3-4-2015

Inductive Development of Reference Process Models Based on Factor Analysis

Alexander Martens

Peter Fettke

Peter Loos

Recommended Citation

Martens, Alexander; Fettke, Peter; and Loos, Peter, "Inductive Development of Reference Process Models Based on Factor Analysis" (2015). *Wirtschaftsinformatik Proceedings 2015*. Paper 30.
http://aisel.aisnet.org/wi2015/30

Inductive Development of Reference Process Models Based on Factor Analysis

Alexander Martens[1], Peter Fettke[1] and Peter Loos[1]

[1] Institute for Information Systems at the German Research Center for Artificial Intelligence (DFKI GmbH) and Saarland University, Saarbruecken, Germany
{alexander.martens,peter.fettke,peter.loos}@iwi.dfki.de

Abstract. Business Process Modeling has become a major task in the field of Business Process Management. There has been a trend within organizations towards adapting and reusing existing reference models for the modeling of individual business processes that belong to the same application domain to save time and costs. Due to the high availability of individual business process models together with the need for further reference models, the inductive development of reference models has gained tremendous importance. Meanwhile, different heuristic approaches have been proposed to the research community, based on median graph computation. By way of contrast, our approach tries to derive reference models based on factor analysis, which does not require an approximated matching between the individual business processes as an additional level of abstraction. The results of the described approach are compared to the results of an existing heuristic approach for a qualitative evaluation of both inductive development concepts.

Keywords: Information Systems, Business Process Management, Business Process Modeling, Inductive Reference Model Development, Factor Analysis

1 Introduction

Nowadays, the inductive development of reference process models has become an important and lively topic of discussion in the business process modeling community. The general problem, justifying the actual need for reference model development, is the lack of reference models in certain application domains as well as the fact that reference models are of great importance for business process modeling as part of the organizational practice. For example, the reuse of generally applicable reference models is accelerating the modeling task itself, resulting in time and cost reductions [1]. At the same time, more and more individual business processes are available, due to the growing maturity of concepts, methods and techniques in the field of business process management (BPM). In parallel, the increasing number of different information systems, e.g. PDM, ERP, SCM, WFM, CRM systems [2], which enable the support of individual business processes in organizations [3], promotes the further ongoing individualization of existing business processes and the emergence of new ones. That is why the inductive development of universally applicable and reusable

12th International Conference on Wirtschaftsinformatik,
March 4-6 2015, Osnabrück, Germany

Martens, A.; Fettke, P.; Loos, P. (2015): Inductive Development of Reference Process Models Based on Factor Analysis, in: Thomas. O.; Teuteberg, F. (Hrsg.): Proceedings der 12. Internationalen Tagung Wirtschaftsinformatik (WI 2015), Osnabrück, S. 438-452

reference models [4] plays a key role in increasing their availability, notwithstanding the fact that inductive development is still at the discussion and research stage. The inductive modeling strategy, in contrast to deductive reference model development, is not based on the consideration of general theories and concepts [5], but it consolidates real-world business process data, e.g. in the form of individual business process models. The aim of the inductive strategy is to identify similarities between the individual business process models in order to derive an abstracting reference model [6]. In the related discipline of process mining, business processes are derived from records of business executions instead.

In literature, common shortcomings of more or less heuristic approaches in the research area of inductive reference model development can be summarized as follows. First, existing approaches are limited by heuristics with negative impacts on the stability and the quality of the approximated results. Second, most of the methods require further prerequisites and assumptions, depending on the characteristics of the used modeling language. Third, similarities are identified based on syntactical comparison of labeled business process elements without considering the semantics. Fourth, the underlying process matching problem of today's methods has to be solved, being NP-complete. Heuristic approaches are used in order to approximate a solution for this problem so as to deal with the complexity of today's real-world business processes in an efficient and effective way. The approximation of a matching which is not unique in general [7] introduces an additional level of abstraction, leading to a loss of information. In the related work stream of process mining, similar concepts, methods and techniques are applied, and similar problems exist, compared to the work stream of inductive development.

Motivated by the unused, but important and promising potential of the inductive strategy, the main contribution of this paper is the presentation of an innovative conceptual idea for the purpose of inductive reference model development based on factor analysis. The goal is to address the mentioned gaps of existing methods by the described approach in the following work that can be classified as design science research [8]. In both the work stream of inductive reference model development and the work stream of process mining, the application of statistical data analysis techniques is fairly unexplored. So far, those techniques are only applied for dimension reduction or clustering purposes common in the area of process mining. The presented method can also be adopted to support the detection of previously unknown process structures, for example in event logs, seen as a possible contribution to this field. Given a set of individual real-world business process examples as input for the development of an inductive reference model, our approach is to perform statistical data analysis in order to find an abstracting reference model that abstracts from differences, while focusing on commonalities. Factor analysis fits to this problem at hand because for this purpose it can be applied directly to a metric-scaled representation of the input in order to reduce the individual business process models to a lower number of latent variables, called factors. The key aspect is that factor analysis allows to draw statistical inferences from the factors back to the input about the degree of its reflection in the result. The abstracting reference model is obtained from those characterizing inferences. A fully-integrated cluster analysis ensures the robustness of the method, because the input is transformed into a homogeneous selection of individual business process models.

The work described in this paper is structured as follows: In section 2, related work is summarized and compared to our approach. In section 3, the mathematical background regarding statistical data analysis is explained in more detail. While section 4 describes the algorithmic idea on a high level, section 5 focuses on the single algorithmic steps in order to provide a full and understandable picture of the implemented algorithm. In section 6, the results are presented and compared with the results of a heuristic method based on the concept of median graph computation. Finally, section 7 concludes the paper.

2 Related Work

In literature, the work stream of inductive reference model development is strongly related to the stream of process mining, sharing common concepts, methods and techniques. For example, heuristic approaches have evolved in both areas. However, the intended aim has to be understood differently. The inductive strategy automatizes the generation of reference models based on a set of individual business process models as input. Instead, process mining extracts valuable knowledge from records of business executions [9], not necessarily in order to derive business process flows. Depending on the mined perspective, which can be differentiated between the process, organizational and data perspective, for example, business processes can be modeled as components in an object-oriented sense that are connected to each other to express dependencies and requirements between them.

In the meantime, different heuristic methods have been proposed to the research community in the area of inductive reference model development. In this context, the powerful concept of genetic algorithms is considered for modeling and optimization purposes more and more often, because it is easy to extend and generally applicable without the need for making assumptions about the optimization problem itself. Yahya et al. [10] have used as the first genetic algorithms in order to reduce the inductive derivation of reference models to the optimization problem of minimal graph-edit distance. However, the presented results do not prove the practical applicability of their work because the presented results are not generated based on real-world business process data, being modeled in no wide-spread modeling language. The application to real-world examples and the evaluation of the results is addressed by the work of Martens et al. [11]. Instead of using genetic algorithms, Ardalani et al. [12] demonstrate another heuristic approach, defining a minimal cost of change function in order to operationalize a reference model also based on minimal graph-edit distance. The reference model is developed iteratively by evaluating this function based on individual business process models.

In the related area of process mining, similar concepts, methods and techniques are applied within various organizations such as public institutions [9], telecom companies [13] and healthcare institutions [2]. Van der Aalst et al. [14] have introduced the α-algorithm in order to discover process models, modeled as Petri net from event logs. Li et al. [15] present a heuristic search algorithm for process variant mining. The formulated optimization problem is approximated using a heuristic measure. The application of clustering techniques is mentioned in some literature to facilitate process mining especially in healthcare [2], [16]. In this context, statistical data analysis is mainly used for the purpose of dimensionality reduction in order to process high-dimensional data [17].

Compared to our approach, heuristic methods are less stable and accurate. For example, the α-algorithm is not showing any robustness against noise in the data. The results are generated based on preconditions and assumptions that are necessary to solve the underlying matching problem, being independent from the formulated optimization problem. The matching is approximated based on heuristics, resulting in an equivalent relation that represents a necessary, but additional level of abstraction. This leads to a loss of information. So far, process element matches are established by syntactic and type-oriented similarity measures. In contrast, our approach extends this bunch of similarity measures by taking into account semantics.

3 Mathematical Background

In this paper, statistical data analysis is applied in order to abstract from differences while gently focusing on commonalities in the business process structures, but not on the basis of assumptions or on the basis of approximations. The intended goal is to obtain more stable and accurate results compared to the use of heuristic methods. In the mathematical sense, a metric-scaled representation of the data is required to decompose the observed variables into a smaller number of latent variables, which reflect the main evidences, while reducing the variability between observations, e.g. business processes. The latent variables guarantee to be independent, when the normally distributed data jointly span a low-dimensional mathematical space. Depending on the application, there are various techniques to search for joint variations in response to latent variables.

Principal component analysis (PCA) is usually applied for the purpose of dimension reduction in conjunction with cluster detection for example. Therefore, the data is projected from a high-dimensional into a low-dimensional mathematical space. The appropriate transformation matrix results from the relevant eigenvectors. The intended goals of factor analysis (FA) are structure detection and causal modeling [18], in that the analysis characterizes the relationship between observed, correlated variables and a lower number of latent, uncorrelated variables. In mathematical terms, the difference between PCA and FA is that FA accounts only for the variance that is common among the observed variables instead of accounting for all of the variance in the data.

As a first step in FA, the data matrix is Z-transformed by mean-centering and standardization. Afterwards, it is decomposed in its eigenvectors and corresponding eigenvalues. Together with the eigenvectors, the factors (or latent variables) form a linear combination, representing the observed variables, and the eigenvalues, accounting for the relevance of the direction of the eigenvector, lead to the factor loads. The factor loads are obtained by scaling the eigenvectors with the square root of the corresponding eigenvalues. In general, different methods are possible for the decomposition, e.g. the standard covariance method or singular value decomposition (SVD). Each method makes different demands on the data, offering both advantages and disadvantages. The main specific characteristic of the application described in this paper is that there are more variables than observations in general. In addition, the number of variables is very high. Under these conditions, the covariance method is not numerically stable and it is not efficient. That is why we are making use of the SVD as an alternative technique that is especially suitable for high-dimensional data. It is numerically more stable and it provides efficient results, even if the number of variables is larger than the number of observations [19]. However, the high random-access memory consumption of our present implementation is seen as a limitation.

4 Overall Approach

Real-world business process models are considered in the following as directed graph structures, consisting of a set of edges that induces a predefined order among a set of nodes. Each node is labeled by a distinct linguistic expression of a natural language. In addition, it may be the case that more meta-information is contained within single entities (e.g. costs for a specific transition), which has been ignored by our work so far. This kind of representation is independent from a concrete modeling language. The following main steps describe our inductive development approach on a high level, as illustrated in figure 1.

In the first step, each directed graph structure is transferred into a matrix representation, which follows the idea behind an adjacency matrix. The dimensions of the matrix represent the set of syntactically distinct labels among all existing nodes that appear in the data. Each matrix entry is assigned a metric value that quantifies the degree of uniqueness regarding the existence of an edge between the respective labels. Therefore it is necessary to compare the source and target labels to the labels of all the source and target nodes, connected by a directed edge in the respective individual graph structure. The idea is that the higher the average of calculated syntactic and semantic similarity values, the higher the chance that an edge occurs. Apart from the calculation of syntactic similarity, the measurability of semantic similarity is a challenge, because labels can be syntactically different and yet still express a similar or the same meaning. For this reason, the approach integrates linguistic background knowledge in the form of dictionaries for the German and English language. In the second step, all single matrices are integrated into a common data matrix by linearization of each matrix as one vector of observed variables. In the third step, the mean-centered and standardized data matrix is decomposed into its eigenvectors and corresponding eigenvalues by the application of SVD (cf. section 3) as an intermediate result in the context of statistical data analysis. The normed eigenvectors form a transformation matrix that is used to project the individual business process models from a high-dimensional into a low-dimensional mathematical space in order to perform a cluster analysis afterwards. The objective is to ensure a homogeneous selection of individual business process models by the removal of outliers to increase the robustness of the overall approach. In terms of business process modeling, an outlier can be understood as a business process model that does not belong to the same application domain as the majority. Based on the results of cluster analysis, a reduced number of factors is expected that account for the variances among the selected individual business process models. The relevant number of factors is derived from the corresponding eigenvalues based on the application of different decision criteria. After calculating the corresponding factor loads, characterizing the relationship between each factor and each observed variable, the commonality vector is calculated. It expresses the degree of how much each observed variable is explained by all the factors. At this point, the legitimate idea of deriving the abstracting reference model directly from the intermediate results of factor analysis does not lead to useful results. The transformation of both the factor loads and eigenvectors back into the original matrix representation results in highly fragmented graph structures, not representing a tangible business process model. Accompanied by the missing mathematical interpretation, the individual business process model that has the smallest distance to the commonality vector is selected in the fourth step. It is interpreted as a model that is closest to the reference model. In the fifth step, it is refined based on the commonality vector to inductively develop an abstracting reference model that can be reused directly by business process designers for the task of business process modeling.

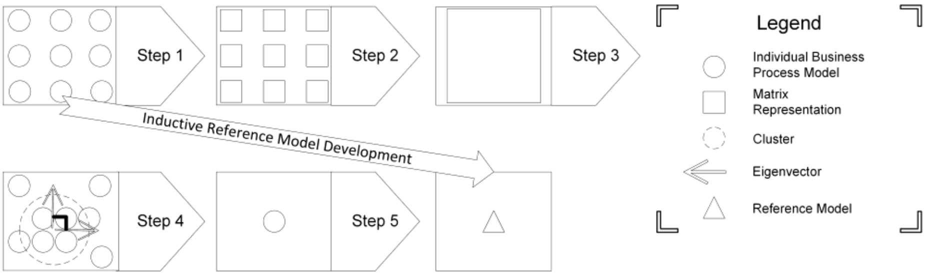

Fig. 1. Overview of the overall approach

5 Algorithm

5.1 Step 1: Represent each Input Model by a Metric-Scaled Matrix

In the first step, the set of distinct labels among all nodes of the individual business process models has to be determined and sorted in an alphabetically ascending order.

$$L = \{l_1, \ldots, l_n\} \tag{1}$$

The distinct labels represent the rows and columns of the $n \times n$ matrix A, representing an individual business process model. The metric-scaled values of the matrix define the uniqueness regarding the existence of a directed edge between a labeled source and target node in the corresponding model $i \in I, c = |I|$ that is described.

Therefore, the corresponding labels are compared to the labels of all the source and target nodes that are connected by a directed edge within the respective graph structure.

$$A_i^{n \times n} = \begin{bmatrix} p_i(l_1, l_1) & \dots & p_i(l_1, l_n) \\ \dots & p_i(l_p, l_q) & \dots \\ p_i(l_n, l_1) & \dots & p_i(l_n, l_n) \end{bmatrix}; p, q \leq n = |L| \tag{2}$$

$$p_i(l_p, l_q) = \sum_{\forall(s,t) \in E_i} \frac{2 * p_i^{source}(l_p, l_s) * p_i^{t \arg et}(l_q, l_t)}{p_i^{source}(l_p, l_s) + p_i^{t \arg et}(l_q, l_t)} \tag{3}$$

In general, the methods used for the determination of similarities between business process elements range from syntactic, semantic, type-oriented and attribute-oriented methods [20] to contextual similarity measures [21]. The type-oriented similarity between two labels is taken under consideration by the following function that compares the types of the corresponding nodes.

$$sim_{type}(x, y) = \begin{cases} 1 & type(x) = type(y) \\ 0 & otherwise \end{cases} \tag{4}$$

The syntactic similarity measure calculates a string edit distance based on the Levenshtein-edit-distance *lev* [22], which is normalized to a similarity in the range [0, 1].

$$sim_{syn}(x, y) = 1 - \frac{lev(x, y)}{\max(x, y)} \tag{5}$$

The semantic similarity measure makes use of a synonym comparison between the identifiers by integrating linguistic background knowledge in the form of open and language-dependent dictionaries, such as WordNet[1] and GermaNet[2] for the English or German language. For the calculation of semantic similarities, each label *l* is split into single words at different delimiters upfront. Afterwards, the stop words as well as punctuations are filtered out, resulting in a set of words w_l. The function $s(w_x, w_y)$ returns the set of all synonyms out of w_x, occurring in w_y.

$$sim_{sem}(x, y) = \frac{2 * |w_x \cap w_y| + |s(w_x, w_y)| + |s(w_y, w_x)|}{|w_x| + |w_y|} \tag{6}$$

In this context, based on the evaluation in [23], the combined similarity between two labeled nodes is measured as follows:

[1] http://wordnet.princeton.edu/

[2] http://www.sfs.uni-tuebingen.de/GermaNet/

$$sim(x,y)=\frac{sim_{syn}(x,y)+sim_{sem}(x,y)}{2};[0\leq sim(x,y)\leq 1];x,y\in L \qquad (7)$$

The presented similarity measures are combined with equal weighting according to [20]. The combined measure complements the calculation of matrix *A* as follows:

$$p_i^{source}(l_p,l_s)=sim(l_p,l_s)*sim_{type}(l_p,l_s) \qquad (8)$$

$$p_i^{t\arg et}(l_q,l_t)=sim(l_q,l_t)*sim_{type}(l_q,l_t) \qquad (9)$$

5.2 Step 2: Transform each Matrix Representation into a Vector

All matrices are transformed into a linearized vector for the purpose of integrating all business process model representations into a common data matrix.

$$A_i^{n\times n}=\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \dots & a_{pq} & \dots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \Rightarrow \vec{x}=\langle a_{11} \quad \dots \quad a_{1n} \quad \dots \quad a_{n1} \quad \dots \quad a_{nn}\rangle \qquad (10)$$

5.3 Step 3: Perform Factor Analysis and Cluster Analysis

The data matrix X^{cxn^2} comprises a set of *c* observations of *n²* variables. The data is arranged as a set of *c* data vectors $\vec{x}_1,\dots,\vec{x}_c$ with each vector representing an individual business process model, described by *n²* variables. Factor analysis (cf. section 3) is used in order to calculate the factor loads for a reduced number of factors. As an intermediate result of factor analysis, the eigenvectors and the corresponding eigenvalues of the decomposed data matrix *X* are obtained. The relevant eigenvectors form a transformation matrix that is used to project the individual business process models based on their linearized vector representation into a low-dimensional mathematical space. The individual models that are located in close proximity form a cluster. Under the assumption that the input data to this approach belongs more or less to the same application domain, the outliers become visible. The corresponding linearized vector representations are removed from the data matrix *X*. Afterwards, factor analysis is restarted without performing cluster analysis again. The resulting factors explain most of the variability within the data set. The most common decision criteria for estimating the relevance of eigenvectors and factors are described below.

- Kaiser criterion: All factors with a corresponding eigenvalue greater or equal to the value of 1 are selected as relevant.
- 90% rule: All factors that explain 90% of the overall variances in terms of the sum total of all eigenvalues are considered to be relevant.

- Scree-plot test: Plotting all eigenvalues leads to the number of relevant factors at the point where the gradient of the plotted curve decreases significantly.

Due to the fact that the mentioned criteria tend to overestimate or underestimate the relevance, the average is considered finally as the relevant number *k*. The influence of a slight decrease or increase of the parameter *k* on the final result is negligible. Based on the factor load matrix $B^{n^2 \times k}$, the commonality vector $\vec{c}$ is derived as follows:

$$c_i = \sum_{j=1}^{k} b_{ij}^2 \tag{11}$$

Each commonality value c_i expresses the degree of how much the observed variable *i* is explained by all the factors. In other words, the commonality describes the relationship between observed variables and their contribution to the explanation of relevant directions with regards to content.

5.4 Step 4: Select an Individual Business Process Model

The inductive reference model has to be derived based on the previous analysis. Therefore, the commonality vector is understood as query vector to determine the individual business process model that fits closest to the reference model in terms of the strongest explanation by all the factors. This is achieved by calculating the Euclidean distance between the linearized vector representation of each input model and the commonality vector.

5.5 Step 5: Refine the Selection Towards an Abstracting Reference Model

In the final step, the refinement process transforms the selected individual business process model into the reference model based on the strength of the dependency between two nodes in response to the identified factors. A strong dependency is marked by commonality value closer to 1, while a weak dependency is expressed by a value closer to 0. As a first evidence, an abstracting reference model is obtained by removing all edges that express the dependency between nodes, being assigned a commonality value below 0.5. Increasing this threshold is a useful method to strengthen the abstraction capability of the method in order to find further reference models on different levels of abstraction. All disconnected nodes are removed post-processing, resulting in a connected graph component, considered as inductive reference model.

6 Evaluation

6.1 Proof of Concept

The described algorithm is prototypically implemented in the Java programming language and integrated in a self-developed application within our research group, providing the functionality of loading business process models that are represented in a common XML-based interchange format as for example EPC Markup Language (EPML) or ARIS Markup Language (AML). For the implementation of the algorithm itself, the JAMA library[3] is used for calculative purposes. The linguistic background knowledge is integrated in form of word nets using the official application programming interfaces. The results of factor analysis are exported for further analysis and verification purposes. Finally, the derived reference model is shown in a special viewer, providing further analysis and export functionalities. The evaluation was performed on today's standard hardware (Intel Core i7 vPro 2.3 GHz processor, 16 GB RAM) under Windows 7 64-bit version and Java Runtime Environment 7 64-bit version.

6.2 Scenarios

The evaluation of the described approach follows three different strategies based on two different real-world business process data sets. Scenarios 1 and 2 are based on the CoSeLoG project data set. Vogelaar et al. [24] retrieved 80 business process models (eight different business processes for 10 Dutch municipalities). The GBA1 business process is used as input for the two different evaluation scenarios. Scenario 3 is founded on a controlled modeling exercise, where different students attending the course "Introduction to Information Systems" in the summer semester 2011 at our institute had to design a business process model by hand, referring to the same process description in the context of the written examination of the course.

- Scenario 1 (S 1): The 10 individual variants of the GBA1 business process are integrated into one reference model. The generated reference model R is evaluated against a manually developed reference model R^* by hand by two independent designers, who were separated locally from each other during modeling without knowing the result of the described algorithm or discussing the modeling task upfront. Afterwards, the different modeled versions were reasonably merged into a single model based on discussions. For designing the reference model by hand, different practices are possible for creating a reference model. On one side, only the matching nodes and edges can be considered in order to build up a reference model. On the other side, the incorporation of all nodes into some kind of super-model is also a possibility. A compromise would be the selection of nodes that occur with a given probability. The designers followed this path, but in addition, nodes that seemed to be of high importance for the overall process were also integrated. Finally, the result was reviewed and classified as a plausible representative.
- Scenario 2 (S 2): In the first step, the first variant R^* describing the GBA1 business process of the Dutch municipality A is used to generate 20 variants by renaming, deleting and exchanging the order of model elements based on a certain probability so as to obtain a diverse synthetic data set of business process models. In the second step, the resulting models are provided as input to the algorithm, inductively

[3] http://math.nist.gov/javanumerics/jama

deriving a reference model. It is expected that the generated reference model R has to be similar to the GBA1 variant *R**, serving as a pattern for automatic variant generation.

- Scenario 3 (S 3): A close look at the 38 solutions that were developed by the students based on the controlled modeling exercise shows that
  - the solutions do not meet partially the essential subject matters,
  - the exercise is partially misunderstood because of misleading formulations and
  - the syntactic rules of the given modeling language are not respected.

  These points are general risks that have to be faced in controlled modeling exercises. An example solution on the subject area was developed independently by experts by following the same textual specification. The modeling task is simplified by the fact that a reference is already given implicitly by such a description. In this context, we have identified the main subject matters and process flows as assessment framework in advance on the basis of the controlled modeling exercise. Afterwards, exam solutions – which either did not represent a valid model or met the established criteria to a very small extent – were dropped as input for our method. The preceding cluster analysis was applied to the remaining ten solutions. Finally, based on the six remaining individual solutions of the students, a reference model *R* is derived inductively. It is compared to the example solution *R**.

In table 1, the differences among the scenario-specific characteristics are summarized.

Table 1. Summary of scenario-specific characteristics

	Avg. # of nodes	Avg. # of edges	Execution time	Memory consumption	Modeling type
Scenario 1	31	33	5 sec.	2.5 GB	natural
Scenario 2	19	20	8 sec.	2.6 GB	synthetic
Scenario 3	43	46	81 sec.	10.1 GB	controlled

6.3 Measures

In this paper, the main question regarding the evaluation is how relevant an obtained reference model is compared to the expected one. This is the reason why it makes sense to transfer the measures precision and recall as noted in equation 12 and equation 13 from the research area of information retrieval to the area of inductive reference model development. In equation 14, the F-measure is defined and interpreted as weighted average (harmonic mean) of precision and recall, rating the accuracy of the retrieved reference model. The values of these measures lie in the interval [0, 1]. A higher value indicates a more relevant and accurate result.

$$precision = \frac{|R^*| \cap |R|}{|R|} \tag{12}$$

$$recall = \frac{|R^*| \cap |R|}{|R^*|} \tag{13}$$

$$F - measure = \frac{2 * precision * recall}{precision + recall} \tag{14}$$

6.4 Results

In general, it is difficult to provide a formal correctness proof for a method. That is the reason why this paper focuses on the evaluation of the presented approach in terms of practical usage and relevance of the results. The results are independent of the use of a certain modeling language as long as the individual business processes are representable as directed graph structures. The generation is possible in real-time and the results are reproducible. Furthermore, statistical data analysis reduces necessary preconditions and assumptions. This makes the method highly relevant for practical usage. The relevance of the results is evaluated based on different measures as defined in subsection 6.3. The evaluation measures are calculated and the results are presented in table 2. A commonality value greater than 0.9 has to be chosen in order to obtain the following stable and promising results for the three different scenarios.

Table 2. Comparison of the FA based approach with a heuristic median graph-based method

			Event	Function	Connector	Edges	All
S 1	Factor Analysis	Precision	0.87	0.91	0.67	0.83	0.82
		Recall	0.67	0.82	0.67	0.65	0.69
		F-measure	0.76	0.86	0.67	0.73	0.75
	Median Graph	Precision	0.89	0.88	0.50	0.86	0.81
		Recall	0.62	0.58	0.33	0.59	0.55
		F-measure	0.73	0.70	0.40	0.69	0.65
S 2	Factor Analysis	Precision	0.77	0.86	0.83	0.64	0.73
		Recall	0.85	0.83	0.83	0.54	0.69
		F-measure	0.80	0.84	0.83	0.59	0.71
	Median Graph	Precision	0.82	0.90	0.50	0.86	0.80
		Recall	0.75	0.82	0.33	0.75	0.69
		F-measure	0.78	0.86	0.40	0.80	0.74
S 3	Factor Analysis	Precision	1.00	1.00	1.00	0.81	0.91
		Recall	0.91	0.87	1.00	0.76	0.84
		F-measure	0.95	0.93	1.00	0.78	0.87

The results of scenario 3 compared to scenarios 1 and 2 state the successful integration of semantics. The results of another actual inductive approach [11] that represent the concept of median graph computation are opposed in table 2 for the scenarios 1 and 2. A direct comparison of both evaluations allows the conclusion that the stability of the described approach is higher, while its results are more accurate.

7 Conclusion

The presented method for the inductive development of reference models based on statistical factor analysis leads to promising results in the different evaluation scenarios using real-world data. The results have been compared to a heuristic approach based on median graph computation [11], dealing with the same research problem. Our approach is able to increase precision and recall by avoiding an approximated heuristic solution of the underlying business process matching problem that is complex and not unique. Furthermore, statistical data analysis avoids necessary preconditions and assumptions. The strong determination ability of this method speeds up the runtime, resulting in higher effectiveness and efficiency. Due to the used methodology of statistical data analysis, the results show a greater stability. In scenario 2, the results are slightly worse in comparison to those of the heuristic median graph-based approach. The problem seems to be that the refinement process is not as flexible as genetic operations. In scenario 3, the incorporation of semantics leads to convincing results, showing its potential also for the integration in other existing inductive development techniques. Rehse et al. [25] have proposed different requirements for inductive development methods that are satisfied by our approach. Overall, this work represents a promising and innovative conceptual idea as an alternative to concepts that are behind existing inductive reference model development approaches. In the current state of research, it provides important advantages in comparison to other concepts. The evaluation shows high-quality results, but a more detailed and qualitative evaluation of the described approach is necessary to prove its relevance for the organizational practice. Therefore, the evaluation must be extended so that further alternative inductive development approaches are included in more complex scenarios. Furthermore, the applied combined similarity measure has to be questioned more closely, opening up possibilities for extension or replacement. While deductive developed reference models are describing more or less a best practice, inductively developed reference models are interpreted more as a common practice. The practical problem behind the inductive development strategy is that background knowledge regarding the impact of single process steps is completely missing. Providing meta-information, e.g. the execution cost of certain process activities, is necessary in order to be able to measure the impact. However, for proven individual business processes as input, the result of inductive development approaches qualifies as a valid reference model.

Our approach suffers from certain limitations such as the high random-access memory consumption and the numerical instability of singular value decomposition that is caused by the high-dimensional data matrix representation (cf. section 3). This is the reason why the way of constructing the data matrix has to be opposed to other possibilities with the aim of reducing its dimensionality. One possibility is to evaluate a fragment-based approach instead of considering the entire set of nodes and edges, representing a graph structure, as basis for statistical data analysis. Another possibility is to unite certain matrix dimensions which represent a common synonym group. The refinement process has to be more flexible to cope with degenerate business process models. At the moment, the missing integration of further operations, such as the adding, the mutation or recombination of structural elements is also a limitation. Furthermore, our approach does not allow for the configuration of a company-specific variant of a reference model. For the future, it will be interesting to consider company-specific constraints in addition to the result of factor analysis during the selection and refinement step.

References

1. Becker, J., Meise, V.: Strategy and Organizational Frame. In: Becker, J., Kugeler, M., Rosemann, M. (eds.) Process Management. A Guide for the Design of Business Processes, pp. 91-132. Springer, Berlin (2011)
2. Mans, R. S., Schonenberg, M. H., Song, M., van der Aalst, W. M. P., Bakker, P. J. M.: Process mining in healthcare – a case study. In: Azevedo, L., Londral, A. R. (eds.) Proceedings of the first international conference on health informatics, pp. 118-125. Funchal, Madeira, Portugal (2008)
3. Houy, C., Fettke, P., Loos, P., van der Aalst, W. M. P., Krogstie, J.: Business Process Management in the Large. Business & Information Systems Engineering (BISE) 3 (6), 385-388 (2011)
4. Fettke, P., Loos, P.: Perspectives on Reference Modeling. In: Fettke, P., Loos, P. (eds.) Reference Modeling for Business Systems Analysis, pp. 1-20. Idea Group, Hershey, PA (2007)
5. Becker, J., Schütte, R.: A Reference Model for Retail Enterprises. In: Fettke, P., Loos, P. (eds.) Reference Modeling for Business Systems Analysis, pp. 182-205. Idea Group, Hershey, PA (2007)
6. Walter, J., Fettke, P., Loos, P.: How to Identify and Design Successful Business Process Models: An Inductive Method. In: Becker, J., Matzner, M. (eds.) Promoting Business Process Management Excellence in Russia - Proceedings and Report of the PropelleR 2012 Workshop, pp. 89-96. Moscow, Russia (2013)
7. Thaler, T., Hake, P., Fettke, P., Loos, P.: Evaluating the Evaluation of Process Matching Techniques. In: Kundisch, D., Suhl, L., Beckmann, L. (eds.) Tagungsband Multikonferenz Wirtschaftsinformatik 2014, MKWI-2014, pp. 1600-1612. Paderborn, Germany (2014)
8. Hevner, A. R., March, S. T., Park, J., Ram, S.: Design Science in Information Systems Research. MIS Quarterly 28 (1), 75-105 (2004)
9. van der Aalst, W. M. P., Reijers, H. A., Weijters, A. J. M. M., van Dongen, B. F., de Medeiros, A. K. A., Song, M., et al.: Business process mining: an industrial application. Information Systems 32 (5), 713-732 (2007)
10. Yahya, B.N., Bae, H., Be, J., Kim, D.: Generating Valid Reference Business Model using Genetic Algorithm. International Journal of Innovative Computing, Information and Control 8, 1463-1477 (2012)
11. Martens, A., Fettke, P., Loos, P.: A Genetic Algorithm for the Inductive Derivation of Reference Models Using Minimal Graph-Edit Distance Applied to Real-World Business Process Data. In: Kundisch, D., Suhl, L., Beckmann, L. (eds.) Tagungsband Multikonferenz Wirtschaftsinformatik 2014, MKWI-2014, pp. 1613-1626. Paderborn, Germany (2014)
12. Ardalani, P., Houy, C., Fettke, P., Loos, P.: Towards a minimal cost of change approach for inductive reference process model development. Proceedings of the 21st European Conference on Information Systems, ECIS-2013, Utrecht, Netherlands (2013)

13. Goedertier, S., Weerdt, J. D., Martens, D., Vanthienen, J., Baesens, B.: Process discovery in event logs: An application in the telecom industry. Applied Soft Computing 11 (2), 1697-1710 (2011)
14. van der Aalst, W. M. P., Weijters, A., Maruster, L.: Workflow Mining: Discovering Process Models from Event Logs. IEEE Transactions on Knowledge and Data Engineering 16 (9), 1128-1142 (2004)
15. Li, C., Reichert, M. U., Wombacher, A.: A Heuristic Approach for Discovering Reference Models by Mining Process Model Variants (2009)
16. Jagadeesh Chandra Bose, R. P., van der Aalst, W. M. P.: Context Aware Trace Clustering: Towards Improving Process Mining Results. In: Proceedings of the SIAM international conference on data mining, pp. 401-412. Sparks, Nevada, USA (2009)
17. Xu, X., Wang, X.: An adaptive network intrusion detection method based on PCA and support vector machines. In: Li, X., Wang, S., Dong, Z. Y. (eds.) Advanced data mining and applications, first international conference. LCNS, vol. 3584, pp. 696-703. Springer (2005)
18. Bartholomew, D. J., Steele, F., Galbraith, J., Moustaki, I.: Analysis of Multi-variate social Science Data. Statistics in the Social and Behavioral Sciences Series (2nd edition). Taylor & Francis (2008)
19. Speed, T.: Statistical Analysis of Gene Expression Microarray Data. Chapman & Hall/CRC (2003)
20. Dijkman, R., Dumas, M., van Dongen, B., Käärik, R., Mendling, J.: Similarity of business process models: Metrics and evaluation. Information Systems 36 (2), 498-516 (2011)
21. Ehrig, M., Koschmider, A., Oberweis, A.: Measuring similarity between semantic business process models. In: Proceedings of the fourth Asia-Pacific conference on Conceptual modelling – Volume 67, pp. 71-80. Australian Computer Society, Inc. Ballarat, Australia (2007)
22. Levenshtein, I. V.: Binary Codes capable of correcting deletions, insertions, and reversals. Cybernetics and Control Theory 10, 707-710 (1966)
23. Hake, P., Fettke, P., Loos, P.: Experimentelle Evaluation automatisierter Verfahren zur Bestimmung der Ähnlichkeit von Knoten in Geschäftsprozessmodellen. In: Kundisch, D., Suhl, L., Beckmann, L. (eds.) Tagungsband Multikonferenz Wirtschaftsinformatik 2014, MKWI-2014, pp. 1061-1074. Paderborn, Germany (2014)
24. Vogelaar, J. J. C. L., Verbeek, H. M. W., Luka, B., van der Aalst, W. M. P.: Comparing Business Processes to Determine the Feasibility of Configurable Models: A Case Study. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) Business Process Management Workshops. Lecture Notes in Business Information Processing, vol. 100, pp. 50-61. Springer, Berlin (2012)
25. Rehse, J.-R., Fettke, P., Loos, P.: Eine Untersuchung der Potentiale automatisierter Abstraktionsansätze für Geschäftsprozessmodelle im Hinblick auf die induktive Entwicklung von Referenzprozessmodellen. In: Alt, R., Franczyk, B. (eds.) Proceedings of the 11th International Conference on Wirtschaftsinformatik, WI-2013, pp. 1277-1291. Leipzig, Germany (2013)