EXPLANATION GENERATION IN BUSINESS PERFORMANCE MODELS
With a Case Study in Competition Benchmarking

Hennie Daniels[1,2] and Emiel Caron[2]
[1]*Center for Economic Research, Tilburg University, P.O. Box 90153, 5000 LE, Tilburg, The Netherlands*

[2]*Erasmus Research Institute of Management (ERIM), Erasmus University*
P.O. Box 90153, 3000 DR, Rotterdam, The Netherlands

Keywords: Decision Support Systems, Finance, Production Statistics, Artificial Intelligence, Explanation.

Abstract: In this paper, we describe an extension of the methodology for explanation generation in financial knowledge-based systems. This offers the possibility to automatically generate diagnostics 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 multi-step look-ahead algorithm is proposed that deals with so-called cancelling-out effects. The extended methodology was tested on a case-study conducted for Statistics Netherlands involving the comparison of financial figures of firms in the Dutch retail branch. The analysis is performed with a diagnostic software application which implements our theory of explanation. Comparison of results of the method described in (Daniels and Feelders, 2001) with the results of the extended method clearly improves the analyses when cancelling-out effects are present in the data.

1 INTRODUCTION

Competition benchmarking or interfirm comparison (IFC) is defined as the regular measuring and comparing of a company's performance against its competitors, against industry leaders or industry and historic averages. The aim is often to learn how the company can improve its own performance. 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 business analysts (a.o. bankers, accountants and consultants). The analyst has to explore large data sets in the domain of business and finance to spot firms that expose exceptional behaviour compared to some norm behaviour. After abnormal behaviour is detected the analyst wants to find the causes, the set of financial variables responsible, that explain the firm's behaviour. Today's information systems for automated financial diagnosis and interfirm comparison have little explanation or diagnostic capabilities. Such functionality can be provided by extending these systems with an explanation formalism, which mimics the work of human analysts in diagnostic processes. In this paper the diagnostic process is fully automated and implemented in a computer program to support decision-makers.

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 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 expert knowledge from financial analysts. Two important consecutive phases in a diagnostic process are *problem identification* (or symptom detection) and *explanation generation* (Verkooijen, 1993). When a discrepancy between actual and norm behaviour is discovered, and is qualified as exceptional with respect to some specified norm, the next step is to explain this discrepancy using our "understanding" of the system. 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 expla-

Daniels H. and Caron E. (2007).
EXPLANATION GENERATION IN BUSINESS PERFORMANCE MODELS - With a Case Study in Competition Benchmarking.
In *Proceedings of the Ninth International Conference on Enterprise Information Systems - AIDSS*, pages 119-128
DOI: 10.5220/0002346101190128

nations based on financial models (Courtney et al., 1987; Daniels and Feelders, 2001; Feelders, 1993; Kosy and Wise, 1984).

The rationale behind this paper is to extend the methodology for automated business diagnosis as described in (Daniels and Feelders, 2001; Feelders, 1993). Firstly, a method for symptom detection is presented that takes into account the probability distribution of the variable under consideration for diagnosis. The detection of symptoms for computerized diagnosis in financial data is not fully developed in earlier methods, where it is described as the process of taking the difference value between the actual and norm value of each variable. Secondly, in this paper we extend their explanation methodology, with a procedure to deal with so-called *cancelling-out* or *neutralisation effects* in data sets. For example, the first half-year positive financial results could partially cancel out the negative financial results of the next half-year in a financial model. If one starts diagnosis with the method described by Daniels and Feelders on the aggregated year level these effects are not identified. However, these effects are quite common in financial data and other economic data sets and could lead to results in the form incomplete explanation trees.

This paper is organised as follows. In the next section we first review the explanation model as described in literature and introduce extensions for it. In section 3 the extensions are illustrated by an extensive case study on interfirm comparison with financial data collected at Statistics Netherlands. In the case study we compare two explanations, in the form of trees of causes, for detected symptoms derived from companies in the Dutch retail industry. We compare the trees generated with and without the look-ahead facility. In section 4 we briefly describe the software implementation of the diagnostic program. Finally, we draw a number of conclusions in section 5. In the appendix the list of variables and data for interfirm comparison used in the case study are included.

2 METHODOLOGY

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

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

where $\langle a, F, r \rangle$ is the event to be explained, C^+ is non-empty set of contributing causes, and C^- 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) that shows the actual behaviour of a company, a property F (e.g. having a low profitability) that shows 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) that shows the norm behaviour. The task is not to explain why a has property F, but rather to explain why a has property F *when the other members of r do not*. This general formalism for explanation constitutes the basis of our extended framework for diagnosis in financial models developed in this paper.

Two principal knowledge structures for diagnosis of business performance are identified:

- Knowledge of general laws, relating variables pertaining to business performance: the *business model*;
- Knowledge of normal behaviour: the *normative model*.

In this section we present a summary and propose some extensions on the general theory and methodology for automated business diagnosis.

2.1 The Business Model

Explanations are usually based on general laws expressing relations between events: such as cause-effect relations or constraints between variables. The general laws on which explanations are based, are represented in the business model M. The business model M represents quantitative financial and operating variables by means of mathematical equations of the form:

$$y = f(\mathbf{x}) \text{ where } \mathbf{x} = (x_1, x_2, \ldots, x_n).$$

The business model is used to propagate both deviating and non-deviating values. In section 3, an example is given of a business model used by Statistics Netherlands for gathering production statistics in the retail branch.

A directed graph, the *explanatory graph* $E(M) = (\mathcal{V}, \mathcal{E})$, is associated with the business model M. The vertex set $\mathcal{V}$ contains as elements all variables appearing in the model. The edge set $\mathcal{E}$ contains a directed

edge from vertex x_i to x_j iff: $x_j = f(\ldots, x_i, \ldots) \in M$. A restriction is placed on the model M to exclude cycles in the explanatory graph $E(M)$. The arcs between the nodes in the graph, which represent the variables in the business model, indicate the direction of influence, or causal direction. Interpreting the $=$ in the equations of model M as a $\leftarrow$ 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).

2.2 The Normative Model

Information seeking or gathering for decision recognition and diagnosis involves both a search for symptoms and a search for causes. Pounds (Pounds, 1969) found that the need for a decision is identified as a perceived difference between the actual situation and some normative model, the expected standard. This model could be based on either trends past or projected, comparable situations inside or outside the organization, expectations of other people or on theoretical models. With the exception of crisis, these differences normally do not present themselves readily to the decision maker but must be filtered from the constant streams of ambiguous data received. 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. The most common "reference objects" to diagnose business performance are: historical reference values, industry averages and plans and budgets.

2.3 Symptom Detection

Diagnosis in a financial model is the explanation for observed exceptional behaviour of a company. The first step in diagnostic process is problem or symptom identification, the detection of abnormal behaviour. 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* and the *reference value* of y, denoted by y^a and y^r, respectively. Note that for the purpose of diagnosis, it is not interesting to explain symptoms with the label $\partial y =$ "*normal*", since it is only required to explain why a variable deviates from its reference value.

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. Here a method is developed that can take into account the probability distribution, e.g. the normal distribution, of the business variable under consideration. In this method first the average value for each variable is estimated based on a statistical model. When a statistical model is used as a normative model then $y^r = \hat{y}$. If we now normalize the residual of the model by the standard deviation σ of the variable in the sample, we get the normalized residual $\partial y/\sigma$. 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 δ ($-\delta$). Usually, we select $\delta = 1.645$ corresponding to a probability of 95% in the standard normal distribution. Automatically, the following series of statistic tests is performed on each variable in the business model to detect symptoms in the data set under consideration:

- if $\partial y/s > \delta$ (one-tailed test) then the symptom is labelled $\partial y =$ "*high*",
- if $\partial y/s < -\delta$ (one-tailed test) then the symptom is labelled $\partial y =$ "*low*" and
- if $-\delta \leq \partial y/s \leq \delta$ then the symptom is labelled $\partial y =$ "*normal*".

2.4 Diagnosis and Explanation

If $\partial y = q$ is identified as a symptom, 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 value of y, a *measure of influence* is defined in literature (Daniels and Feelders, 2001; Feelders, 1993; Kosy and Wise,

1984) as follows:

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

where $f(\mathbf{x}^r_{-i}, x^a_i)$ denotes the value of $f(\mathbf{x})$ with all variables evaluated at their reference values, except x_i. In words, $\inf(x_i, y)$ indicates what the difference between the actual and reference value of y would have been if *only* x_i would have deviated from its reference value. 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* (Daniels and Feelders, 2001; Feelders, 1993). 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 classes of functions satisfy the conjunctiveness constraint, namely *additive* and *monotonic functions*, that frequently occur in business model relations. By monotonicity we mean the monotonicity in all variables separately, on the domain under consideration. The form of the reference function depends on the type of statistical model applied. In the situation that actual and reference function are both additive, then $\inf(x_i, y)$ is correctly interpreted as a quantitative specification of the change in y that is explained by a change in x_i:

If $f = \sum_{k=1}^{n} s_k(x_k)$ (actual function is additive; s_k's are arbitrary functions) and $y^r = \sum_{k=1}^{n} s_k(x^r_k)$ then

$$\partial y = y^a - y^r = \sum_{i=1}^{n} \inf(x_i, y)$$

$$y^a - y^r = \sum_{k=1}^{n} s_k(x^a_k) - \sum_{k=1}^{n} s_k(x^r_k)$$

$$\inf(x_i, y) = \sum_{k \neq i}^{n} s_k(x^r_k) + s_i(x^a_i) - y^r = s_i(x^a_i) - s_i(x^r_i)$$

and therefore: $\sum_{i=1}^{n} \inf(x_i, y) = y^a - y^r$

Furthermore, if f is non-additive but differentiable, $y^r = f(\mathbf{x}^r)$ and $\delta_i = x^a_i - x^r_i$ is small then $\partial y \approx \sum_{i=1}^{n} \inf(x_i, y)$. However, in general ∂y is not necessarily equal to $\partial y = \sum_{i=1}^{n} \inf(x_i, y)$. This occurs when $y^r \neq f(\mathbf{x}^r)$, or when f is non-additive and $\delta_i = x^a_i - x^r_i$ is large. For monotonic functions, the interpretation of $\inf(x_i, y)$ becomes more difficult and context-dependent, but the sign of $\inf(x_i, y)$ is not context-dependent. Therefore sometimes reference values are made *internally consistent* in this situation to maintain the assumption of $y^r = f(\mathbf{x}^r)$.

The definition of the influence-measure makes it possible to operationalize the concepts of contributing and counteracting causes. When explanation is supported by a business model equation, the *set of contributing* (*counteracting*) *causes* C^+ (C^-) consists of measures x_i of $\mathbf{x}$ with $\inf(x_i, y) \times \partial y > 0$ (< 0). In the explanation method, insignificant influences are left out of the explanation by a *filter measure*. 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 $\inf(C^+_p, y) / \inf(C^+, y) \geq T^+$. The parsimonious set of counteracting causes is defined analogously. The fraction T^+ and T^- are numbers between 0 and 1, and will typically 0.85 or so.

Furthermore, in (Daniels and Feelders, 2001; Feelders, 1993) the concept of the *maximal explanation* method is defined. The idea is that for $\partial y = q$, explanation generation is continued (top-down) only for its parsimonious contributing causes, whereas non-parsimonious causes and counteracting causes are not explained any further. 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. Maximal explanation extends the idea of *one-level* explanations, that is based on only one relation from the business model, to *multi-level* explanations. The maximal explanation process results in a so-called *tree of causes* (or explanation tree), where y is the root of the tree and its children, grandchildren, great-grandchildren and so on are parsimonious contributing and counteracting causes. In this way explanations are chained together and a tree of causes is formed.

2.5 Making Hidden Causes Visible by Substitution

The explanation methodology as described in the literature (Daniels and Feelders, 2001; Feelders, 1993; Kosy and Wise, 1984) has the shortcoming that it cannot deal with so-called *cancelling-out* or *neutralisation 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 the case study. For the top-down explanation generation process this means that in some data sets possible significant causes for a symptom will not be detected when cancelling-out effects are present. These non-detected causes by multi-level explanation are called *hidden causes*. Hidden causes are significant causes that are not visible at first due to the

neutralisation of a higher level variable in the business model. In theory, cancelling-out effects may occur at every level in the business model. Therefore, one does not have a clue a priori on what level in the business model detection for these effects should start and whether these effects are significant or not. 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

$$y = f(\mathbf{x}) \in M^{0;1}, \tag{1}$$

$$x_i = g_i(\mathbf{z}) \in M^{1;2}. \tag{2}$$

Where $\mathbf{x} = (x_1, \ldots, x_i, \ldots x_n)$ and $\mathbf{z} = (z_1, \ldots, z_m)$ denote n and m-component vectors. The *depth of the business model* (N) is defined as the number of levels in M or associated directed graph. The root of the tree (y) 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, $M^{LHS(l_p);RHS(l_q)}$ represents the set of equations with the LHS on level l_p and the RHS on level l_q of M. We write $M^{p;q}$ for short, where $p = 0, 1, \ldots, N-1$ and $q = p+1$. The root equation is represented by $M^{0;1}$, the equations for its children are represented by $M^{1;2}$, the equations for its grandchildren are represented by $M^{2;3}$, and so on. Furthermore, suppose that explanation generation with eq. 1 results in sets of parsimonious causes where *variable x_i is not part of*, thus $x_i \notin C_p^+(y)$ and $x_i \notin C_p^-(y)$. 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 ∂y. To make sure that the explanation is complete all successors of x_i have to be investigated for possible cancelling-out effects. Therefore, all children of x_i (the elements of $\mathbf{z}$) are substituted into the RHS of eq. 1 to derive the substituted function

$$y = h_i(\mathbf{x}, \mathbf{z}) \in M^{0;2}. \tag{3}$$

The result of substituting *jointly* all equations at level $M^{q;q+1}$ in the business model into a parent equation $M^{p;q}$ is denoted by $M^{p;q+1}$, this is called *one-step look-ahead*. Subsequently, the substituted equation is *added to the business model M* and considered for explanation generation.

Definition 1. Variable z_j of eq. (3) is a hidden cause when $z_j \in C_p(y)$ under the condition that $x_i \notin C_p(y)$.

In words, variable z_j of eq. (3) (the result of substituting eq. (2) into eq. (1)) is a hidden cause when its influence on y – its grandparent – is significant in explanation generation, under the condition that the influence of variable x_i – its parent – of eq. (1) on y is not significant. Here the influence of z_j on y is given by: $\inf(z_j, y) =$
$f(x_1^r, \ldots, g_i(z_1^r, \ldots, z_j^a, \ldots, z_m^r), \ldots, x_n^r) -$
$f(x_1^r, \ldots, g_i(z_1^r, \ldots, z_j^r, \ldots, z_m^r), \ldots, x_n^r)$,
and the influence of x_i on y is given by: $\inf(x_i, y) =$
$f(x_1^r, \ldots, x_i^a, \ldots, x_n^r) - f(x_1^r, \ldots, x_i^r, \ldots, x_n^r) =$
$f(x_1^r, \ldots, g_i(\mathbf{z}^a), \ldots, x_n^r) - f(x_1^r, \ldots, g_i(\mathbf{z}^r), \ldots, x_n^r)$.
This means that the effect of z_j is neutralized by the effects of other variables in the vector $\mathbf{z}$. Moreover, it is assumed that the derived function h_i satisfies the conjunctiveness constraint. In the special case that the functions f and g_i from eq. (1) and (2) are both additive the following holds: $\inf(x_i, y) = \sum_{j=1}^{m} \inf(z_j, y)$. From this relation it immediately follows that when $x_i \notin C_p^+(y)$ and $z_j \in C_p^+(y)$ at least one variable out of $\mathbf{z}$ is in the set of counteracting causes $C^-(y)$. Or vice versa, when $x_i \notin C_p^+(y)$ and $z_j \in C_p^-(y)$ at least one variable out of $\mathbf{z}$ is in the set of contributing causes $C^+(y)$.

In addition, one-step look-ahead can simply be extended to multi-step look-ahead. For example, *two-step look-head* is defined as one-step look-ahead plus $M^{p;q+2}$, the result of substituting all equations at level $M^{q+1;q+2}$ into $M^{p;q+1}$, *three-step look-head* is defined as two-step look-ahead plus $M^{p;q+3}$, the result of substituting all equations at level $M^{q+2;q+3}$ into $M^{p;q+2}$, and so on. In general, for a business model with depth N, the maximal number of look-ahead steps is $N-1$. In multi-step 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 eq. (1) on y is not significant. Basically, the *multi-step look-ahead method* is an extension of the maximal explanation method (Daniels and Feelders, 2001; Feelders, 1993). In short, the look-ahead method is composed of two consecutive phases: an *analysis* (1) and a *reporting phase* (2). In the analysis phase the explanation generation process starts, similar as for 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*, where all the lower-level equations at level j in the business model are substituted into the higher-level equation under consideration for explanation. In addition, 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 by the multi-

level look-ahead method. As in maximal explanation causes are presented to the analyst in the form of a tree of causes. In fact, the explanation tree generated with maximal explanation needs to be updated when significant hidden causes are present in the substituted equations. In updating the tree new parsimonious causes are added and causes that have become non-parsimonious are removed.

The look-ahead functionality, when activated, is applied as an extension of maximal explanation and executed each time after parsimonious causes have been determined with one-level explanation. When the multi-step look-ahead algorithm is configured with $p = 0$ (explanation starts with the root equation) and maximal number of look-ahead steps $(N-1)$ then all significant (hidden and non-hidden) causes are made visible by substitution and maximal explanation is only used for initialization. The number of look-ahead steps (the horizon) in the business model is user-defined and based on the domain knowledge of the analyst. The pseudo code of the multi-step look-ahead algorithm is given in the Appendix.

3 INTERFIRM ANALYSIS AT STATISTICS NETHERLANDS

The business model and data for IFC in this case study are obtained from Statistics Netherlands (Statistics Netherlands, 2006). Statistics Netherlands is responsible for collecting, processing, and publishing statistics used in practice, by policymakers and for scientific research. The business model M we present is based on the survey structure for gathering production statistics from companies in the Dutch *retail* and *wholesale trade*. In addition, we use production statistics from two consecutive years, the year 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 relations and financial model variables are used

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. $r_{14} = r_{16} + r_{17} + r_{18} + r_{19} + r_{20}$
8. $r_{15} = r_{21} + r_{22}$
9. $r_7 = r_{23} + r_{24} + r_{25} + r_{26} + r_{27} + r_{28} + r_{29} + r_{30} + r_{31} + r_{32} + r_{33} + r_{34}$

$\vdots \qquad\qquad \vdots$

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

In short, three types of business equations are identified in M (with depth $N = 4$): *result* (eq. 1 through 5), *revenue* (eq. 6 through 8), and *cost* (eq. 9 through 19) *equations*. The variable (r_1) in the root result equation gives the company's total result before taxation. This variable is split up into four types of results namely: total operating results (r_2), total financial results (r_3), total results allowances (r_4), and total extraordinary results (r_5). In the Appendix the descriptions of the other variables in M are given.

Several factors that may have an influence on 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 FTE's (full-time employees) and the intervals for the different size classes are: small (1 through 9 employees), medium (10 through 99 employees) and large (from 100 employees and more). In this way homogeneous subsets of the data for analysis are constructed. In addition, we *normalised the data* by dividing all variables in M by the total number of FTE's of each individual company. Reference objects. 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 through 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 selected sample is composed out of 69 fashion shops out of the size class "medium". Problem identification in the data set starts with the variable results for taxation (r_1) 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 pop-

ulation 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 9 firms that are higher (or lower) than the specified threshold value in the sample data set (see Table 1 for a full specification of the norm model). Here we select $\delta = 1.645$ corresponding to a probability of 95% in the standard normal distribution. With these test specifications we derive the following distribution of the number of firms over the three symptom types: 5 firms with symptom high, 60 firms with symptom normal and 4 firms with symptom low.

Table 1: Specification of normative model for example.

slot name	slot entry
variable	results before taxation (r_1)
norm object	industry average (2001)
industry	fashion shops
size class (69 firms)	medium
distribution	$r_1 \sim N(11.30, 832.17)$
threshold	$\alpha = .05$ (two one-tailed tests)

For one of the fashion shops in the sample – the ABC-company – we present complete diagnostics. Moreover, the data is anonimized 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 good compared to its industry average, more than 4 times as good.

3.2 Example Explanation Generation

In this section a comparison is made between the results of two explanations for the symptom: $\langle$ABC-company(2001), $\partial r_1 = high$, branch_average (2001)$\rangle$. We will address the following question:

"*Why are results before tax* (r_1) *relatively high for the ABC-company compared with its branch average?*"

The explanation for this event is generated by maximal explanation method. However this method will not give the complete explanation in the case of cancelling-out effects. Moreover, a comparison between human analysis and the classic explanation method shows noticeable differences when these effects occur. Therefore, we present a second explanation generated with detection for hidden causes switched on. The two explanations and additional explanation trees are both generated automatically by our prototype computer program.

3.2.1 Maximal Explanation Generation

The maximal explanation without look-ahead yields the following results, taking for the fraction $T^+ = T^- = 0.85$. In Table 2 a comparison is made between the actual results before taxation of the ABC-company and the branch average in the year 2001. From the data in the table we infer that $C_p^+ = \{r_2\}$ and $C_p^- = \emptyset$. The variable r_2 (total operating results) explains 90.44% of the difference ∂r_1, and is therefore identified as the single parsimonious contributing cause because its value exceeds the fraction. 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.

Table 2: Actual and norm values for $r_1 = r_2 + r_3 + r_4 + r_5$.

	actual	norm	inf(x_i, y)	diff. %
r_1	61.75	11.30		446.46
r_2	60.42	14.79	45.62	308.52
r_3	1.33	-2.55	3.88	-152.16
r_4	0.00	-0.15	0.15	-100.00
r_5	0.00	-0.79	0.79	-100.00

The diagnostic process is continued only for this parsimonious contributing cause. Further explanation is generated by equation 2 of M, to explain the initial difference in ∂r_1. Explanation generation with the multi-step look-ahead algorithm shows that cancelling-out effects are present in this example. And hidden causes that standard are left undetected by maximal explanation are found in the look-ahead procedure. The next event (analogous to the previous example) to be explained is specified as: $\langle$ABC-company(2001), $\partial r_2 = high$, branch_average(2001)$\rangle$. Table 3 summarizes the results for the explanation of the ABC-company's relative high total operating result. From the data in the table it follows that $C_p^+ = \{r_6, r_7\}$, since both r_6 (explains 45.73%) and r_7 (explains 54.73%) contributed to the difference between norm value and the actual value, and are both needed to explain the desired fraction of inf(C^+, r_2). In words, the total operating results for the ABC-company are relatively high, because of the fact that the total operating rev-

enues (r_6) are high and the total operating costs (r_7) are low in comparison with their branch averages. Obviously, $C_p^- = \emptyset$. Thus, the variable r_2 has two children in the explanation tree. Both children correspond to equations (eq. 6 and 9) in the business model and can therefore be explained further.

Table 3: Actual and norm values for $r_2 = r_6 - r_7$.

	actual	norm	inf(x_i, y)	diff. %
r_2	60.42	14.79		308.52
r_6	329.50	308.64	20.86	6.76
r_7	269.09	293.84	24.76	-8.42

Analogous to the previous example, the new events to be explained are specified as: ⟨ABC-company(2001), $\partial r_6 = high$, branch_average (2001)⟩ and ⟨ABC-company(2001), $\partial r_7 = low$, branch_average(2001)⟩. In other words, we want to determine which lower level revenues and costs variables in the business model contributed significantly to these events. For these equations the influence values are omitted because of space limitations. The previous examples of different one-level explanations are now combined to a complete tree of causes. Fig. 1 summarizes the results of the complete diagnostic process, where dashed lines indicate counteracting causes. Since there is only one symptom to be explained, the diagnosis contains one maximal explanation. Thus, Fig. 1 actually depicts the maximal explanation, as specified in section 2.4, for $\partial r_1 =$ "*high*".

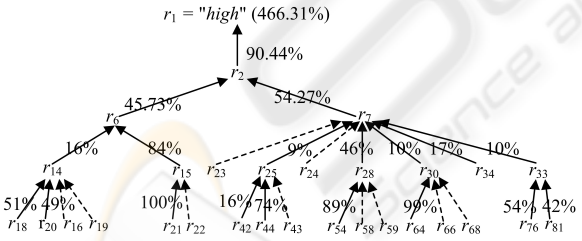

Figure 1: Diagnosis for $S = \{\partial r_1 = high\}$ at ABC-company.

3.2.2 Explanation Generation with Multi-step Look-ahead

In this section, we explain the initial event for the ABC-company with the look-ahead method. The method in the diagnostic program is configured for *one-step look-ahead*. For the threshold value we take again $T^+ = T^- = 0.85$. As before, explanation generation starts again with the root equation. From the data in Table 2 the same set of causes as with maximal explanation is identified. However, instead of proceeding with purely explanation of the parsimonious contributing causes the methods looks for potential cancelling-out effects, one step ahead in the business model. The look-ahead procedure takes into account the effects of all variables one level deep, i.e. the effects of the RHS-variables in equations 2, 3, 4 and 5.

Fig. 2 shows the one step look-ahead with arrows "stepping over" the intermediate nodes, and pointing at the RHS variables of equation 2, 3, 4 and 5, in the partial explanation tree. In this figure, the straight black lines indicate the parsimonious causes that were detected with maximal explanation.

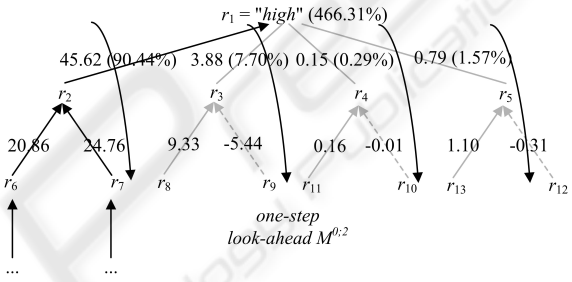

Figure 2: Illustration of one-step look-ahead.

In the analysis phase of the procedure the function substitution is applied to find parsimonious contributing and counteracting causes, which were missed in the local explanation of differences by standard multi-level explanation. Equations 2 through 5 are substituted into the root equation and the following new 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 obtained by substitution is added to the set of business model equations, *changing* the original business model. Because the substituted function is again additive, the conjunctiveness constraint is satisfied. Notice that the specification of the event to explain ∂r_1 remains the same, but now equation $M^{0;2}$ is applied to explain the difference. Table 4 summarizes the results of our extended model of ABC-company's relatively high results before taxation. It follows that $C_p^+ = \{r_6, r_7, r_8\}$ and $C_p^- = \{r_9\}$. We conclude that the effects of causes r_8 and r_9 are significant at the specified fractions for parsimonious sets.

Notice that these *hidden causes* were missing in analysis with maximal explanation. For the tree of causes this means that two new children are added to the root node: a contributing child for r_8 and a counteracting child for r_9. As a result the top branches of

Table 4: Actual and norm values for $M^{0;2}$: $r_1 = (r_6 - r_7) + (r_8 - r_9) + (r_{10} - r_{11}) + (r_{12} - r_{13})$.

	actual	norm	inf(x_i, y)	diff. %
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

the original tree are updated, as can be seen in Fig. 3. Notice that the variable r_3 is not part of the tree of causes (grey line).

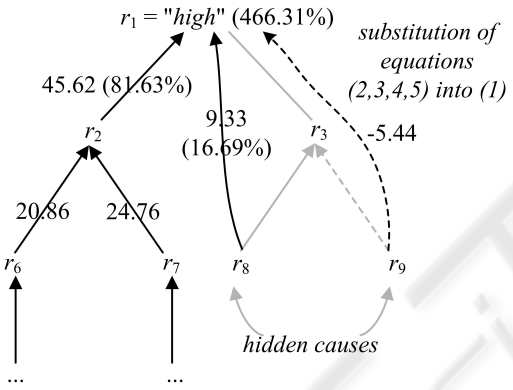

Figure 3: Detection of hidden causes with $M^{0;2}$ of $S = \{\partial r_1 = high\}$.

4 IMPLEMENTATION

In this section we shortly present the software implementation of the prototype diagnosis application in MS Excel in combination with Visual Basic. This application is initially programmed to perform the experiments and analyses for the case study at Statistics Netherlands. However the prototype software can handle data and business models from multiple application domains. Most elements of the program are discussed in the previous parts of this paper. However the procedure *diagnostic component* was not discussed earlier. It contains the method for maximal explanation as well as the multi-step look-ahead algorithm. For the implementation of the procedure we applied *tree programming* to generate the tree of causes.

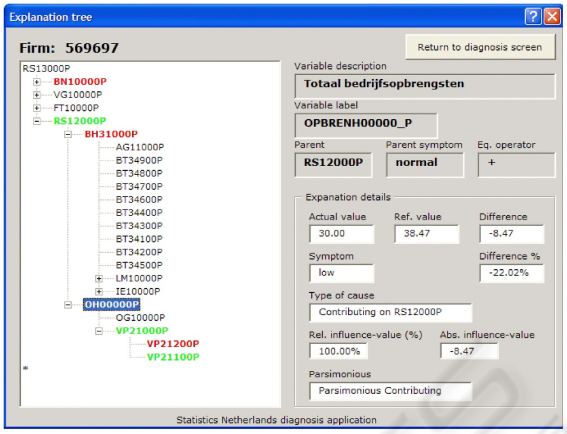

Figure 4: Tree viewer in diagnosis application.

The *tree-viewer* interface of the program is depicted in Fig. 4. In the viewer the whole explanatory graph can be made visible my 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

In this paper, we extended the method for automated business diagnosis in (Daniels and Feelders, 2001; Feelders, 1993) and developed a new implementation. 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. Furthermore, the multi-level look-ahead algorithm is proposed to enhance the explanation methodology so that it can deal with cancelling-out effects, i.e. the common effect that variables cancel each other out somewhere in the business model with the result that their effect on a higher level in the business model is partially or fully neutralized. The extended model is implemented in VB. Within the software implementation 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 diag-

nosis of company performance as a compact tree.

The applicability of the method is illustrated in a case study on interfirm/historic comparison 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 multi-level look-ahead procedure makes significant causes visible that would be missed by the explanation methodology of maximal explanation. In addition, the fully automated diagnostic process makes it possible to detect and explain abnormal company behaviour in large data sets. We believe that this enhanced framework could assist analysts and improve the decision-making process, by automatically generating explanations for exceptional values in various data sets and business models.

REFERENCES

Courtney, J. F., Paradice, D. B., and Mohammed, N. H. A. (1987). A knowledge-based dss for managerial problem diagnosis. *Decision Sciences*, 18(3):373–399.

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

Feelders, A. J. (1993). *Diagnostic reasoning and explanation in financial models of the firm*. PhD thesis, Tilburg University, Department of Economics.

Hesslow, G. (1983). Explaining differences and weighting causes. *Theoria*, 49:87–111.

Humphreys, P. W. (1989). *The chances of explanation*. Princeton University Press, New Jersey.

Kosy, D. W. and Wise, B. P. (1984). Self-explanatory financial planning models. In *Proc. of AAAI-84*, pages 176–181, Austin, TX.

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

Statistics Netherlands (2006). Centraal bureau voor de statistiek.

Verkooijen, W. J. (1993). Automated financial diagnosis: a comparison with other diagnostic domains. *J. Inf. Sci.*, 19(2):125–135.

APPENDIX

Because of space limitations only the description of the variables identified by the explanation model are described here in detail.

Result variables:

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

$\vdots$

r_{22}: net sales other activities

Cost variables:

r_{23}: cost of goods sold

r_{24}: total costs of labour

$\vdots$

r_{34}: depreciations on tangible and intangible fixed assets

Algorithm: Multi-level Look-ahead

1: y is the root node of the tree
2: **for** $p = 0$ to $N-1$ **do**
3: determine parsimonious causes for equation(s) $M^{p;p+1}$
4: add parsimonious causes to the tree as successor nodes
5: **if** look-ahead is activated **then**
6: **for** $i = 1$ to $N-1$ **do**
7: substitute jointly all equations on $M^{p+i;p+i+1}$ into equation $M^{p;p+i}$
8: add derived equation $M^{p;p+i+1}$ to M
9: determine parsimonious causes for $M^{p;p+i+1}$
10: **if** causes on level $p+i+1$ are parsimonious **then**
11: add new parsimonious causes as successor nodes to the tree
12: remove non-parsimonious causes from the tree
13: **if** a node corresponds to counteracting cause **then**
14: it has no successors
15: **if** a node corresponds to variable that cannot be explained in M **then**
16: it has no successors