Analysis for Detecting and Explaining Exceptions in Business Data

Lingzhe Liu
Rotterdam School of Management, Erasmus University, the Netherlands
lliu@rsm.nl

Hennie Daniels
Tilburg University, the Netherlands
daniels@uvt.nl

Abstract
In this paper we describe the concepts of automatic analysis for the exceptional patterns which are hidden in a large set of business data. These exceptions are interesting to be investigated further for their causes and explanations, as they provide important decision support. The analysis process is driven by diagnostic drill-down operations following the equations of the information structure in which the data are organised. Using business intelligence, the analysis method can generate explanations supported by the data. The methodology was tested on a case study and was reflected in considering the practical aspects of its application procedure.

Keywords: Explanatory Analysis, Business Intelligence, Decision Support

1 Introduction
“Management by exceptions” has long been a philosophy for business administration. Managers perceive the environment of a company, form an expectation, and decide on the operations planning. Additional decisions will be made when deviations from the expectation occur. To detect those exceptions, monitor and control loops are devised in the company to continuously collect data about the internal operations and the external environment. Once an exception is detected, the managers need an explanation “why the exception occurred” and make informed decisions on subsequent (re-) actions – whether and how to treat the exception.

Nowadays, the business environment becomes more and more complex, so do the management system and the company itself. The volume of the collected data thus increases tremendously. To avoid information overload, business intelligence (BI) techniques are applied to generate significant information / knowledge from company data. Reports are
generated by aggregating the data before presented to the manager. When the manager is examining the report he/she is looking for extreme or unexpected items and the explanations. The explanation is mostly done by analytics, i.e., reversing the process of report generation, drilling down in a business model, or using additional knowledge possibly from external sources (see Figure 1 in Section 2.1). The process of explanation and causal analysis describes business analysis in general and also indicates the way that BI is applied in decision support concerning exceptions.

Lots of attention has been drawn to the strategic value of business intelligence and analytics in recent years (Bay, Kumaraswamy, Anderle, Kumar, & Steier, 2006), translating analytics into a competitive edge (Davenport, 2006). “Management by exceptions” is then endowed with new implication of “responding to changes proactively”, rather than the old ways of “reactive fire-fighting”. Even though analytics has been vigorously applied in industry, there is hardly any research on the general methodology for doing business analytics systematically.

In this paper we describe the concepts of a general methodology on how to apply statistical methods automatically to analyse the exceptional patterns which are hidden in a large set of business data, based on (Caron, 2012). This paper is organized as follows. In Section 2, we examine the concepts of BI supported business analytics and discuss a general model for the methodology. Section 3 provides a case study to illustrate the application of the model, and the practical aspects of the application are discussed in Section 4. The last section concludes the paper.

2 Business Analytics

The use of analytics in business can be roughly grouped into two parts. First, descriptive analytics captures the pattern of systematic emergence in the company or the environment. The description usually supports prediction. Examples are the data mining algorithms like clustering, classification and association, applied to identify the most profitable customers, or the changes of customer behaviour. Although descriptive analytics does not presume any expectations, the analyst usually looks for “interesting” patterns when interpreting the results. In this process, implicit background knowledge is applied in searching for (mental) exceptions (Keil, 2006). Secondly, diagnostic analytics reason about the causal relations of those patterns. The goal for this type of the analysis is to restore or verify the mechanism of a sequence of events (Keil, 2006), e.g. the operations in the company. The conclusion usually leads to decisions for adjustment and improvement of the system. Exemplary analysis questions are “why the company performance is not as expected” – for improving performance of the managed system, and “why certain exceptions have not been detected by current monitors” – for adjusting the management system. Audit analytics also falls in this category (Bay et al., 2006; Vasarhelyi, Alles, & Kogan, 2004). In the framework we propose (see Section 2.2), we generalize and combine these two types to the detection and the diagnosis phases in an integrated process of business analytics.

We argue that business analytics is a strategic important process of organizational learning that extends the philosophy of "management by exception". The importance of analytics lies
in the necessity of "meta-control" to cope with the internal and external changes. The management system of the company monitors and controls the business processes, which deliver value to customers and form competitive competence. Modern management systems, such as ERPs, automate the routine tasks of detecting and treating operational exceptions, because the business knowledge are codified into the build-in controls of the system (in form of business rules or constraints) in a “plan-do-check-adjust” cycle. With automation, management systems can help with handling these routine tasks in large volume data (big data), e.g. managing thousands of accounts in finance and cost accounting systems. However, their monitor-control capability is limited to the codified rules, so they cannot deal with the “new” changes or the exceptions out-of-scope of the rules. These exceptions are left to the responsibility of human managers. Though the “new” exceptions are on a higher system level than the management system ergo not directly visible, they affect the performance of the managed system (the company): therefore, they must be detectable by analysing the data collected / generated by current management system (ERP). The analysis results in new business knowledge that equips the management system for controlling similar exceptions in the future. Ideally, the managers hope to continuously meta-control the management system, automating the process using BI (Vasarhelyi et al., 2004).

2.1 BI supported Business Analytics

Business Intelligence is the collection of procedures to reduce the volume of information that the manager need to take into account when making decision. The information-reduction is done by organising (extract-transform-load, ETL) transactional data into a multi-dimensional database (data warehouse or OLAP), in which large volume of operational details can be abstracted, aggregated or computed into business reports, using BI techniques (see Figure 1).

![Figure 1: Business analytics supported by BI](image)

This process involves both the business model and the technical information model. On one hand, the organising of information is in essence driven by managerial purpose i.e. the business model. For example, the accounting process, which in general is a BI process, aggregates transaction records in various documents such as journals, general ledgers and
financial statements for operating, financing and investing purposes respectively (Bay et al., 2006). The organization of these documents codifies the business model. For instance, the general ledger, recorded using double-entry book-keeping, is a codified management system which internally controls balance between two accounts involved in each transaction (Bay et al., 2006). On the other hand, the technical model organises information for an analytical purpose. Organising business data in the form of tables helps to highlight contextual similarities among the data, providing important support for the business analyst. For instance, aligning records chronically, e.g. sales in multiple periods, can show the temporal changes and trends in the record set. As a special case, OLAP is a useful tool to analyse multi-dimensional, hierarchical data interactively, with the standard drill-down, roll-up and slice operations (Caron, 2012). From an analytics viewpoint, the business model provides an ontological structure of the information, while the technical model gives a storage structure, also known as data structure in computer science.

2.2 A General Model for Business Analytics
Before the analytic process can be automated, its procedure should first be formalized. The lexical definition of exception is “an instance that does not conform to a rule or generalization” (thefreedictionary.com), which implies the comparison of the actual instance to a norm. Our discussion on business analytics is largely based on previous works of causal analysis and explanations in (Caron & Daniels, 2008, 2009; Caron, 2012; Feelders & Daniels, 2001). The analysis of exceptions takes the canonical format of (Feelders & Daniels, 2001):

\[ (a, F, r) \text{ because } C^+, \text{ despite } C^- \]

where \((a, F, r)\) is the triple for exception detection, and the exception is to be explained by the non-empty set of contributing causes \(C^+\) and the (possibly empty) set of counteracting causes \(C^-\). The diagnosis analysis is to explain why the instance \(a\) (e.g. the ABC-company) has property \(F\) (e.g. having a low profit) when the other members of reference class \(r\) (e.g. other companies in the same branch or industry) do not.

The information structure of \(r\) has the general form of \(y = f(x)\), where \(x = (x_1, x_2, \ldots, x_n)\) is an \(n\)-component vector. In words, certain property value of \(a\) which is important for decision making, denoted by \(y\), is dependent on other property values \(x\) in the information structure of \(r\).

We can use the information structure to estimate the norm value of \(y\), given the actual values of \(x\). Exception-detection is done by studying the difference between the actual and the norm value of \(y\):

\[ y^a = E(y|x^a) + e, \text{ where } e \sim N(0, \sigma). \]

If the difference \(e\) is significant, i.e. \(|e| > \delta \sigma\), \(y^a\) is viewed as a symptom to be explained. The estimation method for \(E(y|x^a)\) depends on \(f(x)\) and the domain application. The user defined threshold parameter \(\delta\) also depends on the application. A more general form of (2) is
\( y^a = \mathbb{E}(y|\text{info}) + e \), where info stands for all kind of information available. For example, Alles et al. (2010) uses the information of sales of prior period \( x_{z-1}^a \) to estimate the profit of current period \( y_z^a \).

The symptom is explained by the influence of each \( x_i \), and the influence is measured as

\[
\inf(x_i, y) = f(x_{z-1}^i, x_i^a) - y^r, \quad i = 1, 2, \ldots, n.
\]  

(3)

where \( f(x_{z-1}^i, x_i^a) \) denotes the value of \( f(x) \) with all variables evaluated at their norm values, except \( x_i \).

For clarity, we distinguish the technical model from the business model in the information structure. For example in OLAP (see equation system (4)), the variables in a business model (shown as the functional relation \( g \)) can be organised into a hierarchy by aggregation, such as summation or average (shown as the functional relation \( h \)). Vertically, all variables in the business model are organised based on the same aggregation relation \( h \). Given that, the variables on a specific level of aggregation follow the same business relation \( g \), just as those variables on other aggregation levels horizontally do.

In (4), the variables \( y \) and \( x \) are organized in an OLAP cube with \( l \) dimensions. Each dimension has a hierarchy of \( q_k \) levels, \( k = 1, 2, \ldots, l \). In a specific dimension \( k \), variables on the hierarchy level \( q_k \) are aggregated from the \( m \) elements in the lower hierarchy level \( (q_k - 1) \), and these elements are denoted respectively as \( y_j \) and \( x_j \), where \( j = 1, 2, \ldots, m \). \( x_j \) is an \( n \)-component vector, whose components are denoted as \( x_{i,j} \).

\[
y = g(x) = g(x_1, x_2, \ldots, x_n)
\]

(4)

\[
y_j^{q_1 \cdots (q_k-1) \cdots q_l} = g(x_j^{q_1 \cdots (q_k-1) \cdots q_l}) = g(x_{1,j}^{q_1 \cdots (q_k-1) \cdots q_l}, \ldots, x_{n,j}^{q_1 \cdots (q_k-1) \cdots q_l})
\]

\[
x_j^{q_1 \cdots (q_k-1) \cdots q_l} = h(y_j^{q_1 \cdots (q_k-1) \cdots q_l}) = \sum_j y_j^{q_1 \cdots (q_k-1) \cdots q_l}
\]

With the information structure available, we can look at lower level of detail for explanation by drilling down. For example, if there is a significant symptom \( e_j^y \) in the OLAP model \( h \), detected by \( y^a = \mathbb{E}_h(y|y_j^a) + e_j^y \), we can drill down the business model \( g \) for explanations, using \( y_j^a = \mathbb{E}_g(y_j|x_j^a) + e_j^x \). A necessary condition to obtain sensible explanations by drilling down is consistency of the normative estimation, i.e.

\[
\mathbb{E}_g(y|x^a) = \mathbb{E}_g(h(y_j)|h(x_j^a)) = h(\mathbb{E}_g(y_j|x_j^a))
\]

(5)
This condition in relation with $g$ usually holds for the OLAP model, but should be checked for (statistical) business models in general. This issue is studied in depth for ANOVA models in OLAP databases (Caron, 2012).

### 3 Case Study: Analysis of Enterprise Performance

The database used for the case study is obtained from Microsoft (Microsoft, 2012). The operations data of the fictional bicycle enterprise adventure works are organised in an OLAP cube (Table 1), \((\text{measurement}, \text{product line}, \text{country}, \text{period})\). We analyse the profits of the specific product line of road bikes. The performance measurements satisfy the business model, \(\text{profit} = \text{sales} - \text{cost}\). Each measurement has an information hierarchy with time (period) and sales territory (country) dimensions.

#### 3.1 Exception Identification

In this case study, we can compute the expected value in the OLAP model with a multi-way ANOVA model (Caron & Daniels, 2009; Daniels & Caron, 2009). The cells in the top level cube can be estimated with a one-way ANOVA, as the cube has only one dimension. The overall effect, i.e. the average of the cells, is regarded as the expected value. The lower level cube has two dimensions, and its cells can be estimated with one of the two-way ANOVA models shown in equations (6) to (9). Basically, the more effects are considered in the ANOVA, the higher accuracy the estimation will have.

\[
\hat{y}(\text{country}, \text{period}) = \mu \quad (6)
\]

\[
\hat{y}(\text{country}, \text{period}) = \mu + \lambda_1(\text{country}) \quad (7)
\]

\[
\hat{y}(\text{country}, \text{period}) = \mu + \lambda_2(\text{period}) \quad (8)
\]

Table 1: Performance figure of the European territory in financial year 2007: the colours indicate the level of exception of a cell, red for exceptionally low values and green for high.

The figures are in thousand dollars.
\[ \hat{\gamma}(\text{country}, \text{period}) = \hat{\mu} + \hat{\lambda}_1(\text{country}) + \hat{\lambda}_2(\text{period}) \]  

(9)

The business model in our case is a one-to-one relation, so we cannot make any estimation.

Under the normality assumption, we choose \( \delta = 1.29, 1.65 \) or 2.33 for the exception threshold, which corresponds to the fact that \( \gamma^a \) has 90%, 95% or 99% probability to be exceptional when \( |e| > \delta \sigma \).

3.2 Explanation

In our top level OLAP cube \( \text{profit}^{0.1,0} \) (road bikes, country, all periods), three cells are marked as exceptions: \( \text{profit}(\text{Dec } 2006) \), \( \text{profit}(\text{Jul } 2006) \) and \( \text{profit}(\text{Aug } 2006) \). Explanation may follow two drill-down paths (Figure 2). On the left path, we first follow the business model to associate exceptions in the cubes \( \text{profit}^{0.1,0} \), \( \text{sales}^{0.1,0} \) and \( \text{cost}^{0.1,0} \). The low profit of \( \text{profit}(\text{Aug } 2006) \) is due to very high cost (prob. 95%) despite mild high sales (prob. 90%). Interestingly, this step of analysis finds two effects of low sales and low cost in Jan 2007, which are cancelled out when combined to profit. No plausible explanation can be found for the other two exceptions, and the latter should be explained in the next step.

We then drill down in the OLAP model. The only possible drill down direction is (country). Reminding the consistency condition (5), we cannot use the two-ANOVA models (8) and (9) for exception identification, because they include the effect of \( \hat{\lambda}_2(\text{period}) \) which cannot be estimated on the top level. Table 1 uses (7) for exception identification in the lower level cube and finds the explanations for the two exceptions in the top level cubes:

- \( \text{profit}(\text{Jul } 2006) \downarrow \Rightarrow \text{profit}(\text{UK, Jul } 2006) \downarrow \) and
- \( \text{profit}(\text{Dec } 2006) \uparrow \leftarrow \text{profit}(\text{France, Dec } 2006) \uparrow \).

However, \( \text{profit}(\text{UK, Jul } 2006) \downarrow \) cannot be further explained by \( \text{sales}(\text{UK, Jul } 2006) \) and \( \text{cost}(\text{UK, Jul } 2006) \), as no exceptions are detected there. This may be due to some kind of incompatibility in the design of the business model: the estimations of \( \sigma(\text{profit}) \), \( \sigma(\text{sales}) \) and \( \sigma(\text{cost}) \) are computed separately instead of simultaneously. In general, it is possible to harmonize the estimation in a properly designed business model, e.g. using the continuity equation developed in (Alles et al., 2010).

Following the right path in Figure 2 gives the same set of causality relations as the left path.
4 Procedure for Analytics

Based on the discussion above, we can summarize a general procedure for business analytics, with considering the practical methodology of data analysis (Feelders & Daniels, 2000):

1. Define problem: define analysis goal and choose the variable which is important for decision.
2. Establish context: abstract and explicitly specify the information structure (or load from a knowledge base, if available). The context is usually connoted by the source of information from which the business report was generated. Sometimes external sources need to be included to enlarge the context, depending on the analysis goal.
3. Identify exceptions: choose appropriate reference class, estimate the norm, and apply it to actual data. Despite the wishes for fully automated analysis, the derivation of the norm remains an interactive process in which several practical aspects demand lots of background knowledge from the analyst (see Section 4.1).
4. Generate explanations: relate the exceptions in different parts of the business system and reason about the causal relations, using equation (3). Method for developing the relations has been well studied in previous works (Caron & Daniels, 2008, 2009), including greedy and top-down explanation.
5. Interpret results: review the explanations. In case the results does not sufficiently supports decision, repeat step 2 to 5.

4.1 Practical Aspects

The following two key tasks are the most intricate in the process of business analysis:

1. How to find an appropriate normative model to detect exceptions, and
2. How to find the real causes to explain the relationship between the exceptions.

4.1.1 Exploration: Finding an Appropriate Norm

Business analysis is in any case an exploratory process. The normative model plays a central role in qualifying a feature as normal or exceptional. The firstly used normative models to detect symptoms are usually the codified business constraints in the management system, such as plans or budgets. Peculiarly, in the subsequent diagnostic analysis to explore a sensible explanation, the choice of the normative model for the lower level of analysis relies to a large extent on the choice of the analysis context, because the analysis goal is usually an open question. For instance, a decrease in profit may due to the drop in internal efficiency or the deteriorated global economy.

In the exploration for the subsequent normative models, statistics are usually applied to the analysis context, i.e. the members of the reference class \( r \). The method for choosing a proper reference class can be “softening” the set of business constraints used in the management system for a particular monitor. For instance, we can use the constraint \( \langle \text{profit}, \text{road bike}, \text{France}, \text{FY2006} \rangle \) to monitor the value of profit in a particular OLAP cube \( \langle \text{product line}, \text{country}, \text{time} \rangle \). Using an un-slice operation, we soften the constraint and retrieve a class of reference values \( \langle \text{profit}, \text{road bike}, \text{France}, \text{all years} \rangle \).
Softening business constraint is a useful technique for analysis. The un-slice operation takes the union of the data sets which correspond to different parts of the system. It thus expands the analysis scope, so that the patterns on a larger system scale can be revealed. For example, in the time dimension, the trend or fluctuation of a variable over time can only be seen on a time period, but not at a time point. Besides, expanding the scope by un-slicing is in itself an attempt of exploration, for instance in searching for those exceptions whose impact only takes effect after a time lag (Alles et al., 2010). This in general helps the analyst to involve extra data by extending the current information structure: in any case, one can always organize the information of the analysis context into an OLAP-like structure, and then start to expand.

The reference class is always defined by a set of constraints. Reminding of the codified business constraints in the first place, the exploration for an appropriate reference class can be regarded as a “meta-control” process that diagnoses and reflects upon the detective power of the current set of constraints, performed by the analyst (see Section 2). The exploration thus iteratively applies the detective and diagnostic processes on the design of the business analysis method. For example, in our case study, the appropriateness of the business model is reflected.

4.1.2 Validation: Finding the Real Cause

Correctness and relevance are two important criteria for evaluating the explanation. The correctness of the models in the information structure is a premise for finding the real cause. If the model doesn’t capture the business correctly, the reference model would be based on a false assumption, and it would then be incapable even in explaining a normal effect. For example, the ANOVA model used in our case study fails to include the season effects among the months in a year. As a result, the model will possibly raise many false alarms. The relevance concerns the usefulness of the explanation for decision support. A counter-example is the explanation presented at the wrong level of detail (also pointed out in Keil, 2006). In our case study, the analysis would be irrelevant if the analysis goal is to determine the contribution of sales region (which is one level lower than country) to the overall performance.

The method for the evaluation of the correctness and relevance generally rely on the background knowledge of the application domain, which is out-of-scope of this paper.

5 Conclusion

Current business databases contain massive amounts of data that carry important explicit and implicit information about the underlying business process. In this paper we have shown how general statistical methods can be applied to automatically detect implicit patterns that are interesting to be investigated further. In many cases the data itself include enough information to discover unusual patterns or trends to be explored further, like in an OLAP database. The process of examination is driven by accounting equations or drill-down equations and can generate explanations supported by the data. In the future we want to investigate the incorporation of heterogeneous external data sources to obtain a richer structure for causal analysis as described in this paper. Another case study in risk management in global supply chains is currently explored.
References