	€67,475.88	€140,382.56	€24,000.00	€143,493.72
Japan	€82,193.36	€17,616.50	€46,438.90	€42,732.26	€34,651.42
Australia	€53,322.30	€30,509.92	€30,249.88	€35,435.34	€50,896.34
Austria	€50,756.32	€30,194.16	€49,656.32	€37,396.14	€27,082.72
China	€29,009.20	€2,160.00	€30,718.76	€18,939.60	€32,497.72
Italy	€57,547.54	€37,344.12	€36,472.14	€57,644.12	€20,764.00
Korea	€26,406.54	€27,986.10	€32,295.96	€17,717.36	€15,839.34
Netherlands	€78,231.78	€35,135.80	€78,432.66	€48,470.84	€52,037.36
Spain	€21,511.98	€12,379.84	€19,006.90	€19,181.52	€13,676.00
Sweden	€85,305.46	€35,611.34	€46,364.44	€26,413.98	€33,008.52
Switzerland	€41,092.18	€29,471.42	€61,137.24	€72,787.04	€46,430.68
Taiwan	€90,016.10	€25,678.68	€28,710.64	€37,197.54	€72,402.50
England	€95,133.02	€28,374.10	€80,476.76	€89,573.12	€63,612.12
Belgium	€31,022.74	€14,325.74	€22,633.70	€23,384.78	€22,231.64
Finland	€37,334.48	€20,304.24	€26,304.24	€23,593.20	€29,735.54
Brazil	€44,680.12	€25,928.96	€38,172.32	€43,835.88	€46,709.52

Figure C.5: Revenues figures in the cube 2001.Q4 × Country × Personal Accessories.ProductType

Year	2000
Product	All-Products
Measure	Profit

Country	Year
	2000
Canada	€266,767.25
Germany	€174,588.79
France	€244,804.01
Mexico	€120,966.25
United States	€637,396.16
Japan	€345,135.30
Australia	€131,634.56
Austria	€241,766.72
China	€220,819.16
Italy	€301,064.74
Korea	€171,534.00
Netherlands	€378,324.70
Spain	€227,834.59
Sweden	€459,965.71
Switzerland	€219,314.94
Taiwan	€211,504.33
England	€357,878.04
Belgium	€176,788.39
Finland	€122,004.56
Brazil	-€111,112.55

Figure C.6: Profit figures in the year 2000 (with a slice), derived from the example financial database, organised per Country (L^3) and All-Products (P^3). Here the historical normative model is based on these figures.

C.3 Aggregated tables for generic explanation

Table C.4: Aggregated table for explanation of $\partial y^{223}(c)$ ="low" in the Product dimension, with a slice on Quarter 3 in the context cube Year.Q3 × Amsterdam × All-Products, where $c' \in c\downarrow$ over the path [233] → [232] → [231] → [230].

	actual	reference	$\inf(y(c'), y^{223}(c))$	rel. inf.
(2001,.,.)	199,690.65	378,324.70	-178,634.05	
(Q3,.,All-Products)	67,569.40	60,119.14	-7,450.26	0.04
(Q3,.,Knives)	1,576.25	12,139.72	-10,563.47	0.06
(Q3,.,Knives.Survival Edge)	594.00	9,368.75	-8,774.75	0.05
...	...	...	...	...
(Q3,.,Tents)	-24,255.78	-37,310.92	13,055.14	-0.07
(Q3,.,Tents.Star Gazer 3)	-5,693.20	-23,481.20	17,788.00	-0.10

Table C.5: Aggregated table for the Product dimension where the actual object is profit(2001, Spain), the norm object is profit(2000, Spain), and the influence values for instances within the Product dimension related to the exceptional cell profit^{223}(2001, Spain, All-Products).

Nr.	ProductLine P^2	ProductType P^1	Product P^0	Actual (2001)	Norm (2000)	Rel. Inf.
	All	All	All	86,248.94	227,834.59	
2	Camp. Equip.	All	All	-28,655.09	11,706,68	0.29
...	...	...	...	...	...	...
7	Camp. Equip.	Tents	All	-29,405.88	-15,903.43	0.10
...	...	...	...	...	...	...
13	Camp. Equip.	Lanterns	All	1,319.52	8,208.17	0.05
...	...	...	...	...	...	...

Table C.6: Aggregated table for the Product dimension where the actual object is profit(2001, Canada), the norm object is profit(2000, Canada), and the influence values for instances within the Product dimension related to the exceptional cell profit^{223}(2001, Canada, All-Products).

Nr.	ProductLine P^2	ProductType P^1	Product P^0	Actual (2001)	Norm (2000)	Rel. Inf.
	All	All	All	141,777.29	266,767.25	
1	Camp. Equip.	All	All	-389,955.25	-184,914.96	1.64
2	Camp. Equip.	Tents	All	-450,053.83	-270,470.19	1.44
...	...	...	...	...	...	...
7	Camp. Equip.	Lanterns	All	16,965.00	32,761.10	0.13
...	...	...	...	...	...	...

Table C.7: Aggregated table for the Product dimension where the actual object is profit(2001, Belgium), the norm object is profit(2000, Belgium), and the influence values for instances within the Product dimension related to the exceptional cell profit^{223}(2001, Belgium, All-Products).

Nr.	ProductLine P^2	ProductType P^1	Product P^0	Actual (2001)	Norm (2000)	Rel. Inf.
	All	All	All	88,679.70	176,788.39	
1	Camp. Equip.	All	All	-28,655.09	11,706,68	0.45
...	...	...	...	...	...	...
4	Camp. Equip.	Tents	All	-47,258.10	-32,126.44	0.17
...	...	...	...	...	...	...
12	Camp. Equip.	Lanterns	All	5,528.30	11,014.41	0.06
...	...	...	...	...	...	...

Table C.8: Aggregated table for the Product dimension where the actual object is profit(2001, Sweden), the norm object is profit(2000, Sweden), and the influence values for instances within the Product dimension related to the exceptional cell profit^{223}(2001, Sweden, All-Products).

Nr.	ProductLine P^2	ProductType P^1	Product P^0	Actual (2001)	Norm (2000)	Rel. Inf.
	All	All	All	369,004.16	459,965.71	
2	Camp. Equip.	All	All	-6,463.02	39,338.77	0.50
3	Camp. Equip.	Tents	All	-119,084.09	-92,667.57	0.29
...	...	...	...	...	...	...
11	Camp. Equip.	Lanterns	All	29,415.65	34,938.77	0.06
...	...	...	...	...	...	...

Appendix D

Matrix representation of OLAP databases

In this appendix, we present the matrix representation of a system of additive drill-down equations in an OLAP database and a number of its properties, as an alternative representation for the notation put forward in Chapter 2. To the best of our knowledge such a matrix representation of OLAP is not yet presented in the current literature. However, we do suppose that in technical OLAP implementations, such as MOLAP (Section 2.1.2), similar matrix representations are used in the software.

D.1 Matrix notation

A system of drill-down equations as formulated in (2.11), can be written as a system of *implicit equations* and represented in matrix form as

$$A\mathbf{z} = \mathbf{0}, \tag{D.1}$$

where A is a $m \times k$ *binary* coefficient matrix of constants, $\mathbf{z}$ is a $k \times 1$ vector of variables, and $\mathbf{0}$ is a $m \times 1$ vector of zeros. The matrix A in (D.1) can be partitioned as $A = [A_1\ A_2]$, where A_1 is the $m \times n$ coefficient submatrix for dependent/non-base variables and A_2 is the $m \times l$ coefficient submatrix for independent/base variables. Moreover, the vector of variables $\mathbf{z}$ in (D.1) is partitioned in a $n \times 1$ vector of dependent variables $\mathbf{y}$ for which we need solutions, and in a $l \times 1$ vector of independent variables $\mathbf{x}$ which are given, and represented as $\mathbf{z}' = [\mathbf{y}\ \mathbf{x}]$. As a result the system of equations

in (D.1) can be written in partitioned form as

$$A_1\mathbf{y} + A_2\mathbf{x} = \mathbf{0}. \tag{D.2}$$

Typically, the system of equations represented in (D.2) is overdetermined, i.e. there are more equations than dependent variables, because each measure is typically associated with multiple dimensions. Therefore, the matrix A_1 is non-square ($m > n$) and not invertible. We will show that in OLAP systems D.2 is uniquely solvable. Notice that, only in the case of a measure that is associated with only one dimension the matrix A_1 is square.

In the next paragraphs we discuss relevant matrix theory on the conditions under which equation (D.2) is consistent and solvable. To deal with the problem of system overspecification, the number of equations in (D.2) is reduced with the following method. For each non-base cube C in L, we write down all the drill-down equations (Equation 2.12) for its cell measure values in a *single*, *arbitrary*, *dimension* D_q. Basically, we write down one equation for each dependent variable in the vector $\mathbf{y}$. In this manner, we derive the reduced form of the submatrix A_1, denoted by A_1^*. The submatrix A_1^* is square ($n \times n$), due to each dependent variable being associated with only a single drill-down equation. The main structure of the submatrix A_2 is not influenced by this procedure, because it represents the coefficients of the base variables in C_B. Only the rows with zero values in A_2 are reduced accordingly. Now the matrix A^* has the following canonical and hierarchical structure:

$$A^* = [A_1^*\ A_2] = \tag{D.3}$$

$$\left(\begin{array}{cccccc|c} -I_1 & P_1 & 0 & 0 & 0 & \dots & 0 \\ 0 & -I_2 & P_2 & 0 & 0 & \dots & 0 \\ 0 & 0 & -I_3 & P_3 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & -I_{m-1} & P_{n-1} & 0 \\ 0 & 0 & \dots & 0 & 0 & -I_m & P_n \end{array}\right) \begin{array}{lll} \text{level} & q & \text{(top)} \\ \text{level} & q-1 & \\ \text{level} & q-2 & \\ \vdots & \vdots & \\ \text{level} & 1 & \\ \text{level} & 0 & \text{(base).} \end{array}$$

The submatrices of A^* correspond to hierarchical levels in L. In particular, each matrix partition in the form of the submatrix $[-I\ P]$ on the diagonal of A^* corresponds with a specific level ($q = i_1 + i_2 + \ldots + i_n$) in L, where I is the identity matrix and P the binary coefficient matrix on that level. For instance, the submatrix $[-I_m\ P_n]$

represents the equations on the base level 0, the submatrix $[-I_{m-1}\ P_{n-1}]$ represents equations on level 1, and so on. Finally, the submatrix $[-I_1\ P_1]$ represents a single equation on the top level q of L.

Analogously, the vector of non-base variables $\mathbf{y}$ is partitioned as:

$$\mathbf{y}' = [y^{\text{level } q}\ \mathbf{y}^{\text{level } q-1}\ \mathbf{y}^{\text{level } q-2}\ \ldots\ \mathbf{y}^{\text{level } 2}\ \mathbf{y}^{\text{level } 1}],$$

where each partition corresponds with the variables on a specific level in L. The variables in each partition are siblings of each other in the lattice.

The reduction method produces for each parent cube C in L a single drill-down path to C_B. The method produces a canonical structure for (D.3) due to the commutativity of drill-down operators (Lemma 2.2.1), which is the same irrespective of the dimension selected for drill-down. Furthermore, if a different dimension would have been selected on some aggregation level in L in the reduction method, only the submatrix P would change to P'. Coefficients of independent variables associated with the selected dimension will be labelled with ones and coefficients of independent variables associated with all the non-selected dimensions will be labelled with zeros in P'. Obviously, the submatrix P_n, with all the coefficients of the base variables, does not change.

The submatrix A_1^* is an upper triangular matrix, and

$$\text{rank}(A_1^*) = \text{rank}(I_1) + \text{rank}(I_2) + \text{rank}(I_3) + \ldots + \text{rank}(I_{m-1}) + \text{rank}(I_m) = n.$$

The determinant of A_1^* is the product of the diagonal entries. From the identity submatrices on the diagonal of A_1^*, we can conclude that $\det(A_1^*) \neq 0$. Hence A_1^* is invertible and therefore the reduced system

$$A_1^*\mathbf{y} + A_2\mathbf{x} = \mathbf{0} \tag{D.4}$$

is uniquely solvable. The solution is given by $\mathbf{y} = -A_1^{*-1}A_2\mathbf{x}$.

The matrix A^* has the property that

$$P_1 \cdot P_2 \cdots P_{n-1} \cdot P_n = \mathbf{1}. \tag{D.5}$$

This property is the result of the fact that the root variable $y^{\max_1 \max_2 \ldots \max_n}(C_T)$ is the sum of all the base variables $\mathbf{x}^{00\ldots0}$ (Theorem 2.3.1). This is illustrated in the

following way. By design the matrix A^* represents single drill-down paths from all non-base cubes (including C_T) to C_B. In these paths, Equation 2.12 is applied on all cells. Therefore, the root variable $y^{\max_1 \max_2 \ldots \max_n}(C_T)$ is the sum of the non-base variables on level $q-1$, the sum of the non-base variables on level $q-2$, the sum of the non-base variables on level $q-3$, and so on. This is shown in the following series of matrix multiplications of the submatrices $P_1 \cdot P_2 \cdots P_{n-1} \cdot P_n$:

$$\begin{aligned} y^{\max_1 \max_2 \ldots \max_n}(C_T) = y^{\text{level } q}(C_T) = &\ P_1 \cdot \mathbf{y}^{\text{level } q-1} \\ = &\ P_1 \cdot P_2 \cdot \mathbf{y}^{\text{level } q-2} \\ = &\ P_1 \cdot P_2 \cdot P_3 \cdot \mathbf{y}^{\text{level } q-3} \\ &\ \vdots \\ = &\ P_1 \cdot P_2 \cdots P_{n-1} \cdot \mathbf{y}^{\text{level } 1} \\ = &\ P_1 \cdot P_2 \cdots P_{n-1} \cdot P_n \cdot \mathbf{x}^{\text{level } 0}. \end{aligned} \tag{D.6}$$

For example, the first row in this series should be interpreted as the matrix representation of a drill-down from $y^{\max_1 \max_2 \ldots \max_n}(C_T)$ to $\mathbf{y}^{\text{level } q-1}(R_q^{-1}(C_T))$ and the last row in this series of matrix operations is equivalent with a series of drill-down operations from $y^{\max_1 \max_2 \ldots \max_n}(C_T)$ over some drill-down path to $\mathbf{x}^{\text{level } 0}(C_B)$. The product of all submatrices $P_1 \cdot P_2 \cdots P_{n-1} \cdot P_n$ in A^*, has to result in a vector of ones, denoted by $\mathbf{1}$, shown by $y^{\max_1 \max_2 \ldots \max_n}(C_T) = P_1 \cdot P_2 \cdots P_{n-1} \cdot P_n \cdot \mathbf{x}^{\text{level } 0} = \mathbf{1} \cdot \mathbf{x}^{00\ldots0}$, because of Theorem 2.3.1.

Now suppose that we write down, for one non-base cube C in L, all the drill-down equations for its cell measure values for *a different dimension* D_q. The structure of the reduced matrix A^* will be the same, because there is still one drill-down equation for each non-base variable. Coefficients in the submatrix $[-I\ P]$ are changed into the coefficients of the submatrix $[-I\ P']$, corresponding to the new dimension. We derive a matrix $A^{*\prime}$ with the same structure as (D.4), however with one or multiple changed submatrices, e.g. $[-I_1\ P_1']$, corresponding with a different dimension for drill-down on the root level q:

$$A^{*\prime} = [A_1^{*\prime}\ A_2] = \left(\begin{array}{cccccc|c} -I_1 & P_1' & 0 & 0 & 0 & \ldots & 0 \\ 0 & -I_2 & P_2' & 0 & 0 & \ldots & 0 \\ 0 & 0 & -I_3 & P_3' & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & -I_{m-1} & P_{n-1}' & 0 \\ 0 & 0 & \ldots & 0 & 0 & -I_m & P_n' \end{array}\right).$$

Notice that the product of all submatrices $P'_1 \cdot P'_2 \cdots P'_{n-1} \cdot P'_n$ in $A^{*\prime}$ also results in a vector of ones: $P'_1 \cdot P'_2 \cdots P'_{n-1} \cdot P'_n = \mathbf{1}$ (see expression D.5) due to the commutativity of drill-down operators (Theorem 2.2.1).

Now we show that the matrix $[A_1^{*\prime}\ A_2]$ can always be obtained from the matrix $[A_1^*\ A_2]$ by means of elementary row operations[1] (Schott 1997). In other words, the matrices $[A_1^{*\prime}\ A_2]$ and $[A_1^*\ A_2]$ are *row-equivalent matrices*. We write $R(A)$ for the matrix obtained by performing row operation R on matrix A. Each row operation R defines an (invertible) elementary matrix $E_R = R(I_n)$ by performing row operation R on the identity matrix I_n. It can easily be shown that $E_R \cdot A = R(A)$ (Schott 1997). If $[A_1^{*\prime}\ A_2]$ and $[A_1^*\ A_2]$ are row-equivalent matrices, then there is a sequence of $E_1, E_2, \ldots, E_k$ matrices such that $A_1^{*\prime} = E_1 E_2 \ldots E_k A_1^*$, in particular an invertible transformation matrix E such that $EA_1^* = A_1^{*\prime}$ and $EA_2 = A_2$. The product of the sequence of elementary matrices, the elementary transformation matrix E, when we write down different drill-down equations for all the cell measure values of non-base cubes C on each level of L, has the form

$$E = \tag{D.7}$$

$$\begin{pmatrix} I_1 & P_1 - P'_1 & (P_1 - P'_1) \cdot P_2 & \ldots & (P_1 - P'_1) \cdot P_2 \cdot P_3 \cdots P_{n-1} \\ 0 & I_2 & P_2 - P'_2 & \ldots & (P_2 - P'_2) \cdot P_3 \cdot P_4 \cdots P_{n-1} \\ 0 & 0 & I_3 & \ldots & (P_3 - P'_3) \cdot P_4 \cdot P_5 \cdots P_{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & I_{m-1} & P_{n-1} - P'_{n-1} \\ 0 & 0 & \ldots & 0 & I_m \end{pmatrix}.$$

The multiplications of submatrices in (D.7) are the result of drill-down operations, as defined in D.6, from one aggregation level to the next in L, combined with their property of commutativity (see Theorem 2.2.1) expressed as $P_1 \cdot P_2 \cdots P_{n-1} \cdot P_n = P'_1 \cdot P_2 \cdots P_{n-1} \cdot P_n$.

Theorem D.1.1. Let $(\mathbf{y}_0, \mathbf{x}_0)$ be the solution for the reduced system of drill-down equations $A_1^*\mathbf{y}_0 + A_2\mathbf{x}_0 = \mathbf{0}$. It is also the solution of the alternative reduced system $A_1^{*\prime}\mathbf{y}_0 + A_2\mathbf{x}_0 = \mathbf{0}$, where $[A_1^*\ A_2] = [A_1^{*\prime}\ A_2]$ are row-equivalent.

[1] Elementary row operations are: row switching, row multiplication, and row addition.

Proof. If the matrix E is the invertible, elementary transformation matrix such that $E \cdot [A_1^* \; A_2] = [A_1^{*\prime} \; A_2]$, then

$$\begin{aligned} &A_1^*\mathbf{y}_0 + A_2\mathbf{x}_0 = \\ &E(A_1^*\mathbf{y}_0 + A_2\mathbf{x}_0) = \\ &EA_1^*\mathbf{y}_0 + EA_2\mathbf{x}_0 = \\ &A_1^{*\prime}\mathbf{y}_0 + A_2\mathbf{x}_0 = \mathbf{0}. \end{aligned}$$

In summary, we have shown that the reduced system of drill-down equations in (D.3) has a unique solution and that this solution also holds for all alternative reduced systems (see Theorem D.1.1) that can be derived from the original system of drill-down equations in (D.2). The original system of equations is uniquely solvable due to Theorem 2.3.1. A solution for this overdetermined system can be computed (Caron and Daniels 2009).

The fact that (D.2) is uniquely solvable implies $\text{rank}(A_1|-A_2\mathbf{x}) = \text{rank}(A_1)$ for all $\mathbf{x}$, see Theorem 6.1 from Schott (1997). So the columns of A_2 are linear combinations of the columns of A_1, so $A_2 = A_1Z$ where Z is a $n \times l$ matrix of constants. Furthermore, since the solution for $\mathbf{y}$ is unique we have $\text{rank}(A_1) = n$ because the null space of A_1 is $N(A_1) = \{\mathbf{0}\}$. So also Z is unique since $A_1Z = A_1Z^*$ would imply $A_1(Z - Z^*) = \mathbf{0}$ and because $N(A_1) = \{\mathbf{0}\}$, we have $Z = Z^*$.

It is also easy to show that

$$Z = A_1^- A_2, \tag{D.8}$$

where A_1^- is the left generalized inverse of A_1. It exists because $\text{rank}(A_1) = n$ and $A_1^- A_1 = I_n$, which is based on Theorem 6.6 from Schott (1997). Notice that $A_1Z = A_2$ implies:

$$A_1A_1^-A_2 = A_1A_1^-A_1A_2 = A_1Z = A_2. \tag{D.9}$$

So $A_1^-A_2$ is a solution of $A_1Z = A_2$ and therefore $Z = A_1^-A_2$ by uniqueness. Using Equation (D.9) it can be shown that the complete system of equations, represented by (D.2), always has a unique solution for a given set of base variables.

D.2 Example systems of drill-down equations

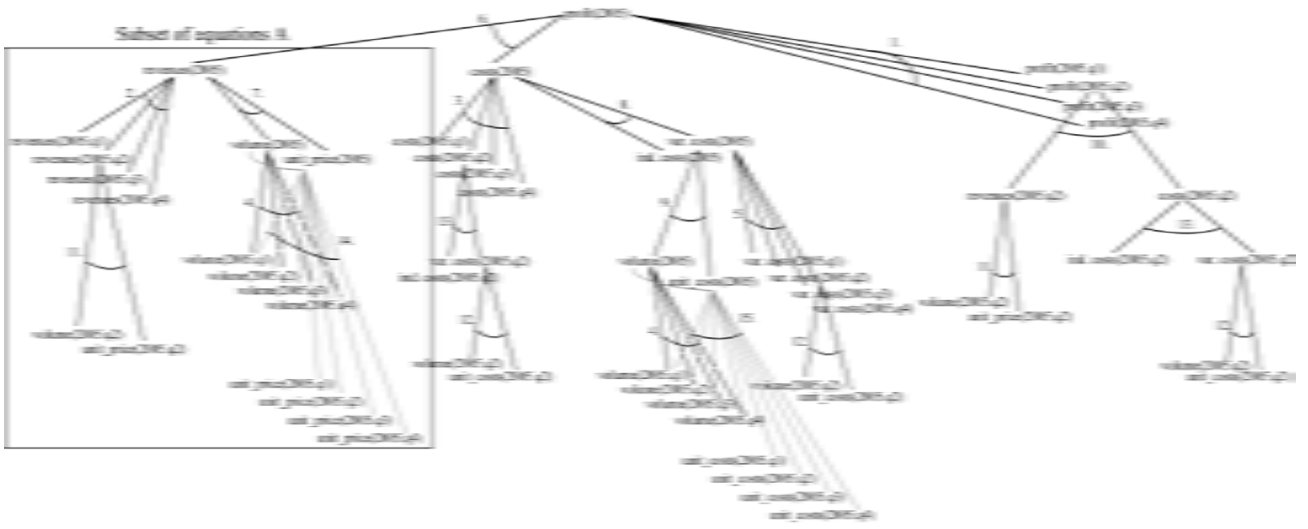


Figure D.1: Mixed systems of drill-down and business equations with subset of equations A, being solvable for what if analysis with the variable volume(2005.Q2).

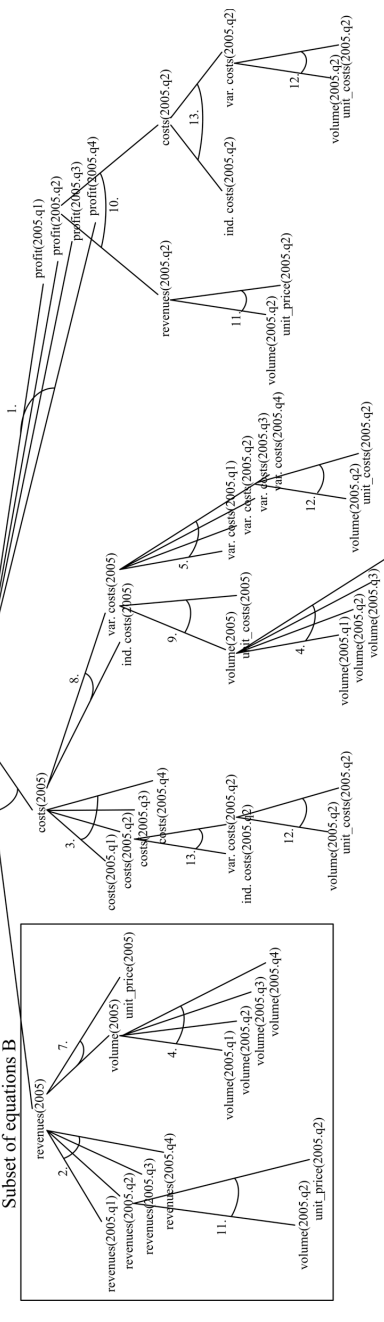


Figure D.2: Mixed system of drill-down and business equations, with subset of equations B, not being solvable for what if analysis with the variable volume(2005.Q2).

Bibliography

Agrawal, R., A. Gupta, and S. Sarawagi (1997). Modeling multidimensional databases. In *ICDE '97: Proceedings of the Thirteenth International Conference on Data Engineering*, Washington, DC, USA, pp. 232–243. IEEE Computer Society.

Apte, C. and S. J. Hong (1986). Using qualitative reasoning to understanding financial arithmetic. In *Proc. of AAAI-86*, Philadelphia, PA, pp. 942–948.

AVc (2006). Foundation for tackling vehicle crime. http://www.stavc.nl.

Baird, B. (1990). *Managerial Decisions Under Uncertainty: An Introduction to the Analysis of Decision Making*. New York: John Wiley & Sons, Inc.

Balmin, A., Y. Papakonstantinou, and T. Papadimitriou (2000). Optimization of hypothetical queries in an olap environment. *Data Engineering, International Conference on 0*, 311.

Barnett, V. and T. Lewis (1994, April). *Outliers in Statistical Data*. Wiley Series in Probability & Statistics. Wiley.

Benjamins, V. R. (1993, June). *Problem Solving Methods for Diagnosis*. Ph. D. thesis, University of Amsterdam, Amsterdam, The Netherlands.

Berndsen, R. and H. A. M. Daniels (1990, May). Qualitative dynamics and causality in a keynesian model. *Journal of Economic Dynamics and Control 14*(2), 435–450.

Binbasioglu, M. and E. J. Zychowicz (1998). Knowledge-based management support: an application of diagnostic reasoning to corporate financing decisions. *International Journal of Intelligent Systems in Accounting, Finance & Management 7*(4), 199–211.

Bishop, Y., S. Fienberg, and P. Holland (1975). *Discrete multivariate analysis theory*. The MIT Press.

Bonge, J. W. (1972). Problem recognition and diagnosis: basic inputs to business policy. *Journal of business policy*, 45–53.

Bouwman, M. J. (1983). Human diagnostic reasoning by computer: An illustration from financial analysis. *Management Science 29*(6), 653–672.

Cabibbo, L. and R. Torlone (1998). A logical approach to multidimensional databases. In *EDBT '98: Proceedings of the 6th International Conference on Extending Database Technology*, London, UK, pp. 183–197. Springer-Verlag.

Cariou, V., J. Cubillé, C. Derquenne, S. Goutier, F. Guisnel, and H. Klajnmic (2007). Built-in indicators to automatically detect interesting cells in a cube. In I. Song, J. Eder, and T. Nguyen (Eds.), *Data Warehousing and Knowledge Discovery*, Volume 4654 of *Lecture Notes in Computer Science*, pp. 123–134. Springer Berlin Heidelberg.

Cariou, V., J. Cubillé, C. Derquenne, S. Goutier, F. Guisnel, and H. Klajnmic (2008). Built-in indicators to discover interesting drill paths in a cube. In *Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery*, DaWaK '08, Berlin, Heidelberg, pp. 33–44. Springer-Verlag.

Caron, E. A. M. and H. A. M. Daniels (2005). Automated business diagnosis in the olap context. In H. A. Fleuren, D. den Hertog, and P. M. K. (Eds.) (Eds.), *Operations Research Proceedings 2004*, Berlin, Germany, pp. 425–433. Springer.

Caron, E. A. M. and H. A. M. Daniels (2007). Explanation of exceptional values in multi-dimensional business databases. *European Journal of Operational Research 188*, 884–897.

Caron, E. A. M. and H. A. M. Daniels (2008). Extensions to the olap framework for business analysis. In J. Cordeiro, B. Shishkov, A. Ranchordas, and M. Helfert (Eds.), *ICSOFT (ISDM/ABF)*, pp. 240–247. INSTICC Press.

Caron, E. A. M. and H. A. M. Daniels (2009). Business analysis in the olap context. In J. Cordeiro and J. Filipe (Eds.), *ICEIS (2)*, pp. 325–330.

Caron, E. A. M. and H. A. M. Daniels (2010). What-if analysis in olap - with a case study in supermarket sales data. In J. Filipe and J. Cordeiro (Eds.), *ICEIS (1)*, pp. 208–213. SciTePress.

Caron, E. A. M. and H. A. M. Daniels (2012). Explanatory analysis in business intelligence systems. In *ECIS 2012 Proceedings*, Barcelona, Spain, pp. 77–89, Paper 87. AIS Electronic Library (AISeL).

Caron, E. A. M. and H. A. M. Daniels (2013). Explanatory business analytics in olap. *International Journal of Business Intelligence Research (IJBIR) 4*(3), 67–82.

Caron, E. A. M. and A. Veenstra (2007). Explanation of exceptional values in multidimensional business databases - with a case study on the analysis of vehicle criminality data. In *Proceedings of international conference on industrial*

engineering and systems management, Beijing, China, pp. pp. cd–rom 11 pages. Tsinghua University Press.

Chandola, V., A. Banerjee, and V. Kumar (2009). Anomaly detection: A survey. *ACM Comput. Surv. 41*(3).

Chandrasekaran, B. and S. Mittal (1984). Deep versus compiled knowledge approaches to diagnostic problem-solving. In M. J. Coombs (Ed.), *Developments in Expert Systems*, pp. 23–34. London: Academic Press.

Chen, Q. (1999). Mining exceptions and quantitative association rules in olap data cube.

Chen, Y., G. Dong, J. Han, B. W. Wah, and J. Wang (2002). Multi-dimensional regression analysis of time-series data streams. In *VLDB '02: Proceedings of the 28th international conference on Very Large Data Bases*, pp. 323–334. VLDB Endowment.

Ciferri, C., R. Ciferri, L. Gomez, M. Schneider, A. Vaisman, and E. Zimayi (2013, APR-JUN). Cube algebra: A generic user-centric model and query language for olap cubes. *International journal of Data Warehousing and Mining 9*(2), 39–65.

Citro, G., G. Banks, and G. F. Cooper (1997). Inkblot: A neurological diagnostic decision support system integrating causal and anatomical knowledge. *Artificial Intelligence in Medicine 10*(3), 257–267.

Clemson, B., T. Yongming, J. Pyne, and R. Unal (1995). Efficient methods for sensitivity analysis. *System Dynamics Review 11*(1), 31–49.

Cliqview Corporation (2010). Cliqview.

Codd, E. F. (1993). Providing olap (on-line analytical processing) to user-analysts: An it mandate. Technical report, E. F. Codd and Associates.

Console, L. and P. Torasso (1989). *Diagnostic problem solving : combining heuristic, approximate and causal reasoning*. London: North Oxford Academic.

Courtney, J. F., D. B. Paradice, and N. H. A. Mohammed (1987). A knowledge-based dss for managerial problem diagnosis. *Decision Sciences 18*(3), 373–399.

Currier, K. (2000). *Comparative Statics Analysis in Economics*. Singapore: World Scientific Publishing Co.

Daniels, H. A. M. and E. A. M. Caron (2007). Explanation generation in business performance models - with a case study in competition benchmarking. In J. Cardoso, J. Cordeiro, and J. Filipe (Eds.), *ICEIS (2)*, pp. 119–128.

Daniels, H. A. M. and E. A. M. Caron (2009). Automated explanation of financial data. *Intelligent Systems in Accounting, Finance & Management 16*(1-2), 5–19.

Datta, A. and H. Thomas (1999). The cube data model: a conceptual model and algebra for on-line analytical processing in data warehouses. *Decis. Support Syst. 27*(3), 289–301.

Davis, R. and W. Hamscher (1988). Model-based reasoning: troubleshooting. pp. 297–346.

de Kleer, J. and J. S. Brown (1986). Theories of causal ordering. *Artif. Intell. 29*(1), 33–61.

de Kleer, J., A. K. Mackworth, and R. Reiter (1992). Characterizing diagnoses and systems. *Artif. Intell. 56*(2-3), 197–222.

de Kleer, J. and B. C. Williams (1989, August). Diagnosis with behavioral modes. pp. 1324–1330. IJCAI-89: Morgan Kaufmann,.

de Kleer, J. and B. C. Williams (1992). Diagnosing multiple faults. pp. 100–117.

Deming, W. and F. Stephan (1940). On a least square adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist. 11*, 427–444.

Emory, C. W. and P. Niland (1968). *Making Management Decisions*. Boston, MA, USA: Houghton Mifflin.

Everitt, B. S. (1994). *The Analysis of Contingency Tables*. London: Chapman & Hall.

Feelders, A. J. (1993). *Diagnostic reasoning and explanation in financial models of the firm*. Ph. D. thesis, Tilburg University, Department of Economics.

Feelders, A. J., H. Daniels, and M. Holsheimer (2000). Methodological and practical aspects of data mining. *Inf. Manage. 37*(5), 271–281.

Feelders, A. J. and H. A. M. Daniels (2001). A general model for automated business diagnosis. *European Journal of Operational Research 130*, 623–637.

Fridson, M. S. and F. Alvarez (2002). *Financial Statement Analysis: A Practitioner's Guide*. New York: Wiley; 3 edition.

Gartner (2011). Market share analysis: Business intelligence, analytics and performance management, worldwide, 2010. http://www.gartner.com.

Genesereth, M. R. (1984). The use of design descriptions in automated diagnosis. *Artificial Intelligence 24*(1-3), 411–436.

Giacometti, A., P. Marcel, and E. Negre (2008). A framework for recommending olap queries. In *Proceedings of the ACM 11th international workshop on Data warehousing and OLAP*, DOLAP '08, New York, NY, USA, pp. 73–80. ACM.

Giacometti, A., P. Marcel, E. Negre, and A. Soulet (2011, APR-JUN). Query recommendations for olap discovery-driven analysis. *International journal of data warehousing and mining 7*(2), 1–25.

Granger, C. W. J. (2001). Testing for causality: a personal viewpoint. pp. 48–70.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics 11*(1), 1–21.

Gyssens, M. and L. V. S. Lakshmanan (1997). A foundation for multi-dimensional databases. In M. Jarke, M. J. Carey, K. R. Dittrich, F. H. Lochovsky, P. Loucopoulos, and M. A. Jeusfeld (Eds.), *VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece*, pp. 106–115. Morgan Kaufmann.

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics 29*, 205–220.

Hamscher, W. (1990). Explaining unexpected financial results. Technical report 11, Price Waterhouse Technology Centre. Also in working notes of the AAAI Spring Symposium on Automated Abduction.

Hamscher, W. (1992, Dec). Model-based reasoning in financial domains. Working Papers 27, Price Waterhouse. available at http://ideas.repec.org/p/wop/prwawp/027.html.

Hamscher, W. (1994). Crosby: Financial data interpretation as model-based diagnosis. *Ann. Math. Artif. Intell. 11*(1-4), 511–524.

Han, J. (1995). Mining knowledge at multiple concept levels. In *CIKM '95: Proceedings of the fourth international conference on Information and knowledge management*, New York, NY, USA, pp. 19–24. ACM Press.

Han, J. and M. Kamber (2005). *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Hawkins, D. (1994). *Identification of Outliers*. Chapman and Hall, London.

Heckman, J. (2000). Causal parameters and policy analysis in economics: A twentieth century retrospective. *The Quarterly Journal of Economics 115*(1), 45–97.

Hesslow, G. (1983). Explaining differences and weighting causes. *Theoria 49*, 87–111.

Hoaglin, D. C., F. Mosteller, and e. J. W. Tukey (1983). *Understanding Robust and Exploratory Data Analysis*. New York: Wiley series in probability.

Hoaglin, D. C., F. Mosteller, and J. W. Tukey (1988). *Exploring data tables, trends and shapes*. New York: Wiley series in probability.

Horner, J., I. Y. Song, and P. P. Chen (2004). An analysis of additivity in olap systems. In *DOLAP '04: Proceedings of the 7th ACM international workshop on Data warehousing and OLAP*, New York, NY, USA, pp. 83–91. ACM Press.

Hsu, K. C. and M. Z. Li (2011, APR). Techniques for finding similarity knowledge in olap reports. *Expert systems with applications 38*(4), 3743–3756.

Humphreys, P. W. (1989). *The chances of explanation*. New Jersey: Princeton University Press.

IBM Cognos Software (2012). Ibm cognos business intelligence, powerplay.

Inmon, W. H. (1996). *Building the data warehouse*. John Wiley.

Judd, P., C. Paddock, and J. Wetherbe (1981). Decision impelling differences: An investigation of management by exception reporting. *Information & Management 4*, 259–267.

Kimball, R. (1996). *The data warehouse toolkit: practical techniques for building dimensional data warehouses*. New York, NY, USA: John Wiley & Sons, Inc.

Knorr, E. M. and R. T. Ng (1998, 24–27). Algorithms for mining distance-based outliers in large datasets. In *Proc. 24th Int. Conf. Very Large Data Bases, VLDB*, pp. 392–403.

Kosy, D. W. (1989). Applications of explanation in financial modeling. pp. 487–509.

Kosy, D. W. and B. P. Wise (1984). Self-explanatory financial planning models. In *Proc. of AAAI-84*, Austin, TX, pp. 176–181.

Koutsoukis, N. S., G. Mitra, and C. Lucas (1999). Adapting on-line analytical processing for decision modelling: the interaction of information and decision technologies. *Decis. Support Syst. 26*(1), 1–30.

Kuznetsov, S. D. and Y. A. Kudryavtsev (2009, SEP). A mathematical model of the olap cubes. *Programming and Computer Software 35*(5), 257–265.

Lakshmanan, L., A. Russakovsky, and V. Sashikanth (2007). What if olap queries with changing dimensions perspectives are everything. In *VLDB '07: Proceedings of the 33th International Conference on Very Large Data Bases*, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lavrac, N., P. A. Flach, and B. Zupan (1999). Rule evaluation measures: A unifying view. In *ILP '99: Proceedings of the 9th International Workshop on Inductive Logic Programming*, London, UK, pp. 174–185. Springer-Verlag.

Lehner, W. (1998). Modelling large scale olap scenarios. In *EDBT '98: Proceedings of the 6th International Conference on Extending Database Technology*, London, UK, pp. 153–167. Springer-Verlag.

Lenz, H. J. and A. Shoshani (1997). Summarizability in OLAP and statistical data bases. In *Statistical and Scientific Database Management*, pp. 132–143.

Lin, S. and D. E. Brown (2003). Criminal incident data association using the olap technology. In H. Chen, R. Miranda, D. D. Zeng, C. C. Demchak, J. Schroeder, and T. Madhusudan (Eds.), *Intelligence and Security Informatics, First NSF/NIJ Symposium, ISI 2003, Tucson, AZ, USA, June 2-3, 2003, Proceedings*, Volume 2665 of *Lecture Notes in Computer Science*, pp. 13–26. Springer.

Lin, S. and D. E. Brown (2006). An outlier-based data association method for linking criminal incidents. *Decision Support Systems 41*(3), 604–615.

Lucas, P. J. F. (1997). Model-based diagnosis in medicine. *Artificial Intelligence in Medicine 10*(3), 201–208.

Meador, C. L., M. Guyote, and P. G. W. Keen (1984, June). Setting priorities for dss development. *MIS Quarterly 8*(2), 117–129.

Mendenhall, W., J. E. Reinmuth, and R. J. Beaver (1993). *Statistics for Management and Economics*. Duxbury Press.

Milgrom, P. and C. Shannon (1994, January). Monotone comparative statics. *Econometrica 62*(1), 157–80.

Miller, R. A., H. E. Pople, and J. D. Myers (1982). Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. *N Engl J Med 307*(8), 468–476.

Mintzberg, H., D. Raisinghani, and A. Théorêt (1976). The structure of "unstructured" decision processes. *Administrative Science Quarterly 21*(2), 246–275.

Mohammed, N. H. A., J. F. Courtney, and D. B. Paradice (1988). A prototype dss for structuring and diagnosing managerial problems. *IEEE Transactions on Systems, Man and Cybernetics 18*, 899–907.

Pannell, D. J. (1997, May). Sensitivity analysis of normative economic models: theoretical framework and practical strategies. *Agricultural Economics 16*(2), 139–152.

Pedersen, T. B., C. S. Jensen, and C. E. Dyreson (1999). Extending practical pre-aggregation in on-line analytical processing. In *VLDB '99: Proceedings of the 25th International Conference on Very Large Data Bases*, San Francisco, CA, USA, pp. 663–674. Morgan Kaufmann Publishers Inc.

Pedersen, T. B., C. S. Jensen, and C. E. Dyreson (2001). A foundation for capturing and querying complex multidimensional data. *Inf. Syst. 26*(5), 383–423.

Pendse, N. (2006). Olap report. http://www.olapreport.com.

Pounds, W. F. (1969). The process of problem finding. *Industrial Management Review 11*(1), 1–19.

PROTECT (2006). Transumo. http://protect.transumo.nl.

Pyle, D. (1999, March). *Data Preparation for Data Mining (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann.

Raisinghani, M. (2004). *Business Intellignce in the Digital Economy*. Hershey, PA: The Idea Group.

Reiter, R. (1987). A theory of diagnosis from first principles. *Artificial Intelligence 32*(1), 57–95.

Samuelson, P. A. (1941). The stability of equilibrium: Comparative statics and dynamics. *Econometrica 9*(2), 97–120.

SAP (2006). Sap corporation. http://www.sap.com.

Sarawagi, S. (2001). idiff: Informative summarization of differences in multidimensional aggregates. *Data Min. Knowl. Discov. 5*(4), 255–276.

Sarawagi, S., R. Agrawal, and N. Megiddo (1998). Discovery-driven exploration of olap data cubes. In *Conf. Proc. EDBT '98*, London, UK, pp. 168–182. Springer-Verlag.

Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.

Schott, J. (1997). *Matrix Analysis for Statistics*. New York: Wiley.

Shortliffe, E. H. (1976). *Computer-Based Medical Consultations:* MYCIN. New York: Elsevier.

Shoshani, A. (1997). Olap and statistical databases: similarities and differences. In *PODS '97: Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, New York, NY, USA, pp. 185–196. ACM Press.

Simon, H. A. (1960). *The New Science of Management Decision*. New York: Harper texbookBrothers.

Snedecor, G. W. and W. C.Cochran (1980). *Statistical Methods*. Iowa: Iowa State University Press.

Statistics Netherlands (2009). Centraal bureau voor de statistiek. http://www.cbs.nl.

Thalhammer, T., M. Schrefl, and M. Hohania (2001). Data & knowledge engineering. *Active data warehouses: complementing OLAP with analysis rules 39*, 241–269.

The R Foundation for statistical computing (2011). R statistical software.

Thomsen, E. (1997). *OLAP solutions: building multidimensional information systems*. New York, NY, USA: John Wiley & Sons, Inc.

Thomsen, E. (2002). *OLAP solutions: building multi-dimensional information systems*. New York: John Wiley & Sons, Inc.

Tukey, J. W. (1949). One degree of freedom for nonadditivity. *Biometrics 5*, 232–242.

Turban, E., J. E. Aronson, T. P. Liang, and R. Sharda (2007). *Decision Support and Business Intelligence Systems*. New Jersey, USA: Pearson, Prentice Hall.

Usman, M., R. Pears, and A. C. M. Fong (2013a, MAR). A data mining approach to knowledge discovery from multidimensional cube structures. *Knowledge-based systems 40*, 36–49.

Usman, M., R. Pears, and A. C. M. Fong (2013b, NOV 1). Discovering diverse association rules from multidimensional schema. *Expert systems with applications 40*(15), 5975–5996.

van Buuren, S., E. V. van Mulligen, and J. P. L. Brand (1994). Routine multiple imputation in statistical databases. In *Proceedings of the Seventh International Working Conference on Scientific and Statistical Database Management*, Washington, DC, USA, pp. 74–78. IEEE Computer Society.

Vassiliadis, P. (1998). Modeling multidimensional databases, cubes and cube operations. In M. Rafanelli and M. Jarke (Eds.), *10th International Conference on Scientific and Statistical Database Management, Proceedings, Capri, Italy, July 1-3, 1998*, pp. 53–62. IEEE Computer Society.

Vassiliadis, P. and T. K. Sellis (1999). A survey of logical models for OLAP databases. *SIGMOD Record 28*(4), 64–69.

Verkooijen, W. J. (1993). Automated financial diagnosis: a comparison with other diagnostic domains. *J. Inf. Sci. 19*(2), 125–135.

Weld, D. S. and J. de Kleer (1990). *Readings in Qualitative Reasoning about Physical Systems*. San Mateo, CA: Morgan Kaufmann.

Wierenga, B. and G. van Bruggen (2001, may-june). Developing a customized decision-support system for managers. *Interfaces 31*(3), 128–145.

Williams, B. C. and J. de Kleer (1991). Qualitative reasoning about physical systems: a return to roots. *Artif. Intell. 51*(1-3), 1–9.

Curriculum Vitae

Emiel Caron was born in Dongen on January 21, 1975. He obtained his master's degree in Information Management, with a specialization in Information Technology (IT), from Tilburg University in 2000. In the same year he started to work as a business intelligence (BI) consultant for Pink-Roccade (now KPN) on data warehousing and BI projects. In 2002, he started his PhD research at the Decision & Information Sciences department of the Rotterdam School of Management. Later he joined the former department of Economics & Informatics at the Rotterdam School of Economics as an assistant professor and lecturer in business information systems. Emiel's PhD research is in the field of business intelligence and analytics and focuses on the explanation of exceptional values in multi-dimensional databases. In his PhD trajectory he received two research grants from the Netherlands Organisation for Scientific Research to work with the Natural Computing Research & Applications Group at University College Dublin. Results of his research have been presented at various international conferences and have been published in various conference proceedings and journals (*European Journal of Operational Research*, *Intelligent Systems in Accounting, Finance and Management*, and *International Journal of Business Intelligence Research*). He is also a member of the Erasmus Centre for Business Intelligence.

Since 2012, Emiel has been working as an IT developer and researcher at the Centre for Science and Technology Studies of Leiden University. He has become interested in the application of business intelligence and data mining techniques on large scientometric and bibliometric databases. Moreover, he teaches the course business intelligence in the master ICT in Business of the Leiden Institute of Advanced Computer Science and supervises several MSc students.

For more information, please visit http://www.emielcaron.nl.

ERASMUS RESEARCH INSTITUTE OF MANAGEMENT (ERIM)

ERIM PH.D. SERIES RESEARCH IN MANAGEMENT

The ERIM PhD Series contains PhD dissertations in the field of Research in Management defended at Erasmus University Rotterdam and supervised by senior researchers affiliated to the Erasmus Research Institute of Management (ERIM). All dissertations in the ERIM PhD Series are available in full text through the ERIM Electronic Series Portal: http://hdl.handle.net/1765/1 ERIM is the joint research institute of the Rotterdam School of Management (RSM) and the Erasmus School of Economics at the Erasmus University Rotterdam (EUR).

DISSERTATIONS LAST FIVE YEARS

Acciaro, M., *Bundling Strategies in Global Supply Chains*. Promoter(s): Prof.dr. H.E. Haralambides, EPS-2010-197-LIS, http://hdl.handle.net/1765/19742

Agatz, N.A.H., *Demand Management in E-Fulfillment*. Promoter(s): Prof.dr.ir. J.A.E.E. van Nunen, EPS-2009-163-LIS, http://hdl.handle.net/1765/15425

Alexiev, A., *Exploratory Innovation: The Role of Organizational and Top Management Team Social Capital*. Promoter(s): Prof.dr. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-208-STR, http://hdl.handle.net/1765/20632

Asperen, E. van, *Essays on Port, Container, and Bulk Chemical Logistics Optimization*. Promoter(s): Prof.dr.ir. R. Dekker, EPS-2009-181-LIS, http://hdl.handle.net/1765/17626

Bannouh, K., *Measuring and Forecasting Financial Market Volatility using High-Frequency Data*, Promoter(s): Prof.dr.D.J.C. van Dijk, EPS-2013-273-F&A, http://hdl.handle.net/1765/38240

Benning, T.M., *A Consumer Perspective on Flexibility in Health Care: Priority Access Pricing and Customized Care*, Promoter(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2011-241-MKT, http://hdl.handle.net/1765/23670

Ben-Menahem, S.M., *Strategic Timing and Proactiveness of Organizations*, Promoter(s): Prof.dr. H.W. Volberda & Prof.dr.ing. F.A.J. van den Bosch, EPS-2013-278-S&E, http://hdl.handle.net/1765/ 39128

Betancourt, N.E., *Typical Atypicality: Formal and Informal Institutional Conformity, Deviance, and Dynamics*, Promoter(s): Prof.dr. B. Krug, EPS-2012-262-ORG, http://hdl.handle.net/1765/32345

Bezemer, P.J., *Diffusion of Corporate Governance Beliefs: Board Independence and the Emergence of a Shareholder Value Orientation in the Netherlands*. Promoter(s): Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2009-192-STR, http://hdl.handle.net/1765/18458

Binken, J.L.G., *System Markets: Indirect Network Effects in Action, or Inaction*, Promoter(s): Prof.dr. S. Stremersch, EPS-2010-213-MKT, http://hdl.handle.net/1765/21186

Blitz, D.C., *Benchmarking Benchmarks*, Promoter(s): Prof.dr. A.G.Z. Kemna & Prof.dr. W.F.C. Verschoor, EPS-2011-225-F&A, http://hdl.handle.net/1765/22624

Borst, W.A.M., *Understanding Crowdsourcing: Effects of Motivation and Rewards on Participation and Performance in Voluntary Online Activities*, Promoter(s): Prof.dr.ir. J.C.M. van den Ende & Prof.dr.ir. H.W.G.M. van Heck, EPS-2010-221-LIS, http://hdl.handle.net/1765/ 21914

Budiono, D.P., *The Analysis of Mutual Fund Performance: Evidence from U.S. Equity Mutual Funds*, Promoter(s): Prof.dr. M.J.C.M. Verbeek, EPS-2010-185-F&A, http://hdl.handle.net/1765/18126

Burger, M.J., *Structure and Cooptition in Urban Networks*, Promoter(s): Prof.dr. G.A. van der Knaap & Prof.dr. H.R. Commandeur, EPS-2011-243-ORG, http://hdl.handle.net/1765/26178

Byington, E., *Exploring Coworker Relationships: Antecedents and Dimensions of Interpersonal Fit, Coworker Satisfaction, and Relational Models*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2013-292-ORG, http://hdl.handle.net/1765/1

Camacho, N.M., *Health and Marketing; Essays on Physician and Patient Decision-making*, Promoter(s): Prof.dr. S. Stremersch, EPS-2011-237-MKT, http://hdl.handle.net/1765/23604

Carvalho, L., *Knowledge Locations in Cities; Emergence and Development Dynamics*, Promoter(s): Prof.dr. L. van den Berg, EPS-2013-274-S&E, http://hdl.handle.net/1765/ 38449

Carvalho de Mesquita Ferreira, L., *Attention Mosaics: Studies of Organizational Attention*, Promoter(s): Prof.dr. P.M.A.R. Heugens & Prof.dr. J. van Oosterhout, EPS-2010-205-ORG, http://hdl.handle.net/1765/19882

Chen, C.M., *Evaluation and Design of Supply Chain Operations Using DEA*, Promoter(s): Prof.dr. J.A.E.E. van Nunen, EPS-2009-172-LIS, http://hdl.handle.net/1765/16181

Cox, R.H.G.M., *To Own, To Finance, and to Insure; Residential Real Estate Revealed*, Promoter(s): Prof.dr. D. Brounen, EPS-2013-290-F&A, http://hdl.handle.net/1765/1

Defilippi Angeldonis, E.F., *Access Regulation for Naturally Monopolistic Port Terminals: Lessons from Regulated Network Industries*, Promoter(s): Prof.dr. H.E. Haralambides, EPS-2010-204-LIS, http://hdl.handle.net/1765/19881

Deichmann, D., *Idea Management: Perspectives from Leadership, Learning, and Network Theory*, Promoter(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2012-255-ORG, http://hdl.handle.net/1765/ 31174

Desmet, P.T.M., *In Money we Trust? Trust Repair and the Psychology of Financial Compensations*, Promoter(s): Prof.dr. D. De Cremer & Prof.dr. E. van Dijk, EPS-2011-232-ORG, http://hdl.handle.net/1765/23268

Diepen, M. van, *Dynamics and Competition in Charitable Giving*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2009-159-MKT, http://hdl.handle.net/1765/14526

Dietvorst, R.C., *Neural Mechanisms Underlying Social Intelligence and Their Relationship with the Performance of Sales Managers*, Promoter(s): Prof.dr. W.J.M.I. Verbeke, EPS-2010-215-MKT, http://hdl.handle.net/1765/21188

Dietz, H.M.S., *Managing (Sales)People towards Performance: HR Strategy, Leadership & Teamwork*, Promoter(s): Prof.dr. G.W.J. Hendrikse, EPS-2009-168-ORG, http://hdl.handle.net/1765/16081

Dollevoet, T.A.B., *Delay Management and Dispatching in Railways*, Promoter(s): Prof.dr. A.P.M. Wagelmans, EPS-2013-272-LIS, http://hdl.handle.net/1765/38241

Doorn, S. van, *Managing Entrepreneurial Orientation*, Promoter(s): Prof.dr. J.J.P. Jansen, Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2012-258-STR, http://hdl.handle.net/1765/32166

Douwens-Zonneveld, M.G., *Animal Spirits and Extreme Confidence: No Guts, No Glory*, Promoter(s): Prof.dr. W.F.C. Verschoor, EPS-2012-257-F&A, http://hdl.handle.net/1765/31914

Duca, E., *The Impact of Investor Demand on Security Offerings*, Promoter(s): Prof.dr. A. de Jong, EPS-2011-240-F&A, http://hdl.handle.net/1765/26041

Duursema, H., *Strategic Leadership; Moving Beyond the Leader-follower Dyad*, Promoter(s): Prof.dr. R.J.M. van Tulder, EPS-2013-279-ORG, http://hdl.handle.net/1765/ 39129

Eck, N.J. van, *Methodological Advances in Bibliometric Mapping of Science*, Promoter(s): Prof.dr.ir. R. Dekker, EPS-2011-247-LIS, http://hdl.handle.net/1765/26509

Eijk, A.R. van der, *Behind Networks: Knowledge Transfer, Favor Exchange and Performance*, Promoter(s): Prof.dr. S.L. van de Velde & Prof.dr.drs. W.A. Dolfsma, EPS-2009-161-LIS, http://hdl.handle.net/1765/14613

Essen, M. van, *An Institution-Based View of Ownership*, Promoter(s): Prof.dr. J. van Oosterhout & Prof.dr. G.M.H. Mertens, EPS-2011-226-ORG, http://hdl.handle.net/1765/22643

Feng, L., *Motivation, Coordination and Cognition in Cooperatives*, Promoter(s): Prof.dr. G.W.J. Hendrikse, EPS-2010-220-ORG, http://hdl.handle.net/1765/21680

Gertsen, H.F.M., *Riding a Tiger without Being Eaten: How Companies and Analysts Tame Financial Restatements and Influence Corporate Reputation*, Promoter(s): Prof.dr. C.B.M. van Riel, EPS-2009-171-ORG, http://hdl.handle.net/1765/16098

Gharehgozli, A.H., *Developing New Methods for Efficient Container Stacking Operations*, Promoter(s): Prof.dr.ir. M.B.M. de Koster, EPS-2012-269-LIS, http://hdl.handle.net/1765/ 37779

Gijsbers, G.W., *Agricultural Innovation in Asia: Drivers, Paradigms and Performance*, Promoter(s): Prof.dr. R.J.M. van Tulder, EPS-2009-156-ORG, http://hdl.handle.net/1765/14524

Gils, S. van, *Morality in Interactions: On the Display of Moral Behavior by Leaders and Employees*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2012-270-ORG, http://hdl.handle.net/1765/ 38028

Ginkel-Bieshaar, M.N.G. van, *The Impact of Abstract versus Concrete Product Communications on Consumer Decision-making Processes*, Promoter(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2012-256-MKT, http://hdl.handle.net/1765/31913

Gkougkousi, X., *Empirical Studies in Financial Accounting*, Promoter(s): Prof.dr. G.M.H. Mertens & Prof.dr. E. Peek, EPS-2012-264-F&A, http://hdl.handle.net/1765/37170

Gong, Y., *Stochastic Modelling and Analysis of Warehouse Operations*, Promoter(s): Prof.dr. M.B.M. de Koster & Prof.dr. S.L. van de Velde, EPS-2009-180-LIS, http://hdl.handle.net/1765/16724

Greeven, M.J., *Innovation in an Uncertain Institutional Environment: Private Software Entrepreneurs in Hangzhou*, China, Promoter(s): Prof.dr. B. Krug, EPS-2009-164-ORG, http://hdl.handle.net/1765/15426

Hakimi, N.A, *Leader Empowering Behaviour: The Leaders Perspective: Understanding the Motivation behind Leader Empowering Behaviour*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2010-184-ORG, http://hdl.handle.net/1765/17701

Hensmans, M., *A Republican Settlement Theory of the Firm: Applied to Retail Banks in England and the Netherlands (1830-2007)*, Promoter(s): Prof.dr. A. Jolink & Prof.dr. S.J. Magala, EPS-2010-193-ORG, http://hdl.handle.net/1765/19494

Hernandez Mireles, C., *Marketing Modeling for New Products*, Promoter(s): Prof.dr. P.H. Franses, EPS-2010-202-MKT, http://hdl.handle.net/1765/19878

Heyden, M.L.M., *Essays on Upper Echelons & Strategic Renewal: A Multilevel Contingency Approach*, Promoter(s): Prof.dr. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2012-259-STR, http://hdl.handle.net/1765/32167

Hoever, I.J., *Diversity and Creativity: In Search of Synergy*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2012-267-ORG, http://hdl.handle.net/1765/37392

Hoogendoorn, B., *Social Entrepreneurship in the Modern Economy: Warm Glow, Cold Feet*, Promoter(s): Prof.dr. H.P.G. Pennings & Prof.dr. A.R. Thurik, EPS-2011-246-STR, http://hdl.handle.net/1765/26447

Hoogervorst, N., *On The Psychology of Displaying Ethical Leadership: A Behavioral Ethics Approach*, Promoter(s): Prof.dr. D. De Cremer & Dr. M. van Dijke, EPS-2011-244-ORG, http://hdl.handle.net/1765/26228

Huang, X., *An Analysis of Occupational Pension Provision: From Evaluation to Redesign*, Promoter(s): Prof.dr. M.J.C.M. Verbeek & Prof.dr. R.J. Mahieu, EPS-2010-196-F&A, http://hdl.handle.net/1765/19674

Hytnen, K.A. *Context Effects in Valuation, Judgment and Choice*, Promoter(s): Prof.dr.ir. A. Smidts, EPS-2011-252-MKT, http://hdl.handle.net/1765/30668

Jaarsveld, W.L. van, *Maintenance Centered Service Parts Inventory Control*, Promoter(s): Prof.dr.ir. R. Dekker, EPS-2013-288-LIS, http://hdl.handle.net/1765/ 39933

Jalil, M.N., *Customer Information Driven After Sales Service Management: Lessons from Spare Parts Logistics*, Promoter(s): Prof.dr. L.G. Kroon, EPS-2011-222-LIS, http://hdl.handle.net/1765/22156

Jaspers, F.P.H., *Organizing Systemic Innovation*, Promoter(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2009-160-ORG, http://hdl.handle.net/1765/14974

Jiang, T., *Capital Structure Determinants and Governance Structure Variety in Franchising*, Promoter(s): Prof.dr. G. Hendrikse & Prof.dr. A. de Jong, EPS-2009-158-F&A, http://hdl.handle.net/1765/14975

Jiao, T., *Essays in Financial Accounting*, Promoter(s): Prof.dr. G.M.H. Mertens, EPS-2009-176-F&A, http://hdl.handle.net/1765/16097

Kaa, G. van, *Standard Battles for Complex Systems: Empirical Research on the Home Network*, Promoter(s): Prof.dr.ir. J. van den Ende & Prof.dr.ir. H.W.G.M. van Heck, EPS-2009-166-ORG, http://hdl.handle.net/1765/16011

Kagie, M., *Advances in Online Shopping Interfaces: Product Catalog Maps and Recommender Systems*, Promoter(s): Prof.dr. P.J.F. Groenen, EPS-2010-195-MKT, http://hdl.handle.net/1765/19532

Kappe, E.R., *The Effectiveness of Pharmaceutical Marketing*, Promoter(s): Prof.dr. S. Stremersch, EPS-2011-239-MKT, http://hdl.handle.net/1765/23610

Karreman, B., *Financial Services and Emerging Markets*, Promoter(s): Prof.dr. G.A. van der Knaap & Prof.dr. H.P.G. Pennings, EPS-2011-223-ORG, http://hdl.handle.net/1765/ 22280

Kwee, Z., *Investigating Three Key Principles of Sustained Strategic Renewal: A Longitudinal Study of Long-Lived Firms*, Promoter(s): Prof.dr.ir. F.A.J. Van den Bosch & Prof.dr. H.W. Volberda, EPS-2009-174-STR, http://hdl.handle.net/1765/16207

Lam, K.Y., *Reliability and Rankings*, Promoter(s): Prof.dr. P.H.B.F. Franses, EPS-2011-230-MKT, http://hdl.handle.net/1765/22977

Lander, M.W., *Profits or Professionalism? On Designing Professional Service Firms*, Promoter(s): Prof.dr. J. van Oosterhout & Prof.dr. P.P.M.A.R. Heugens, EPS-2012-253-ORG, http://hdl.handle.net/1765/30682

Langhe, B. de, *Contingencies: Learning Numerical and Emotional Associations in an Uncertain World*, Promoter(s): Prof.dr.ir. B. Wierenga & Prof.dr. S.M.J. van Osselaer, EPS-2011-236-MKT,

http://hdl.handle.net/1765/23504

Larco Martinelli, J.A., *Incorporating Worker-Specific Factors in Operations Management Models*, Promoter(s): Prof.dr.ir. J. Dul & Prof.dr. M.B.M. de Koster, EPS-2010-217-LIS, http://hdl.handle.net/1765/21527

Li, T., *Informedness and Customer-Centric Revenue Management*, Promoter(s): Prof.dr. P.H.M. Vervest & Prof.dr.ir. H.W.G.M. van Heck, EPS-2009-146-LIS, http://hdl.handle.net/1765/14525

Liang, Q., *Governance, CEO Indentity, and Quality Provision of Farmer Cooperatives*, Promoter(s): Prof.dr. G.W.J. Hendrikse, EPS-2013-281-ORG, http://hdl.handle.net/1765/1

Loos, M.J.H.M. van der, *Molecular Genetics and Hormones; New Frontiers in Entrepreneurship Research*, Promoter(s): Prof.dr. A.R. Thurik, Prof.dr. P.J.F. Groenen & Prof.dr. A. Hofman, EPS-2013-287-S&E, http://hdl.handle.net/1765/ 40081

Lovric, M., *Behavioral Finance and Agent-Based Artificial Markets*, Promoter(s): Prof.dr. J. Spronk & Prof.dr.ir. U. Kaymak, EPS-2011-229-F&A, http://hdl.handle.net/1765/ 22814

Maas, K.E.G., *Corporate Social Performance: From Output Measurement to Impact Measurement*, Promoter(s): Prof.dr. H.R. Commandeur, EPS-2009-182-STR, http://hdl.handle.net/1765/17627

Markwat, T.D., *Extreme Dependence in Asset Markets Around the Globe*, Promoter(s): Prof.dr. D.J.C. van Dijk, EPS-2011-227-F&A, http://hdl.handle.net/1765/22744

Mees, H., *Changing Fortunes: How Chinas Boom Caused the Financial Crisis*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2012-266-MKT, http://hdl.handle.net/1765/34930

Meuer, J., *Configurations of Inter-Firm Relations in Management Innovation*: A Study in Chinas Biopharmaceutical Industry, Promoter(s): Prof.dr. B. Krug, EPS-2011-228-ORG, http://hdl.handle.net/1765/22745

Mihalache, O.R., *Stimulating Firm Innovativeness: Probing the Interrelations between Managerial and Organizational Determinants*, Promoter(s): Prof.dr. J.J.P. Jansen, Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2012-260-S&E, http://hdl.handle.net/1765/32343

Milea, V., *New Analytics for Financial Decision Support*, Promoter(s): Prof.dr.ir. U. Kaymak, EPS-2013-275-LIS, http://hdl.handle.net/1765/ 38673

Moonen, J.M., *Multi-Agent Systems for Transportation Planning and Coordination*, Promoter(s): Prof.dr. J. van Hillegersberg & Prof.dr. S.L. van de Velde, EPS-2009-177-LIS, http://hdl.handle.net/1765/16208

Nederveen Pieterse, A., *Goal Orientation in Teams: The Role of Diversity*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2009-162-ORG, http://hdl.handle.net/1765/15240

Nielsen, L.K., *Rolling Stock Rescheduling in Passenger Railways: Applications in Short-term Planning and in Disruption Management*, Promoter(s): Prof.dr. L.G. Kroon, EPS-2011-224-LIS,

http://hdl.handle.net/1765/22444

Niesten, E.M.M.I., *Regulation, Governance and Adaptation: Governance Transformations in the Dutch and French Liberalizing Electricity Industries*, Promoter(s): Prof.dr. A. Jolink & Prof.dr. J.P.M. Groenewegen, EPS-2009-170-ORG, http://hdl.handle.net/1765/16096

Nijdam, M.H., *Leader Firms: The Value of Companies for the Competitiveness of the Rotterdam Seaport Cluster*, Promoter(s): Prof.dr. R.J.M. van Tulder, EPS-2010-216-ORG, http://hdl.handle.net/1765/21405

Noordegraaf-Eelens, L.H.J., *Contested Communication: A Critical Analysis of Central Bank Speech*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2010-209-MKT, http://hdl.handle.net/1765/21061

Nuijten, A.L.P., *Deaf Effect for Risk Warnings: A Causal Examination applied to Information Systems Projects*, Promoter(s): Prof.dr. G. van der Pijl & Prof.dr. H. Commandeur & Prof.dr. M. Keil, EPS-2012-263-S&E, http://hdl.handle.net/1765/34928

Nuijten, I., *Servant Leadership: Paradox or Diamond in the Rough? A Multidimensional Measure and Empirical Evidence*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2009-183-ORG, http://hdl.handle.net/1765/21405

Oosterhout, M., van, *Business Agility and Information Technology in Service Organizations*, Promoter(s): Prof,dr.ir. H.W.G.M. van Heck, EPS-2010-198-LIS, http://hdl.handle.net/1765/19805

Oostrum, J.M., van, *Applying Mathematical Models to Surgical Patient Planning*, Promoter(s): Prof.dr. A.P.M. Wagelmans, EPS-2009-179-LIS, http://hdl.handle.net/1765/16728

Osadchiy, S.E., *The Dynamics of Formal Organization: Essays on Bureaucracy and Formal Rules*, Promoter(s): Prof.dr. P.P.M.A.R. Heugens, EPS-2011-231-ORG, http://hdl.handle.net/1765/23250

Otgaar, A.H.J., *Industrial Tourism: Where the Public Meets the Private*, Promoter(s): Prof.dr. L. van den Berg, EPS-2010-219-ORG, http://hdl.handle.net/1765/21585

Ozdemir, M.N., *Project-level Governance, Monetary Incentives and Performance in Strategic R&D Alliances*, Promoter(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2011-235-LIS, http://hdl.handle.net/1765/23550

Peers, Y., *Econometric Advances in Diffusion Models*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2011-251-MKT, http://hdl.handle.net/1765/ 30586

Pince, C., *Advances in Inventory Management: Dynamic Models*, Promoter(s): Prof.dr.ir. R. Dekker, EPS-2010-199-LIS, http://hdl.handle.net/1765/19867

Porras Prado, M., *The Long and Short Side of Real Estate, Real Estate Stocks, and Equity*, Promoter(s): Prof.dr. M.J.C.M. Verbeek, EPS-2012-254-F&A, http://hdl.handle.net/1765/30848

Potthoff, D., *Railway Crew Rescheduling: Novel Approaches and Extensions*, Promoter(s): Prof.dr. A.P.M. Wagelmans & Prof.dr. L.G. Kroon, EPS-2010-210-LIS, http://hdl.handle.net/1765/21084

Poruthiyil, P.V., *Steering Through: How Organizations Negotiate Permanent Uncertainty and Unresolvable Choices*, Promoter(s): Prof.dr. P.P.M.A.R. Heugens & Prof.dr. S. Magala, EPS-2011-245-ORG, http://hdl.handle.net/1765/26392

Pourakbar, M. *End-of-Life Inventory Decisions of Service Parts*, Promoter(s): Prof.dr.ir. R. Dekker, EPS-2011-249-LIS, http://hdl.handle.net/1765/30584

Pronker, E.S., *Innovation Paradox in Vaccine Target Selection*, Promoter(s): Prof.dr. H.R. Commandeur & Prof.dr. H.J.H.M. Claassen, EPS-2013-282-S&E, http://hdl.handle.net/1765/1

Rijsenbilt, J.A., *CEO Narcissism; Measurement and Impact*, Promoter(s): Prof.dr. A.G.Z. Kemna & Prof.dr. H.R. Commandeur, EPS-2011-238-STR, http://hdl.handle.net/1765/ 23554

Roelofsen, E.M., *The Role of Analyst Conference Calls in Capital Markets*, Promoter(s): Prof.dr. G.M.H. Mertens & Prof.dr. L.G. van der Tas RA, EPS-2010-190-F&A, http://hdl.handle.net/1765/18013

Rosmalen, J. van, *Segmentation and Dimension Reduction: Exploratory and Model-Based Approaches*, Promoter(s): Prof.dr. P.J.F. Groenen, EPS-2009-165-MKT, http://hdl.handle.net/1765/15536

Roza, M.W., *The Relationship between Offshoring Strategies and Firm Performance: Impact of Innovation, Absorptive Capacity and Firm Size*, Promoter(s): Prof.dr. H.W. Volberda & Prof.dr.ing. F.A.J. van den Bosch, EPS-2011-214-STR, http://hdl.handle.net/1765/22155

Rubbaniy, G., *Investment Behavior of Institutional Investors*, Promoter(s): Prof.dr. W.F.C. Vershoor, EPS-2013-284-F&A, http://hdl.handle.net/1765/ 40068

Rus, D., *The Dark Side of Leadership: Exploring the Psychology of Leader Self-serving Behavior*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2009-178-ORG, http://hdl.handle.net/1765/16726

Schellekens, G.A.C., *Language Abstraction in Word of Mouth*, Promoter(s): Prof.dr.ir. A. Smidts, EPS-2010-218-MKT, http://hdl.handle.net/1765/21580

Shahzad, K., *Credit Rating Agencies, Financial Regulations and the Capital Markets*, Promoter(s): Prof.dr. G.M.H. Mertens, EPS-2013-283-F&A, http://hdl.handle.net/1765/39655

Sotgiu, F., *Not All Promotions are Made Equal: From the Effects of a Price War to Cross-chain Cannibalization*, Promoter(s): Prof.dr. M.G. Dekimpe & Prof.dr.ir. B. Wierenga, EPS-2010-203-MKT, http://hdl.handle.net/1765/19714

Spliet, R., *Vehicle Routing with Uncertain Demand*, Promoter(s): Prof.dr.ir. R. Dekker, EPS-2013-293-LIS, http://hdl.handle.net/1765/1

Srour, F.J., *Dissecting Drayage: An Examination of Structure, Information, and Control in Drayage Operations*, Promoter(s): Prof.dr. S.L. van de Velde, EPS-2010-186-LIS, http://hdl.handle.net/1765/18231

Stallen, M., *Social Context Effects on Decision-Making; A Neurobiological Approach*, Promoter(s): Prof.dr.ir. A. Smidts, EPS-2013-285-MKT, http://hdl.handle.net/1765/ 39931

Sweldens, *S.T.L.R., Evaluative Conditioning 2.0: Direct versus Associative Transfer of Affect to Brands*, Promoter(s): Prof.dr. S.M.J. van Osselaer, EPS-2009-167-MKT, http://hdl.handle.net/1765/16012

Tarakci, M., *Behavioral Strategy; Strategic Consensus, Power and Networks*, Promoter(s): Prof.dr. P.J.F. Groenen & Prof.dr. D.L. van Knippenberg, EPS-2013-280-ORG, http://hdl.handle.net/1765/ 39130

Teixeira de Vasconcelos, M., *Agency Costs, Firm Value, and Corporate Investment*, Promoter(s): Prof.dr. P.G.J. Roosenboom, EPS-2012-265-F&A, http://hdl.handle.net/1765/37265

Tempelaar, M.P., *Organizing for Ambidexterity: Studies on the Pursuit of Exploration and Exploitation through Differentiation*, Integration, Contextual and Individual Attributes, Promoter(s): Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-191-STR, http://hdl.handle.net/1765/18457

Tiwari, V., *Transition Process and Performance in IT Outsourcing: Evidence from a Field Study and Laboratory Experiments*, Promoter(s): Prof.dr.ir. H.W.G.M. van Heck & Prof.dr. P.H.M. Vervest, EPS-2010-201-LIS, http://hdl.handle.net/1765/19868

Trster, C., *Nationality Heterogeneity and Interpersonal Relationships at Work*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2011-233-ORG, http://hdl.handle.net/1765/23298

Tsekouras, D., *No Pain No Gain: The Beneficial Role of Consumer Effort in Decision Making*, Promoter(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2012-268-MKT, http://hdl.handle.net/1765/ 37542

Tzioti, S., *Let Me Give You a Piece of Advice: Empirical Papers about Advice Taking in Marketing*, Promoter(s): Prof.dr. S.M.J. van Osselaer & Prof.dr.ir. B. Wierenga, EPS-2010-211-MKT, hdl.handle.net/1765/21149

Vaccaro, I.G., *Management Innovation: Studies on the Role of Internal Change Agents*, Promoter(s): Prof.dr. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-212-STR, hdl.handle.net/1765/21150

Verheijen, H.J.J., *Vendor-Buyer Coordination in Supply Chains*, Promoter(s): Prof.dr.ir. J.A.E.E. van Nunen, EPS-2010-194-LIS, http://hdl.handle.net/1765/19594

Venus, M., *Demystifying Visionary Leadership; In Search of the Essence of Effective Vision Communication*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2013-289-ORG, http://hdl.handle.net/1765/ 40079

Verwijmeren, P., *Empirical Essays on Debt, Equity, and Convertible Securities*, Promoter(s): Prof.dr. A. de Jong & Prof.dr. M.J.C.M. Verbeek, EPS-2009-154-F&A, http://hdl.handle.net/1765/14312

Visser, V., *Leader Affect and Leader Effectiveness; How Leader Affective Displays Influence Follower Outcomes*, Promoter(s): Prof.dr. D. van Knippenberg, EPS-2013-286-ORG, http://hdl.handle.net/1765/40076

Vlam, A.J., *Customer First? The Relationship between Advisors and Consumers of Financial Products*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2011-250-MKT, http://hdl.handle.net/1765/30585

Waard, E.J. de, *Engaging Environmental Turbulence: Organizational Determinants for Repetitive Quick and Adequate Responses*, Promoter(s): Prof.dr. H.W. Volberda & Prof.dr. J. Soeters, EPS-2010-189-STR, http://hdl.handle.net/1765/18012

Wall, R.S., *Netscape: Cities and Global Corporate Networks*, Promoter(s): Prof.dr. G.A. van der Knaap, EPS-2009-169-ORG, http://hdl.handle.net/1765/16013

Waltman, L., *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*, Promoter(s): Prof.dr.ir. R. Dekker & Prof.dr.ir. U. Kaymak, EPS-2011-248-LIS, http://hdl.handle.net/1765/26564

Wang, Y., *Information Content of Mutual Fund Portfolio Disclosure*, Promoter(s): Prof.dr. M.J.C.M. Verbeek, EPS-2011-242-F&A, http://hdl.handle.net/1765/26066

Wang, Y., *Corporate Reputation Management; Reaching Out to Find Stakeholders*, Promoter(s): Prof.dr. C.B.M. van Riel, EPS-2013-271-ORG, http://hdl.handle.net/1765/ 38675

Weerdt, N.P. van der, *Organizational Flexibility for Hypercompetitive Markets*: Empirical Evidence of the Composition and Context Specificity of Dynamic Capabilities and Organization Design Parameters, Promoter(s): Prof.dr. H.W. Volberda, EPS-2009-173-STR, http://hdl.handle.net/1765/16182

Wolfswinkel, M., *Corporate Governance, Firm Risk and Shareholder Value of Dutch Firms*, Promoter(s): Prof.dr. A. de Jong, EPS-2013-277-F&A, http://hdl.handle.net/1765/ 39127

Wubben, M.J.J., *Social Functions of Emotions in Social Dilemmas*, Promoter(s): Prof.dr. D. De Cremer & Prof.dr. E. van Dijk, EPS-2009-187-ORG, http://hdl.handle.net/1765/18228

Xu, Y., *Empirical Essays on the Stock Returns, Risk Management, and Liquidity Creation of Banks*, Promoter(s): Prof.dr. M.J.C.M. Verbeek, EPS-2010-188-F&A, http://hdl.handle.net/1765/18125

Yang, J., *Towards the Restructuring and Co-ordination Mechanisms for the Architecture of Chinese Transport Logistics*, Promoter(s): Prof.dr. H.E. Harlambides, EPS-2009-157-LIS, http://hdl.handle.net/1765/14527

Zaerpour, N., *Efficient Management of Compact Storage Systems*, Promoter(s): Prof.dr. M.B.M. de Koster, EPS-2013-276-LIS, http://hdl.handle.net/1765/1

Zhang, D., *Essays in Executive Compensation*, Promoter(s): Prof.dr. I. Dittmann, EPS-2012-261-F&A, http://hdl.handle.net/1765/32344

Zhang, X., *Scheduling with Time Lags*, Promoter(s): Prof.dr. S.L. van de Velde, EPS-2010-206-LIS, http://hdl.handle.net/1765/19928

Zhou, H., *Knowledge, Entrepreneurship and Performance: Evidence from Country-level and Firm-level Studies*, Promoter(s): Prof.dr. A.R. Thurik & Prof.dr. L.M. Uhlaner, EPS-2010-207-ORG, http://hdl.handle.net/1765/20634

Zwan, P.W. van der, T*he Entrepreneurial Process: An International Analysis of Entry and Exit*, Promoter(s): Prof.dr. A.R. Thurik & Prof.dr. P.J.F. Groenen, EPS-2011-234-ORG, http://hdl.handle.net/1765/23422

EXPLANATION OF EXCEPTIONAL VALUES IN MULTI-DIMENSIONAL BUSINESS DATABASES

Multi-dimensional or OnLine Analytical Processing (OLAP) databases are a popular business intelligence (BI) technique in the field of enterprise information systems for business analytics and decision support. In this dissertation, OLAP database functionality is extended to support the business analyst in the exploration of OLAP data. The database is augmented with novel functionality for the detection of exceptional values, explanation generation, and sensitivity analysis. We describe how exceptional values at any level in the data, can be automatically detected by statistical and managerial models. It is also shown how exceptional values can be explained by underlying causes. This is realized by a generic model for diagnosis of atypical values. By applying it, a full explanation tree of causes at successive levels can be generated. If the tree is too large, the analyst can use appropriate filtering measures to prune the tree to a manageable size. The purpose of the methods and algorithms presented here, is to provide OLAP databases with more powerful explanatory analytics and reporting functions. This methodology has a wide range of applications, such as variance analysis in accounting, competition benchmarking, analysis of sales and financial data, and the analysis of any other data that possess a multi-dimensional hierarchical structure. The method is demonstrated in several case studies. For example, the explanatory analysis of a sales data cube is discussed, and computerized competition benchmarking with financial data about Dutch retail companies is illustrated.

ERiM

The Erasmus Research Institute of Management (ERIM) is the Research School (Onderzoekschool) in the field of management of the Erasmus University Rotterdam. The founding participants of ERIM are the Rotterdam School of Management (RSM), and the Erasmus School of Economics (ESE). ERIM was founded in 1999 and is officially accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW). The research undertaken by ERIM is focused on the management of the firm in its environment, its intra- and interfirm relations, and its business processes in their interdependent connections.

The objective of ERIM is to carry out first rate research in management, and to offer an advanced doctoral programme in Research in Management. Within ERIM, over three hundred senior researchers and PhD candidates are active in the different research programmes. From a variety of academic backgrounds and expertises, the ERIM community is united in striving for excellence and working at the forefront of creating new business knowledge.

Erasmus Research Institute of Management - ERiM

ERIM PhD Series

Research in Management

Erasmus Research Institute of Management - ERiM
Rotterdam School of Management (RSM)
Erasmus School of Economics (ESE)
Erasmus University Rotterdam (EUR)
P.O. Box 1738, 3000 DR Rotterdam,
The Netherlands

Tel. +31 10 408 11 82
Fax +31 10 408 96 40
E-mail info@erim.eur.nl
Internet www.erim.eur.nl

Design & layout: B&T Ontwerp en advies (www.b-en-t.nl) Print: Haveka (www.haveka.nl)