ECIS 2012 Proceedings

European Conference on Information Systems (ECIS)

5-15-2012

EXPLANATORY ANALYSIS IN BUSINESS INTELLIGENCE SYSTEMS

Emiel Caron
Erasmus University Rotterdam

Hennie Daniels
Tilburg University

EXPLANATORY ANALYSIS IN BUSINESS INTELLIGENCE SYSTEMS

Caron, Emiel, Rotterdam School of Management, Erasmus University Rotterdam, ERIM Institute of Advanced Management Studies, P.O. Box 1738, 3000 DR, Rotterdam, The Netherlands, caron@emielcaron.nl

Daniels, Hennie, Center for Economic Research, Tilburg University, P.O. Box 90153, 5000 LE, Tilburg, The Netherlands, and Rotterdam School of Management, daniels@uvt.nl

Abstract

In this paper we describe a method for the discovery of exceptional values in business intelligence (BI) systems, in particular OLAP information systems. We also show how exceptional values can be explained by underlying causes. OLAP applications offer a support tool for business analysts and accountants in analyzing financial data because of the availability of different views and managerial reporting facilities. The purpose of the methods and algorithms presented here, is to extend OLAP based systems with more powerful analysis and reporting functions. We describe how exceptional values at any level in the data, can be automatically detected by statistical models. Secondly a generic model for diagnosis of atypical values is realized in the OLAP context. 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. This 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. The method is demonstrated in a case study on financial data.

Keywords: Exception reporting, Variance analysis, Business Analytics, OLAP, Explanation.

1 Introduction

Modern firms can store millions of transaction data in company databases, and consequently the potential of obtaining valuable new business insights from business data has increased enormously. The proliferation of sophisticated software with new analysis tools and the online availability of data will alter the way of working of business and financial analysts. Large amounts of transaction data are nowadays stored in a company data warehouse and multi-dimensional data items like sales(2008, product, region) can be extracted from the data warehouse and organized in so called OLAP cubes for analysis. 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 inspection of multidimensional data cubes. In principle the analyst can explore the data by using the standard operators in OLAP like drill-down, roll-up and slice (Han and Kamber 2005). But as the data sets become large, browsing through the data in search for atypical values may become a complicated and tedious task. Moreover when it comes to an efficient in depth examination of the underlying causes, there is still a shortage of tools to intelligently prune a large tree of causes to its essential branches. In this paper we propose several extensions of the OLAP framework for intelligent variance analysis. Remarkable differences of actual versus reference values; like actual versus budget, actual versus historical, etc., are automatically detected by statistical models or normative models. In the next step these differences are explained by generating the most important causes at lower level data. The latter process is guided by several heuristic rules to reduce information overload.

The remainder of this paper is organised as follows. In Section 2, we summarize the most important OLAP database concepts and notations. In Section 3, we show our methodology for explanatory analysis of exceptional values. This section is structured as follows. In Subsection 3.1, we show how exceptional values in OLAP databases are defined and computed using various normative models. In Subsection 3.2, we present a general methodology for the explanation of such values based on the internal structure of the database. In Subsection 3.3, we propose techniques to prune (the tree of) explanations to its essential parts. In Subsection 3.4, we discuss how to construct consistent chains of reference objects for various types of normative models. In Section 4, we present a case study with financial sales data. Finally, we draw some conclusions in Section 5. A formal mathematical representation of OLAP databases is given in Appendix A which is used throughout the paper.

2 OLAP information systems

An important and popular front-end business intelligence 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 that form an essential part of information systems, like DSS, MIS, and ERP systems. Manipulation and presentation of such information through interactive multi-dimensional tables and graphical displays provide important support for the business analyst.

The highly normalized form of the relational data model for OLTP databases is inappropriate in an OLAP database for performance reasons (Kimball 1996). Therefore, OLAP database implementations typically employ a *star model*, 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.

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. Aggregating measures up to a certain dimension level with aggregation functions like SUM, COUNT, and AVERAGE, creates a multi-dimensional view of the data, also known as the data or *OLAP cube*.

Drill-down equations are formed by the application of a specific aggregation function *f* on a measure $y(C)$, somewhere in the lattice *L* (see Appendix A for details of the formal notation). The aggregation we consider here is the common SUM function. The measure *y* is an *additive drill-down measure* if for every cell $c \in C$, where *C* is a cube in the lattice *L*, we have

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

The latter equation is a used for expanding a dimension that is of interest. A business model *M* is a system of relations between measures. These relations can be derived from any business domain. Relations between measures are denoted by

$$y^{i_1 i_2 \ldots i_n}(C) = f(\mathbf{x}^{i_1 i_2 \ldots i_n}(C)) \, , \tag{2}$$

where and *y* and $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, are measures on the same cube *C*. *M* represents a *system of business model equations*, where each equation is defined by an instance of the above equation. A example of an equation from a financial database is given by: profit(C) = revenues(C) − costs(C).

3 Identification and explanation of exceptions

3.1 Exceptional values

Exception identification is a comparison activity carried out by business analysts. 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). The *residual of a cell* $\partial y(c)$ in some cube *C* is defined as, the difference between its *actual value*, $y^a(c)$, and some *reference value,* $y^r(c)$: $\partial y(c) = y^a(c) - y^r(c)$. The computation of the reference value is based on a normative model .The size of $\partial y(c)$ is the exception score for that cell. To identify relevant exceptions we only exceptions with a score exceeding a *threshold* δ . 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.

The normative behaviour in a multi-dimensional database, is usually defined by goals that have been formulated by management. Here we discuss two classes of normative models (*R*) that are relevant:

- *R* is a *managerial normative model* (Pounds 1969):
  - *Planning and budget models*, the plan or determined budget is the expectation;
  - *Historical models*, expectations 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.
- *R* is a *statistical normative model.* Decision-makers may also apply more abstract normative models in the form of statistical models, to compute or estimate the expected value of important measures (Daniels and Caron 2009). When applying a statistical model the expected behaviour represents the statistically normal case (Feelders and Daniels 2001). We distinguish between two broad classes of statistical models that can be applied in an OLAP database:
  - *Multi-way ANOVA models*, expectations for *continues measures* are computed by multi-way ANOVA models;
  - *Contingency table models*, expectations for *discrete measures* are computed by the independency model or the log-linear model.

3.2 Methodology for explanation

If an exceptional cell value $\partial y(c)$ is identified, 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) a top-down explanation method for both systems of equations and 2) a special greedy explanation method if only additive drill-down equations apply.

1) To determine the contributing and counteracting causes for in ∂y we define a *measure of influence* as follows:

$$\inf(x_i, y) = f(\mathbf{x}^r_{-i}, x^a_i) - y^r \ , \tag{3}$$

Where $f(\mathbf{x})$ is a relation as defined in (1) or (2) and where $f(\mathbf{x}^r_{-i}, x^a_i)$ denotes the value of $f(\mathbf{x})$ with all variables evaluated at their norm values, except x_i. The inf-measure represents a form of ceteris paribus reasoning where the x_i's play the role of causes that produced *y*. The *set of contributing* (*counteracting*) *causes* Cb (Ca) is defined as measures x_i of **x** with $\inf(x_i, y) \times \Delta y > 0 \ (< 0)$. 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.

The above definitions produce "one-level" explanations, explanations based on a single business model equation or drill-down equation. In general however, it is meaningful to continue an explanation of ∂y to lower levels in the OLAP hierarchy or by continuing in the business model. Causes can be chained together, from one level to the next in these systems, until *maximal explanation* is obtained.

2) In the special class that the measure *y* is an additive drill-down measure (like in OLAP), equation (3) can be transformed into a simple form:

Proposition 1. (Transitivity): If $C_p = [i_1 i_2 \ldots i_n] = [\mathbf{i}]$ and $C_q = [j_1 j_2 \ldots j_n] = [\mathbf{j}]$ are cubes in *L* where $C_q \le C_p$, $c \in C_p$ and $c' \in C_q$, and *y* is an additive drill-down measure then:

$$\inf(y^{a;\mathbf{j}}(c'), y^{a;\mathbf{i}}(c)) = y^{a;\mathbf{j}}(c') - y^{r;\mathbf{j}}(c') \ . \tag{4}$$

The proof of the proposition is given in (Caron 2012) .

This proposition states that, in a system of additive drill-down equations *the influence of a variable* $y^{\mathbf{j}}(c')$ *on any ancestor variable y 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 on the upset of a cell, because we only have to compute the difference between the actual and reference value of a cell, instead of repeatedly applying equation (1). This property is used in a *greedy algorithm for the explanation*. The inputs for the algorithm is an exceptional cell and a table with actual, norm and influence values for elements in the exceptional cell's downset. In the second step the causes are determined in the aggregated table by selecting the *n* largest causes and filtered by some filter measure (see Subsection 3.3). The output of the algorithm is the tree of largest causes.

3.3 Reducing information overload

Because every drill- down equation in the multi-dimensional database yields a possible explanation, the number of explanations generated for a single symptom can be very large. By leaving out insignificant influences we can reduce information overload to a large extent. We propose three *generic reduction methods* (RM_1-RM_3) to cut down the number of explanations.

RM_1) Small influences are left out in the explanation by a *measure of parsimony*. The parsimonious set of contributing causes, denoted by Cb_p, is the smallest subset of the set of contributing causes, such that its influence on *y* exceeds a particular fraction T^+ of the influence of the complete set. The fraction

T^+ is a number between 0 and 1, and will typically 0.9 or so. Alternatively, in the case of the greedy algorithm the analyst might select the number of significant causes he wants so see for a particular symptom. In this way the analyst can simply select the *n* largest causes. For example, the analyst can generate a top-10 list of largest causes for only the Product dimension.

RM$_2$) The number of explanations is reduced by applying a *measure of specificity* for each applicable equation. This measure quantifies the "interestingness" of the explanation step. The measure is defined as:

$$\text{specificity } (S) = \frac{\#\text{ possible causes}}{\#\text{ actual causes}}. \tag{5}$$

The number of possible causes is the number of right-hand side elements of each equation, and the number of actual causes is the number of elements in the parsimonious set of causes. By using this measure of specificity, we can diminish the number of explanation paths if only the most specific dimensions are explored.

RM$_3$) The analyst can also manually select a few dimensions for further exploration and ignore those which seem less interesting.

3.4 Consistency of reference objects

A correct interpretation of the influence measure (see expression (3)) is only possible if and only if the *consistency constraint* is fulfilled. This constraint says 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. Here we describe under what conditions the constraint is satisfied. We discuss how *consistent chains of reference objects* can be formed for the different types of normative models. Actual values in the OLAP context are consistent because they satisfy the drill-down equations (equation (1)) or business model equations (equation (2)) by definition. Often reference values are computed directly from the actual values in the OLAP database. In the case that $y = f(\mathbf{x})$, and it is given that $y^r = R(y)$, $\mathbf{x}^r = R(\mathbf{x})$, and $f \circ R = R \circ f$, i.e. the computation of reference values is commutative, then the reference values are consistent. Now there is a natural canonical way to construct a consistent chain of reference values 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 both actual as 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 both actual as reference values for the business measures.

For each type of normative model a consistent chains of reference values can be constructed. Here we consider two important cases.

1) *R* is selected as a historical model. In this case the reference objects are basically internal, and directly available in the database, because the Time dimension is in principle always part of the star model. The historical reference objects, in the case of pairwise comparison, are determined by a specific slice operation on the Time dimension, where, for example, the previous year is selected as the normative model. Because the reference objects are just cells in a cube *C*, the consistency of reference values in drill-down equations is guaranteed by definition.

2) *R* is selected as a statistical model. Statistical normative models, in general, do not produce a consistent chain of reference values, because many statistical models have multiplicative terms that result in reference values that are not commutative, i.e. $f \circ R \neq R \circ f$. An exception to this general rule are some additive ANOVA models. Suppose that A_1 and A_2 are additive ANOVA models.

Reference values are now computed by $y^r = A_1(y^a)$ and $\mathbf{x}^r = A_2(\mathbf{x}^a)$. Consistency holds if and only if $y^r = A_1(y^a) = A_1(f(\mathbf{x}^a)) = f(A_2(\mathbf{x}^a)) = f(\mathbf{x}^r)$, thus $A_1 \circ f = f \circ A_2$. The construction of a consistent chain of reference values is guaranteed, if and only if, the additive 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 used for the parent cell.

Proposition 2. (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

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

The proof of this proposition is given in (Caron 2012).

4 Case study: explanatory analysis of financial OLAP data

The database used for the case study consists of 42.063 records and is obtained from Cognos (Cognos 2008). The central fact table represents the financial data set. It contains the measures like profit, revenues, costs, etc. The financial data set has dimensions tables, like Time (T), Product (P), Location (L), etc.. The hierarchies for these dimensions are given by T[Month] $\prec$ T[Quarter] $\prec$ T[Year] $\prec$ T[All-Times], P[Product] $\prec$ P[ProductType] $\prec$ P[ProductLine] $\prec$ P[All-Products], and L[Name] $\prec$ L[Position] $\prec$ L[City] $\prec$ L[Country] $\prec$ L[All-Locations].

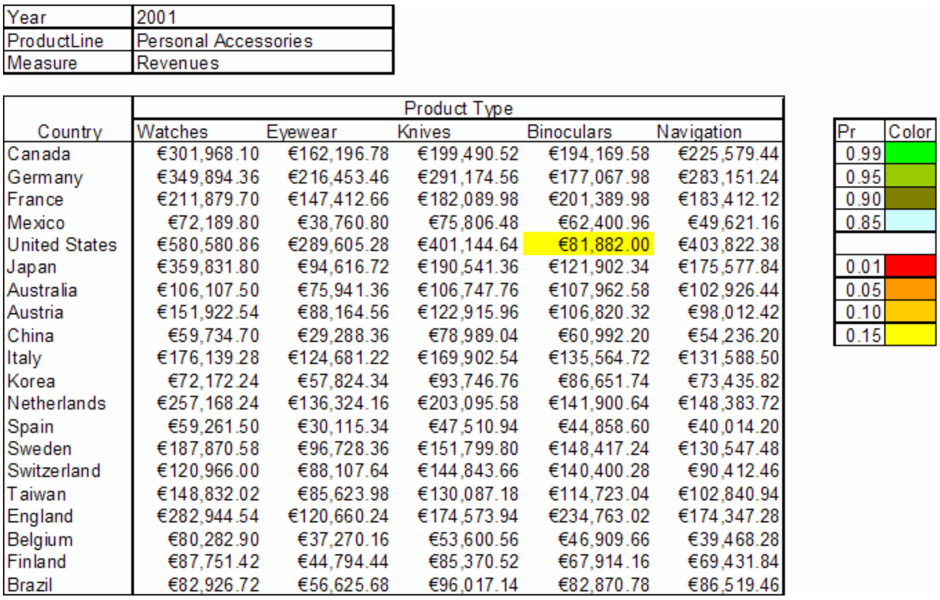
4.1 Exception identification

First the applicability of the method for statistical exception identification is shown in an example. Here we apply the method on the cube Year × Country × ProductLine, with slices $S^{\text{Year=2001}}$ and $S^{\text{ProductLine=Personal Accessories}}$, for the measure y^{231}= revenues^{231}. The resulting cube 2001 × Country × Personal Accessories is denoted by *C*. The cube's initial actual data is presented in Figure 1, it describes the revenues figures of the GoSales company in 20 countries where the company is active for 5 types of product accessories in the year 2001.

The algorithm for exception identification is configured with *R* selected as a simple additive ANOVA model. Here the additive two-way ANOVA model

$$\hat{y}^{231}(\text{2001,Country, Personal Accessories}) = \hat{\mu} + \hat{\lambda}_1(\text{Country}) + \hat{\lambda}_2(\text{Personal Accessories})$$

is applied to compute the reference values. All the residuals 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 δ = 1.036 and -δ = 1.036, we find that the cell *c* = (United States, Binoculars) in the year 2001 is the only (low) exception with the residual $\partial y(c)/s$= -1.212, because -1.212 < -1.036. This exceptional cell is indicated with a yellow color in Figure 1. The analyst now might want to explore this deviating cell in more detail, to find he reasons for the deviation in the cell's downset.



Year	2001
ProductLine	Personal Accessories
Measure	Revenues

Country	Product Type				
	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 1: Revenue figures, derived from the financial database, organised per type of Personal Accessories (P^1) and Country (L^3) with a slice on the year 2001 (T^2). Here the cell (United States, Binoculars) is identified as a moderate ``low exception" .

4.2 Top-down explanation

Here we address the 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?*" The answer to this question is given with top-down explanation in the downset {↓c}, in particular in the Time dimension over the path p = [231]→ [131] → [031]. In this case the analyst wants to explain the exception solely in the Time dimension over the path p, i.e. on the Quarter and Month level (see RM_3). As an additional reduction method, RM_1 is applied here with fraction $T^+ = T^- = 0.9$, to remove the effect of marginal causes. For explanation of the event, for each cell on the path p in {↓c}, both the actual as the reference value are required. 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 $c' = R_T^{-1}(c)$ produces the actual values for cells on the Quarter level:

$$y^{231}(c) = \sum_{i=1}^{4} y^{131}(2001.\text{Q}_i, \text{U.S.A.}, \text{Binoculars}) \,.$$

Moreover, the reference values for cells in p are computed here 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 computed with the additive ANOVA model A_1

$$\hat{y}^{131}(c') = \hat{\mu} + \hat{\lambda}_1(2001.\text{Quarter}) + \hat{\lambda}_2(\text{Country}) + \hat{\lambda}_3(\text{Personal Accessories}) \,,$$

in the context cube 2001.Quarter × Country × Personal Accessories. A_1 is a specialized additive ANOVA model for the quarters, which is a specialization of ANOVA model A_0 within an unfolded Time dimension. The model contains the effects of the ANOVA model that was used for the parent cell, plus the 2001.Quarter-effects. The two conditions for Proposition 2 are fulfilled, and therefore the drill-down equation for the reference values holds:

$$\hat{y}^{231}(c) = \sum_{i=1}^{4} \hat{y}^{131}(2001.\text{Q}_i, \text{U.S.A.}, \text{Binoculars}) \,.$$

Next in Table 1 comparison is made between the actual and the reference values for the cell c in the Time dimension, on the level Quarter. In this table the influence values are computed by expression (4), because the measure revenues is additive. The inf-measure is correctly interpreted as a quantitative specification of the change in $y^{231}(c)$ that is explained by a change in $y^{131}(c')$ by consistency.

	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

Table 1. *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.*

In the table relative influences are computed by $(y^{a}(c)\text{-}y^{r}(c))/\inf(y(c'),(c))$. From the data in the table it can be concluded that 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, $\text{Ca}_{\text{p}} = \varnothing$.

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 by the algorithm for each month by estimating an additive ANOVA model. Therefore, for each cell $c" = R_T^{-1}(c')$ its reference value is computed by the specialized ANOVA model A_2

$$\hat{y}^{031}(c") = \hat{\mu} + \hat{\lambda}_1(\text{2001.Month}) + \hat{\lambda}_2(\text{Country}) + \hat{\lambda}_3(\text{Personal Accessories}),$$

in the context cube 2001.Month × Country × Personal Accessories. The model A_2 is a specialization of model A_1 within the Time dimension, from the Quarter to the Month level. In this way, consistent reference values are formed for each quarter Q_i given by

$$\hat{y}^{131}(\text{2001.Q}_i, \text{U.S.A., Binoculars}) = \sum_{i=1}^{3} \hat{y}^{031}(\text{2001.Q}_i.\text{Month}_{\text{j}}, \text{U.S.A., Binoculars}),$$

where i = 1,2,3,4 and j = 1,2,3. The values are consistent because, the ANOVA model applied on the Month level, is a specialization of the ANOVA model applied on the Quarter level, and therefore the conditions for Proposition 2 are met.

As an example, comparison is made in Table 2 between the actual and the reference values for the cell (2001.Q4, U.S.A, Binoculars) and its children on the Month level.

	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

Table 2. *Data for explanation of y^{131}(2001.Q4, U.S.A, Binoculars) = "low" in the Time dimension, on the level Month in the cube 2001.Quarter.Month × Country × Personal Accessories.*

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, $\text{Ca}_{\text{p}} = \varnothing$. 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).

The explanation tree in the lower part of Figure 2, summarizes the results of this explanation. In this figure, the straight lines indicate parsimonious contributing causes and (possible) dotted lines indicate counteracting causes, the numbers on the lines indicate the relative values for the influence measures, and the ratios indicate the specificity value (S) of the explanation step (see RM_2). In addition, we give a business interpretation of the complete explanation tree for 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 managerial inspection.

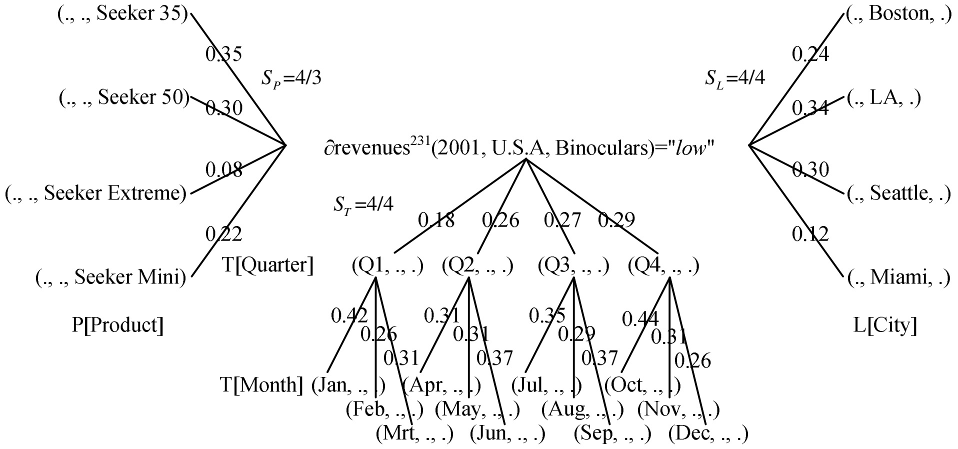


Figure 2: Explanations trees that partially explain the exceptional cell revenues(2001, U.S.A, Binoculars) = ``low" in the Product (P), Time (T), and Location (L) dimension.

Furthermore, in Figure 2, three partial explanation trees are depicted, 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 for each dimension. For all the one-step explanation that are possible in the downset of the exceptional cell c, 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. The 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. In the algorithm this mechanism can be continued until it reaches the base cube. By the application of the measure of specificity the business analyst is automatically guided through the exceptional cell's downset $\{\downarrow c\}$. This reduction method selects in each explanation step, the dimension for explanation that is the most specific.

4.3 Greedy explanation

Here we identify exceptions in the cube C = 2001 × Country for the measure profit[1], labelled by the variable y, with a historical normative model, in this case the profit figures of the previous year, represented by the cube C' = 2000 × Country. The following cells in C are marked as exceptions: a moderate high exception is the cell (2001, China) and the cells (2001, Canada), (2001, The

[1] Notice that the actual data for this cube is not presented in this paper because of space limitations.

Netherlands), (2001, Spain), (2001, Sweden), and (2001, Belgium). The largest exception is found in the cell c = (2001, The Netherlands), where $\partial y(c) = y^a(c) - y^r(c') = 199{,}690.65 - 378{,}324.70 = -178{,}634.05$. Subsequently, we want to explore the exceptional cell $\partial y(c)$ in more detail, to identify possible causes for this exception in $\downarrow\{c\}$. In words, we address the following business question: "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?" Here the exceptional cell $\partial y(c)$ is explained with greedy explanation in only the Product dimension.

In Table 3, the data for greedy explanation for the city of Amsterdam is presented. 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. Interestingly, is the cause y^{220}(., ., Camping Equipment.Tents.Star Dome) which is relatively large contributing cause on the lowest level of the Product dimension.

Nr.	ProductLine	ProductType	Product	Actual	Norm	Rel. Inf.
	P^2	P^1	P^0	(2001)	(2000)	
	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 3. Aggregated table for the Product dimension where the actual object is the year 2001, the norm object is the year 2000, and the influence values for instances within the Product dimension are related to the exceptional cell profit^{223}(c).

In Figure 3, the results are depicted in an explanation tree, which reports specifically the 10 largest contributing causes for the exception in the Product dimension (see RM$_1$).

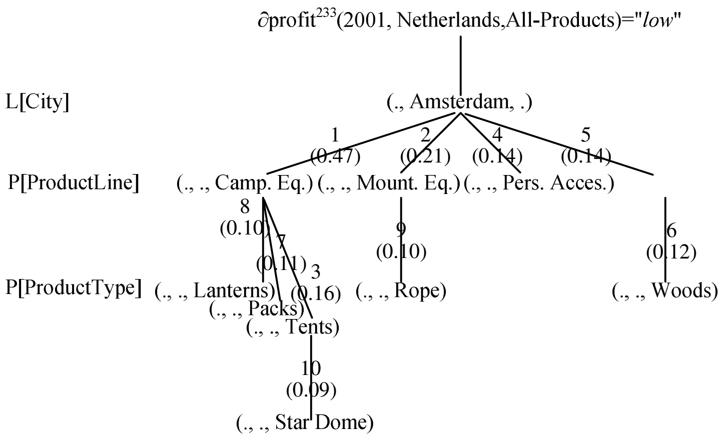


Figure 3: Greedy explanation in the Product dimension reporting the 10 largest causes.

4.4 Scalability

Although transaction databases can be very large the kind of analysis discussed in the paper is mostly performed on aggregated data. The method for explanation as described in this paper is scalable since all operations are linear in the number of records in an OLAP data cube. Note that here the ANOVA models for computation of the reference values also have linear complexity. If other more complex statistical models are applied like 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

$$\frac{(n_1 + n_2 + \ldots + n_k)!}{n_1!n_2!\ldots n_k!}$$

paths, where n_k is the number of possible levels in dimension *k*. In this case the complexity is still linear in the size of the dataset, 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.

5 Conclusion

In this paper we proposed some new methods for investigation and evaluation of financial data that are stored in multi-dimensional OLAP databases. Exceptional values are automatically discovered using statistical or normative models. Interesting dimensions to be expanded are computed from a business model and can be analyzed in further detail and displayed in a tree of causes. Several strategies to reduce information overload are presented and applied in the case study. We believe that the methodology put forward here, can be effectively employed in a wide range BI systems. Example applications are: interfirm comparison (Daniels and Caron 2009), sales analysis (Caron 2012), crime analysis (Caron and Veenstra 2007), analysis of variance in accounting, and the generation of fishbone diagrams. The method can also be applied 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 paper. Larger deviations serve as a trigger for audit activities in which case the explanation method automatically generates important dimensions that can be explored in further detail. Computerized diagnosis in the business and management domain is an important research area, studied in the fields of Operations Research and Artificial Intelligence. In our opinion, this paper contributes substantially to the integration of diagnostic support in business information systems.

Appendix A

In OLAP databases dimensions can be represented by $D_1^{i_1}, D_2^{i_2}, \ldots, D_n^{i_n}$ where each domain $D_k^{i_k}$ represents a dimension *k*, e.g. Time, Location, Product and so on, from the associated business process, with a set of *dimension levels* $i_k = \{0, 1, \ldots, \max_k\}$. For example, the Time dimension might have the following levels: Day, Week, Month, Quarter, Season, and Year. In dot-notation an example hierarchy for the Time dimension is represented as Year.Quarter.Month. The key structure in the multi-dimensional database is the data cube. A *cube C* is defined as the Cartesian product over the levels of the subsets of the 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 = {2008, 2009}^{2} × Germany^3 × Product^2 is an example of a cube. Also every pivot table in MS Excel is an example of a 2 dimensional cube.

A cube C is composed out of one or more cells. A *cell* c is defined as an instance element of a cube C

$$c = (d_1^{i_1}, d_2^{i_2}, \ldots, d_n^{i_n}), \text{ where } d_1^{i_1} \in X_1^{i_1}, d_2^{i_2} \in X_2^{i_2}, \ldots, d_n^{i_n} \in X_n^{i_n}.$$

A number of *navigational operations* are available for the manual exploration OLAP cubes, e.g. down, roll-up, slice and dice, allowing interactive querying and analysis of the data. The *drill-down operator* is defined by

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

A *roll-up* operator in dimension q, given by R_q^{+1}, is defined similarly. This is the inverse of the drill down operator and aggregates a cube to a higher level for dimension q.

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

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

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

References

Cognos Software Corporation (2008). Cognos 8 business intelligence, Powerplay.

Caron, E.A.M. (2012). Explanation of exceptional values in multi-dimensional databases. Ph.D. thesis, Erasmus University Rotterdam (Forthcoming, available from http://www.emielcaron.nl).

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, 11 pages. Tsinghua University Press.

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.

Feelders, A. J. and H. A. M. Daniels (2001). A general model for automated business diagnosis. European Journal of Operational Research 130, 623–637.

Han, J. and M. Kamber (2005). Data Mining: Concepts and Techniques. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

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.

Pounds, W. F. (1969). The process of problem finding. Industrial Management Review 11 (1), 1–19.