AUTOMATED EXPLANATION OF FINANCIAL DATA

H. A. M. DANIELS and E. A. M. CARON

Center for Economic Research, Tilburg University, PO Box 90153, 5000 LE, Tilburg, The Netherlands
Erasmus Research Institute of Management, Erasmus University Rotterdam, PO Box 1738, 3000 DR, Rotterdam, The Netherlands

SUMMARY

We describe a methodology for explanation generation in financial knowledge-based systems. This offers the possibility to generate explanations and diagnostics automatically to support business decision tasks. The central goal is the identification of specific knowledge structures and reasoning methods required to construct computerized explanations from financial data and models. A multistep look-ahead algorithm is proposed that deals with so-called cancelling-out effects, which are a common phenomenon in financial data sets. Our method is an extension of the traditional variance decomposition in accounting. The method was tested on a case-study conducted for Statistics Netherlands involving the comparison of financial figures of firms in the Dutch retail branch.

1. INTRODUCTION

Competition benchmarking or interfirm comparison (IFC) is defined as the regular 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, the company can assess its performance against objective standards and see where the company is strong or weak. Currently, 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 with some norm behaviour. After abnormal behaviour is detected, the analyst wants to find the causes, i.e. the set of financial variables responsible. The traditional methods frequently used in accounting are variance decomposition and analysis of ratios in a Du Pont model (Fridson and Alvarez, 2002). Today’s information systems for automated financial diagnosis and IFC have little explanation or diagnostic capabilities. Such functionality can be provided by extending these systems with an explanation formalism, which supports the work of human analysts in diagnostic processes. In this paper, we describe how the diagnostic process is fully 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, with the Du Pont and, for example, OnLine Analytical Processing (OLAP) business databases as special cases.

Diagnosis is generally defined as finding the best explanation of observed symptoms of a system under study. This definition assumes that we know which behaviour we may expect from a correctly working system. Diagnosis of business performance is defined in Feelders (1993) as explaining the difference between the actual performance of a company and its norm performance. The norm
performance or normative model can be derived from some statistical model or can be dictated by the financial analysts.

There are many contributions on medical diagnosis and diagnosis of technical devices; see Verkooijen (1993) for an overview. A limited number of approaches have been proposed for the automatic generation of explanations based on financial models (Kosy and Wise, 1984; Hamscher, 1994; Feelders, 1993; Binbasioglu and Zychowicz, 1998; Feelders and Daniels, 2001). Hamscher (1994) proposes a method that automatically constructs explanations for financial results. Moreover, Hamscher (1994) discusses the motivations and foundations of model-based reasoning and diagnosis in financial domains, and surveys several existing artificial intelligence programs for explanation. Binbasioglu and Zychowicz (1998) present a diagnostic knowledge-based system for analysing 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. Feelders and Daniels (2001) describe a formal framework for explanation and diagnosis of company performance with both qualitative and quantitative information. Their method reduces the sets of contributing and counteracting causes to parsimonious sets to avoid the inclusion of insignificant causes. Kosy and Wise (1984) 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.

The rationale behind this paper is to extend the methodology for automated business diagnosis as described in Feelders and Daniels (2001) and Feelders (1993) and its implementation. From the accounting viewpoint, it can also be considered as a generalization and automation of the variance decomposition method. First, a method for symptom detection is presented that takes into account the probability distribution of the variable under consideration for diagnosis. Second, we extend the explanation methodology with a procedure to deal with so-called cancelling-out or neutralization effects in data sets. For example, the first half-year positive financial results could partially cancel out negative financial results of the next half-year. These effects are quite common in financial data and other data sets and could lead to incomplete explanation trees. Finally, the diagnostic program described in Feelders (1993) was implemented in Prolog. This type of implementation has some advantages in terms of knowledge representation, but it also has some clear disadvantages in terms of applicability in an office environment and presentation of program output. Therefore, we have implemented the extended explanation model in MS Excel in combination with Visual Basic.

This paper is part of our continued work on extending multidimensional databases or OLAP data cubes with diagnostic support for business analysis. OLAP is a relatively recent development in the field of decision support systems (DSSs). It not only integrates earlier generations of DSSs, but goes further and introduces spreadsheet-like multidimensional data views and graphical presentation capabilities (Koutsoukis et al., 1999). OLAP databases have a variety of business functions; for example, finance departments use OLAP for applications such as budgeting, activity-based costing and financial performance analysis (Thomsen, 2002). Currently, researchers are working on extending the OLAP framework with diagnostic support (Sarawagi, 2001; Shim et al., 2002; Caron and Daniels, 2007), offering the possibility to generate explanations automatically for exceptional cell values. The explanation methodology described here can easily be integrated for this purpose in an OLAP business database.

This paper is organized as follows. Section 2 reviews the explanation model as described in the literature and proposes improvements for it. In Section 3, the extensions are illustrated in a
case-study on IFC with financial data about Dutch retail companies collected at Statistics Netherlands. Section 4 briefly describes the software implementation of the diagnostic program. Finally, we draw a number of conclusions in Section 5.

2. EXPLANATION

2.1. The Explanation Model

Our exposition of diagnostic reasoning and causal explanation is largely based on the notion of explanations in Feelders and Daniels (2001) and Feelders (1993), which is essentially based on the notion of aleatory explanations (Humphreys, 1989) and the theory of explaining differences (Hesslow, 1983). Causal influences can appear in two forms: contributing and counteracting. The canonical format for causal explanations is given by

\[\langle a, F, r \rangle \quad \text{because } C^+, \text{ despite } C^-\]

where \(\langle a, F, r \rangle\) is the event to be explained, \(C^+\) is a nonempty set of contributing causes and \(C^-\) is a (possibly empty) set of counteracting causes. The explanation itself consists of the causes to which \(C^+\) jointly refers. \(C^-\) is not part of the explanation, but gives a clearer notion of how the members of \(C^+\) actually brought about the symptom. In words, the explanandum is a three-place relation between an object \(a\) (e.g. the ABC-company—the actual behaviour of a company), a property \(F\) (e.g. having a low profitability—the deviation for a particular variable from its norm value) and a reference class \(r\) (e.g. other companies in the same branch or industry—the norm behaviour). The task is to explain why \(a\) has property \(F\) while the other members of \(r\) do not.

2.2. Knowledge Structures for Explanation

Explanations are usually based on general laws expressing relations between events: such as cause–effect relations or constraints between variables. These general laws are represented in the business model \(M\). For example, \(M\) could be a Du Pont ratio model or an OLAP hierarchy. Here, the business model equations are written as

\[y = f(x), \quad \text{where } x = (x_1, x_2, \ldots, x_n)\]

The business model used by Statistics Netherlands for gathering production statistics in the retail branch is given in Section 3.

A directed graph \(G(M) = (v, \varepsilon)\), the explanatory graph, is associated with the business model \(M\). The vertex set \(v\) contains as elements all variables appearing in the model. The edge set \(\varepsilon\) contains a directed edge from vertex \(x_i\) to \(x_j\) iff: \(x_j = f(x_1, x_2, \ldots) \in M\). A restriction is placed on the model \(M\) to exclude cycles in the explanatory graph \(G(M)\). The arcs between the nodes, which represent the variables in the business model, indicate the direction of influence, or causal direction. Interpreting the equality sign in the equations of model \(M\) as \(\leftarrow\) gives the causal direction as used by economists, accountants or financial analysts. Thus, in the model, the effects appear on the left-hand side (LHS) of the equations and the causes on the right-hand side (RHS). However, as we shall see, the diagnostic reasoning direction is the reverse of the causal direction. In other words, the
explanation generation process takes part from the whole (the LHS variables) to the parts (the RHS variables). Moreover, the normative model specifies which reference object(s) should be used to compare. It also specifies the variables with respect to which the comparison should be made.

2.3. Symptom Detection

Diagnosis in a financial model is the explanation for the observed exceptional behaviour of a company. The central question in problem identification for business diagnosis is: ‘Which firms deviate significantly from their branch average or historic average?’ Suppose the normative model contains a reference value for variable $y$. The data set may contain several reference values, besides the actual values for business variables. For diagnosis of company performance, the event to be explained with actual object $a$ and reference object $r$ will always be clear from the context; therefore, the explanation formalism is simplified to: $\partial y = q$ occurred because $C^+$, despite $C^-$. In this expression, $\partial y = y^a - y^r = q$, where $q \in \{\text{low, normal, high}\}$, specifies an event in the financial data set, i.e. the occurrence of a quantitative difference between the actual value and the reference value of $y$, denoted by $y^a$ and $y^r$ respectively.

Problem identification is a process where a value $g(y^a, y^r)$ is computed for each variable, where $g$ is some user-defined function such as percentage or absolute difference. When a statistical model is used as a normative model, then $y^r$ is computed from the statistical model $y^r = E(y)$, where $E(y)$ denotes the expected value of $y$. The normalized residual is denoted by $\partial y/\sigma$, where $\sigma$ is the standard deviation. The exact population parameters of the distribution are usually unknown; therefore, they are estimated and replaced by the sample mean and sample variance. Correspondingly, the problem of looking for exceptional company behaviour is equivalent to the problem of looking for exceptional normalized residuals. Statistically defined, a variable is a symptom or exceptional value if it is higher (lower) than some user-defined threshold $\delta$ ($-\delta$). Usually, we select $\delta = 1.645$, corresponding to a probability of 95% in the standard normal distribution.

2.4. Diagnosis and Explanation

If $\partial y = q$ is identified as a symptom, then we want to explain the difference $\partial y = y^a - y^r$. An explanation is based on the financial equations of the business model. To determine the contributing and counteracting causes that explain the quantitative difference between the actual and reference values of $y$, a measure of influence is defined in the literature (Feelders and Daniels, 2001) as follows:

$$\inf(x_i, y) = f(x^a_i, x^r_i) - y^r$$

where $f(x^a_i, x^r_i)$ denotes the value of $f(x)$ with all variables evaluated at their reference values, except $x_i$. In words, $\inf(x_i, y)$ indicates what the difference between the actual and reference values of $y$ would have been if only $x_i$ had deviated from its reference value.

Here, it is assumed that $y^a = f(x^a_1, x^a_2, \ldots, x^a_n)$ and $y^r = f(x^r_1, x^r_2, \ldots, x^r_n)$. Furthermore, the function $f$ has to satisfy the so-called conjunctiveness constraint (Feelders and Daniels, 2001). This constraint captures the intuitive notion that the influence of a single variable should not turn around when it is considered in conjunction with the influence of other variables. Two important classes of functions satisfy the conjunctiveness constraint, namely additive and monotonic functions. Relations in financial models are almost always additive or monotone. By monotonicity we mean monotonicity...
in all variables separately, on the domain under consideration. The form of the reference function depends on the type of normative model applied. In the situation that the actual and reference functions are both additive, then \( \inf(x, y) \) is interpreted as a quantitative specification of the change in \( y \) that is explained by a change in \( x \): 
\[
\frac{\partial y}{\partial y} = y_a - y_r = \sum_{i=1}^{n} \inf(x_i, y).
\]

The set of contributing (counteracting) causes \( C^+ (C^-) \) consists of measures \( x_i \) of \( x \) with \( \inf(x_i, y) \times \partial y > 0 (\partial y < 0) \). In the explanation method, insignificant influences are left out of the explanation by a filter measure. This corresponds to the concept of materiality in accounting. Therefore, the set of causes is reduced to the so-called parsimonious or significant set of causes. The parsimonious set of contributing causes \( C^+_p \) is the smallest subset of the set of contributing causes such that 
\[
\frac{\inf(C^+_p, y)}{\inf(C^+, y)} \geq T^+.
\]
The parsimonious set of counteracting causes is defined analogously. The fractions \( T^+ \) and \( T^- \) are numbers between 0 and 1, and will typically be 0.85 or so.

Causess can be chained together until a maximal explanation is obtained. The idea is that, for \( \partial y = q \), explanation generation is continued (top down) only for its parsimonious contributing causes. This process is continued until a contributing cause is encountered that cannot be explained within the business model \( M \), because the business model does not contain a relation in which this contributing cause appears on the LHS. The maximal explanation process results in a so-called tree of causes or explanation tree.

2.5. Making Hidden Causes Visible by Substitution

One of the shortcomings of the method described above is that it cannot deal with so-called cancelling-out or neutralization effects. Cancelling-out is the phenomenon that the effects of two or more lower level variables in the business model cancel each other out so that their joint influence on a higher level variable in the business model is partly or fully neutralized. These effects are quite common in financial models, as we shall see in a case-study. The problems with these effects were first mentioned by Kosy and Wise (1984); however, no solution was presented in their article. Nondetected causes by multilevel explanation are called hidden causes. Hidden causes are significant causes that are not visible at first due to the neutralization of a higher level variable in the business model. In theory, cancelling-out effects may occur at every level in the business model. Of course, financial analysts would like to be informed about significant hidden causes, and would consider an explanation tree without mentioning these causes as incomplete and not accurate.

Suppose that we are explaining a symptom \( \partial y = q \) with the following equations out of business model \( M \):
\[
y = f(x) \in M^{0,1} \quad (1)
\]
\[
x_i = g_i(z) \in M^{1,2} \quad (2)
\]

where \( x = (x_1, \ldots, x_n, \ldots, x_p) \) and \( z = (z_1, \ldots, z_m) \) denote \( n \)- and \( m \)-component vectors. The depth of the business model \( d \) is defined as the number of levels in \( M \) or the associated directed graph. The root \( y \) of the tree is on level 0, the children of the root (variables \( x_1, x_2, \ldots, x_n \)) are on level 1, the grandchildren of the root are on level 2, and so on. Furthermore, \( M^{p,q} \) is the set of equations where the variables at level \( p \) appear on the LHS and are expressed in the variables at level \( q \) that appear on the RHS of the equations. For variables that cannot be expressed in variables at level \( q \) we use variables at a higher level closest to level \( q \). Figure 1 shows an explanatory graph with two example equations, \( y = f(x_1, x_2) \) and \( x_2 = g_2(z_1, z_2) \), in the LHS.
Furthermore, suppose that explanation generation with equation (1) results in sets of parsimonious causes where variable $x_i$ is not part of: thus $x_i \not\in C_p^+(y)$ and $x_i \not\in C_p^-(y)$. 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 $\partial y$. To make sure that the explanation is complete, all successors of $x_i$ (the elements of $z$) are substituted into the RHS of equation (1) to derive the substituted function

$$y = h_i(x, z) \in M^{0,2}$$

The result of substituting jointly all equations at level $M^{p,q+1}$ in the business model into a parent equation in $M^{p,q}$ is denoted by $M^{p,q+1}$, this is called one-step look-ahead. In the RHS of Figure 1, the equation $y = h_3(x_1, z_1, z_2)$ is depicted as the explanatory graph $M^{0,2}$. Subsequently, the substituted equation is added to the business model $M$ and considered for explanation generation.

We define a hidden cause as follows.

**Definition 1.** Variable $z_j$ of equation (3) is a hidden cause when $z_j \in C_p(y)$ and $x_i \not\in C_p(y)$, where $z_j$ is a successor of $x_i$. Here, the influence of $z_j$ on $y$ is given by

$$\inf(z_j, y) = f(x^r, g_i(z^r, z^r_j)) - f(x^r, g(z^r))$$

and the influence of $x_i$ on $y$ is given by

$$\inf(x_i, y) = f(x^r, x_i^r) - f(x^r) = f(x^r, g(z^r)) - f(x^r, g(z^r))$$

This means that the effect of $z_j$ is neutralized by the effects of other variables in the vector $z$. Moreover, 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 (1) and (2) are both additive we have $\inf(x_i, y) = \sum_{j=1}^n \inf(z_j, y)$. From this relation it immediately follows that, when $x_i \not\in C_p^+(y)$ and $z_j \in C_p^+(y)$, at least one variable out of $z$ is in the set of counteracting causes $C^-(y)$. Or vice versa, when $x_i \not\in C_p^-(y)$ and $z_j \in C_p^-(y)$, at least one variable out of $z$ is in the set of contributing causes $C^+(y)$.

One-step look-ahead can simply be extended to multistep look-ahead. For example, two-step look-ahead is defined as one-step look-ahead plus $M^{p,q+2}$, the result of substituting all equations at level $M^{p,q+2}$ into $M^{p,q+1}$, and so on. In general, for a business model with depth $d$, the maximal number of look-ahead steps is $d - 1$. In multistep look-ahead, a successor of variable $x_i$ is a hidden cause if its influence on $y$ is significant after substitution, when the influence of variable $x_i$ of equation (1) on $y$ is not significant.
The look-ahead algorithm is composed of two consecutive phases: (1) an analysis and (2) a reporting phase. In the analysis phase, the explanation generation process starts (similar to the maximal explanation) with the root equation in the business model by determining parsimonious causes. However, instead of proceeding with strictly parsimonious causes, all non-parsimonious contributing and counteracting 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 function substitution. Moreover, the substituted functions are added to $M$ and considered for explanation generation. In the reporting phase, the explanation tree is updated when hidden causes are detected. In updating the tree, new parsimonious causes (e.g. RHS elements of equation (2)) are added and causes that have become non-parsimonious (e.g. RHS elements of equation (1)) are removed. This method is applied in a multistep look-ahead algorithm with two repetitive phases. The pseudo code of the algorithm is given in Appendix B.

3. CASE-STUDY: INTERFIRM ANALYSIS AT STATISTICS NETHERLANDS

The business model and data for IFC in this case-study are obtained from Statistics Netherlands (Statistics Netherlands, C.B.S., 2007). Statistics Netherlands is responsible for collecting, processing and publishing statistics to be used in practice, by policymakers and for scientific research. The business model $M$ we present here is derived from the production statistics for companies in the Dutch retail and wholesale trade. We use production statistics from two consecutive years, namely the years 2001 and 2002. For both years, data sets with more than 5000 different retail and wholesale companies are classified into branch sections. The following business model equations are used for production statistics:

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. $r_{15} = r_{21} + r_{22}$

$: :$

19. $r_{33} = r_{75} + r_{76} + r_{77} + r_{78} + r_{79} + r_{80} + r_{81}$

In short, three types of business equation are identified in $M$ (with depth $d = 4$): result, equations 1–5; revenue, equations 6–8; cost, equations 9–19. 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 result; 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 (like total operating revenues $r_6$, financial revenues $r_8$, deductions from allowances $r_{10}$ and extraordinary profits $r_{12}$) and a costs component (like 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 collection of interests received, revenues from participations, payments of dividends, and profits from investments and other financial gains. The additions to allowances \( r_{14} \) are the sum of additions to internal provident funds, like 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 the business model \( M \). The descriptions of the other variables in \( M \) are given in Appendix A. In the case-study example, \( M \) consists mainly of additive and difference relations. Obviously, our explanation methodology can also handle multiplicative and division relations (Feelders and Daniels, 2001).

For diagnosis of business performance we have to determine appropriate reference objects. Several factors that influence the business diagnosis results have to be taken into account, like the Standard Industry Classification (SIC) for the retail and wholesale industry, and the size of the company. Therefore, computerized selections on the data set are made, like supermarkets, liquor stores, do-it-yourself shops, etc. Within these subsets we make a further selection on the size class (small, medium and large) of the companies. The company size classes are based on the number of employees of the firm in full-time employees (FTEs). The intervals for the different size classes are: small, 1–9 employees; medium 10–99 employees; large, \( \geq 100 \) employees. In this way, homogeneous subsets of the data for analysis are constructed. In addition, for the analysis, data are normalized by dividing all variables in \( M \) by the total number of FTEs of each individual company. The reference object 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 \((r_1 - r_81)\) in the business model. Moreover, for historic comparisons, the reference objects for the business model variables are the values in one or more previous time periods; for example, we can benchmark the results for the current year with the results of the previous year for a certain company.

3.1. Symptom Detection

Analysis is performed on a specific homogeneous sample selected out of the original data set with production statistics for the year 2001. The sample selected is composed of 69 fashion shops out of the size class ‘medium’. Problem identification in the data set starts with the variable \( r_1 \), total result before taxation, on the root level of the business model. This variable has a normal distribution (tested with the Shapiro–Wilks normality test) with mean 11.30 (the industry average) and standard deviation 28.85. The exact population parameters of the distribution are unknown; therefore, they are estimated and replaced by the sample mean and sample variance. The central question in problem identification for this case study is: ‘Which firms deviate significantly from their branch average in 2001?’ The symptom detection module of the diagnosis application identifies nine firms that are higher or lower than the specified threshold value in the sample data set; see Table I for 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: five firms with symptom ‘high’, 60 firms with symptom ‘normal’ and four firms with symptom ‘low’.

For one of the fashion shops in the sample (the ABC-company) we present complete diagnostics. Moreover, the data are anonymized, because Statistics Netherlands does not allow exposure of data
on the micro level. 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 with its industry average – more than four times as good.

### 3.2. Example Explanation Generation for ABC-Company

We analyse the symptom

\[<\text{ABC-company}(2001), \partial r_1 = \text{high}, \text{branch average}(2001)>\]

using the explanation method with one-step look-ahead. Explanation generation with look-ahead starts with the root equation in \(M\). We take \(T^+ = T^- = 0.85\). In Table II, a comparison is made between the actual total result before taxation of the ABC-company and the branch average in the year 2001. From the data in Table II we infer that \(C^+_p = \{r_2\}\) and \(C^-_p = \phi\). The variable \(r_2\) (total operating results) explains 90.44% of the difference \(\partial r_1\) and, therefore, is identified as the single parsimonious contributing cause because its value exceeds the fraction. Thus, 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. Therefore, the variable \(r_2\) is the single child node of its parent (root node) \(r_1\) in the explanation tree.

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 2–5 in \(M\). This is illustrated graphically in the partial explanation tree depicted in Figure 2, where the curved black arrows ‘step over’

<table>
<thead>
<tr>
<th>Slot name</th>
<th>Slot entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>Total result before taxation (r_1)</td>
</tr>
<tr>
<td>Norm object</td>
<td>Industry average (2001)</td>
</tr>
<tr>
<td>Industry</td>
<td>Fashion shops</td>
</tr>
<tr>
<td>Size class (69 firms)</td>
<td>Medium</td>
</tr>
<tr>
<td>Distribution</td>
<td>Normal distribution (r_1 \sim N(11.30, 832.17))</td>
</tr>
<tr>
<td>Threshold</td>
<td>(\alpha = 0.05) (two one-tailed tests)</td>
</tr>
</tbody>
</table>

### Table II. Actual and norm values for \(r_1 = r_2 + r_3 + r_4 + r_5\)

<table>
<thead>
<tr>
<th></th>
<th>Actual</th>
<th>Norm</th>
<th>(\text{inf}(x, y))</th>
<th>Difference (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(r_1)</td>
<td>61.75</td>
<td>11.30</td>
<td></td>
<td>446.46</td>
</tr>
<tr>
<td>(r_2)</td>
<td>60.42</td>
<td>14.79</td>
<td>45.62</td>
<td>308.52</td>
</tr>
<tr>
<td>(r_3)</td>
<td>1.33</td>
<td>-2.55</td>
<td>3.88</td>
<td>-152.16</td>
</tr>
<tr>
<td>(r_4)</td>
<td>0.00</td>
<td>-0.15</td>
<td>0.15</td>
<td>-100.00</td>
</tr>
<tr>
<td>(r_5)</td>
<td>0.00</td>
<td>-0.79</td>
<td>0.79</td>
<td>-100.00</td>
</tr>
</tbody>
</table>
In the analysis phase, function substitution is applied to find parsimonious causes, which were missed in the local explanation of differences. Equations 2–5 are substituted into the root equation and the following equation for explanation generation is derived:

\[ M^{0.2}: r_1 = (r_6 - r_7) + (r_8 - r_9) + (r_{10} - r_{11}) + (r_{12} - r_{13}) \]

This equation is added to the set of business model equations. Notice that the specification of the event to explain \( \partial r_1 \) remains the same, but now equation \( M^{0.2} \) is applied to explain the difference.

Table III summarizes the results of our extended model of ABC-company’s relatively high total result before taxation. From the data in Table III, it follows that \( C^+ = \{r_6, r_7, r_8\} \) and \( C^- = \{r_9\} \).

Table III. Actual and norm values for \( M^{0.2} \): 
\[
\begin{array}{cccc}
\text{Component} & \text{Actual} & \text{Norm} & \text{inf}(x, y) & \text{Difference} \% \\
\hline
r_1 & 61.75 & 11.30 & & 466.31 \\
r_6 & 329.50 & 308.64 & & 20.86 & 6.76 \\
r_7 & 269.09 & 293.84 & & 24.76 & -8.42 \\
r_8 & 11.17 & 1.84 & & 9.33 & 507.07 \\
r_9 & 9.83 & 4.39 & & -5.44 & 123.92 \\
r_{10} & 0.00 & 0.16 & & 0.16 & -100.00 \\
r_{11} & 0.00 & 0.01 & & -0.01 & -100.00 \\
r_{12} & 0.00 & 0.31 & & -0.31 & -100.00 \\
r_{13} & 0.00 & 1.10 & & 1.10 & -100.00 \\
\end{array}
\]
The diagnostic process is continued for all significant contributing causes. Thus, the new events to be explained are specified as

\[ \langle \text{ABC-company}(2001), \partial r_6 = \text{high}, \text{branch average}(2001) \rangle \]

and

\[ \langle \text{ABC-company}(2001), \partial r_7 = \text{low}, \text{branch average}(2001) \rangle \]

For these model equations the influence values are omitted from the paper because of space limitations. The previous examples of different one-level explanations are now combined to a complete tree of causes. Figure 4 summarizes the results of the diagnostic process.

The following economic interpretation is given to the explanation tree in Figure 4. Why are the ABC-company’s total results before taxation relatively high compared with 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 are due to a combination of comparatively high...
total operating results $r_2$ and comparatively high financial revenues $r_8$, 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.

4. SOFTWARE IMPLEMENTATION

The software is implemented in MS Excel in combination with Visual Basic. The most important part of the program is the diagnostic component. It contains the method for maximal explanation and the multi-step look-ahead algorithm. For the implementation of the procedure we applied tree programming to generate the tree of causes. The tree-viewer interface of the program is depicted in Figure 5.

![Figure 5. Tree viewer in diagnosis application](image_url)
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 colour. By clicking on the cause under consideration, the details for the cause become visible in the right panel of the screen.

5. SUMMARY AND CONCLUSION

We have extended the method for automated business diagnosis in Feelders and Daniels (2001) and Feelders (1993). The explanation model is extended in two ways: in the symptom detection phase, the probability distribution of business model variables is taken into account; and in the explanation generation phase, hidden causes can be made visible by function substitution. The problem of looking for exceptional company behaviour in financial data sets is translated into the problem of looking for exceptional normalized residuals. In this way, a statistical definition for a symptom is derived. Furthermore, the multilevel look-ahead algorithm improves the explanation methodology so that it can deal with cancelling-out effects. In the software, special attention is given to presentation of the program output, where symptoms and causes are presented graphically as a tree of causes. In this manner, a manager or financial analyst can view and access the results of the explanation process for diagnosis of company performance as a compact tree. The applicability of the method is illustrated in a case-study on IFC in the Dutch retail and wholesale trade, based on production statistics obtained from Statistics Netherlands. In the case-study, it is shown that, in the presence of cancelling-out effects, the extended model with the look-ahead procedure makes significant causes visible that would have been missed by the explanation methodology of maximal explanation.

Currently, we are working on an extension of the explanation method in the framework of OLAP. This functionality is integrated with conventional OLAP business databases and is a powerful tool for automated analysis of multidimensional data.

ACKNOWLEDGEMENTS

We thank Jeffrey Hoogland for his support at Statistics Netherlands.

APPENDIX A: LIST OF VARIABLES

Because of space limitations, only the meaning of the variables identified by the explanation model are described here in detail. The variable descriptions have been translated from the original Dutch surveys. The complete list of variables and their definitions is available upon request.

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

**APPENDIX B: ALGORITHM PSEUDO CODE**

*Algorithm: Multilevel look-ahead*

1. \( y \) is the root node of the tree
2. for \( p = 0 \) to \( d - 1 \) do
3. determine parsimonious causes for equation(s) \( M^{p+1} \)
4. add parsimonious causes to the tree as successor nodes
5. if look-ahead is activated then
6. for \( i = 1 \) to \( d - 1 \) do
7. substitute jointly all equations on \( M^{p+1}\) into equation \( M^{p+1} \)
add derived equation $M^{p+i+1}$ to $M$

determine parsimonious causes for $M^{p+i+1}$

if causes on level $p+i+1$ are parsimonious then

add new parsimonious causes as successor nodes to the tree

remove non-parsimonious causes to the tree

if a node corresponds to counteracting cause then

it has no successors

if a node corresponds to variable that cannot be explained in $M$ then

it has no successors

REFERENCES


