Universität Stuttgart iF Institut für Industrielle Fertigung und Fabrikbetrieb

Fraunhofer IPA

Universität Stuttgart
Germany

Master thesis

Development of a model for the efficient optimization of customer satisfaction on administrative processes using Bayesian networks

by:

Francisco Javier REIG BLANCO

Matr.Nr.: 2763594

Supervisor:

Michael Kempf

Stuttgart, February 2013

-Statutory declaration-

I hereby declare on my honor that I have written this thesis / study work independently and used no other sources and aids.

Place, Date: _____ Signature: _____

Name, author

Contents

Figures

Tables

1. Starting Situation

1.1. Presentation of the problem

Traditionally, thinking about the quality of a product or service used to mean that such product or service had no defects and fulfilled all the functionality requirements. However, nowadays, quality is not only measured by these parameters but also many other facts related to what is understood as customer satisfaction. In other words, the state of the product or service is not the only important issue. It is also a crucial fact the level of satisfaction assigned by the customer to the product. In this thesis we focus our study on the latter, customer satisfaction, because of its increasing importance in terms of quality evaluation.

Then, the first step consists on knowing exactly what customer satisfaction is. Farris, Bendle, Pfeifer and Reibstein (2010) define it as: "the number of customers or percentage of total customers, whose reported experience with a firm, its products or its services (ratings) exceeds specified satisfaction goals". In other words, it consists on measuring how the products or services offered by a firm accomplish the customers' expectations or exceeds them.

The next step is then to know how this customer satisfaction level can be measured. There are several methods. The most popular is the one developed by Parasuraman, Zeithandl and Berry (1988). They provide the basis for the measurement of customer satisfaction with a service by using the gap between the customer's expectation of performance and their perceived experience of performance. In order to perform their study, they use their own methodology called SERVQUAL.

However, concerning this methodology, we can find some detractors that consider it excessively complex, subjective and statistically unstable. It is because of this that we propose an alternative methodology to deal with customer satisfaction. This is the use of Bayesian networks, which are easily interpretable for the managers. Besides, they give a more objective interpretation of the results as well as a better statistic stability.

1.2. Objective

The essential objective of this work is to optimize customer satisfaction in administrative processes. In order to do so, we look for a methodology robust enough to apply it to any administrative process. We use a particular scenario to illustrate this. More specifically, our example deals with customer satisfaction concerning the preparation, execution and accounting of business travels.

To achieve the above mentioned objective, we will use Bayesian networks (BN). It is said that this type of networks can be two-way applied. In other words they can be used, at the same time, to predict as well as to diagnose. We apply them to the analysis of customer satisfaction with data obtained from a survey of administrative processes. Besides, these networks present a very useful property: they allow new cases to be introduced in any moment and the net is automatically updated according to them. Therefore, these kind of nets are constantly learning from the data that they receive. As a consequence, the network improves with every new learning process getting closer to the real world they represent.

This tool (BN) should enable us to find the most important indicators for overall satisfaction in order to improve these indicators as far as possible. Moreover, BN must help us to determine which the strengths of the administrative processes are, and more importantly, which ones have a greater impact in the overall satisfaction.

2. State of the art

As described in the previous section, this project focuses on two different fields: Bayesian networks and customer satisfaction. Therefore, section 2 is divided in two subsections: one presenting Bayesian networks and another dealing with customer satisfaction. The former introduces the theoretical background on Bayesian networks in order to provide the reader with the necessary knowledge to understand why this methodology in the most appropriate for our study. It tries to address the following issues:

What are Bayesian networks?

An empirical example on Bayesian networks

Independency assumptions

Uses of Bayesian networks

The second subsection analyses the existent literature on customer satisfaction and its optimization by means of the cited methodology.

2.1. Bayesian networks

2.1.1. What are Bayesian networks?

Bayesian networks, also known as Belief networks, are a formalism to deal with uncertain knowledge. They allow both the representation of uncertain knowledge as well as the development of inference within it. Those are networks representing different scenarios in which the real state of the variables is unknown and there exist some causality between them. Thus, we need probabilities to provide an exhaustive description of the network. Therefore, a Bayesian network is a probabilistic model representing a set of variables and its dependencies. Formally, it is a directed acyclic graph (DAG) whose nodes represent random variables and its links contain information about their conditional dependencies.

The nodes can represent a measurable parameter, a latent variable (deduced from other observable variables) or a hypothesis. In any case, independency assumptions (see section 2.1.3) must be fulfilled. The variables value is not fixed; it depends on a probability quantifying the belief concerning the existence of a particular case in that variable. For instance, let a variable be the fuel tank of a car. It could present 3 states: *full*, *half* or *empty*. These states must be mutually exclusive and exhaustive and its probabilities need to sum up to one (100%). Therefore, it is impossible to have two states of the variable at the same time or not being in any of them. In other words, the variable will be at only one stage at a moment. In many studies variables can only be at two different states (binary variables). However, we use variables with more than two stages, as Bayesian networks can perfectly deal with them. More specifically, our variables present four different possible states corresponding to four degrees in a satisfaction scale. These variables can be either continuous or discrete but most of the algorithms only accept the discrete ones. Moreover, those using continuous variables discretize the variables using intervals. We now turn to the previous example again. A possible node could be to measure the level of the tank. The values of the variable can be full, half or empty with probability 0,1; 0,8 or 0,1[1] respectively, showing a strong belief of finding half the tank and not finding it full or empty. Those beliefs are represented in column vector as follows:

$$P(\text{fuel tank}) = \begin{pmatrix} P(\text{fuel tank} = \textit{full}) \\ P(\text{fuel tank} = \textit{half}) \\ P(\text{fuel tank} = \textit{empty}) \end{pmatrix} = \begin{pmatrix} 0{,}1 \\ 0{,}8 \\ 0{,}1 \end{pmatrix}$$

Apart from the nodes, we can also find links. These represent the relation between the nodes they connect. They allow the information to flow across the variables (or nodes), given their dependencies. This information flow is the so called *inference* and it is the key issue in this study. It is so because it shows the information on how the net changes when updating it with new data. The relationship between variables is quantified in a **c**onditional

[1] The number in brakets represents the probability of the tank being in any of the three possible states

probability table (CPT). As Bayesian nets represent causal relation between variables, the links go in a specific direction showing the cause-effect relation between such variables.

2.1.2. An empirical example of a Bayesian nets

In order to illustrate better what a Bayesian Network is, we now turn to present a more sophisticated example than the one in section 2.1.1 extracted from Charniak (1991). Let us suppose that we arrive home after work and we want to know if our family is at home or not. Before entering, there are a few facts that can give us some clues to answer our question:

- Usually, my wife leaves the outside light switched on when she is away but she also does it when she is waiting for someone.
- When there is nobody at home, the dog is at the back yard. However, in case the dog is ill, he is usually out too.
- When the dog is at the back yard he barks when someone is coming despite it might not be him but the neighbor's dog.

We can gather this data in the following graph:

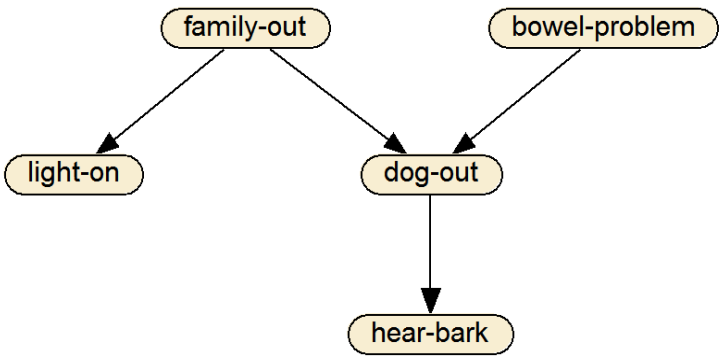


Figure 1: Causal Graph

We can use Bayesian Networks either to analyze what happens in the modeled world (if my family is not at home, then the dog will probably be outside) or to predict events (if the dog is outside it may be nobody at home). Although sometimes, depending on the net characteristics and size, it will be difficult to predict behaviors (if the dog is out but the light is on, will there be someone at home?).

In order to have a fully defined network we need to define the "prior probabilities" of the root nodes (those variables without antecessors) and the conditional probabilities of the other nodes gathered into tables (**CPT:** conditional probabilities tables). The following figure shows the full defined net of our example:

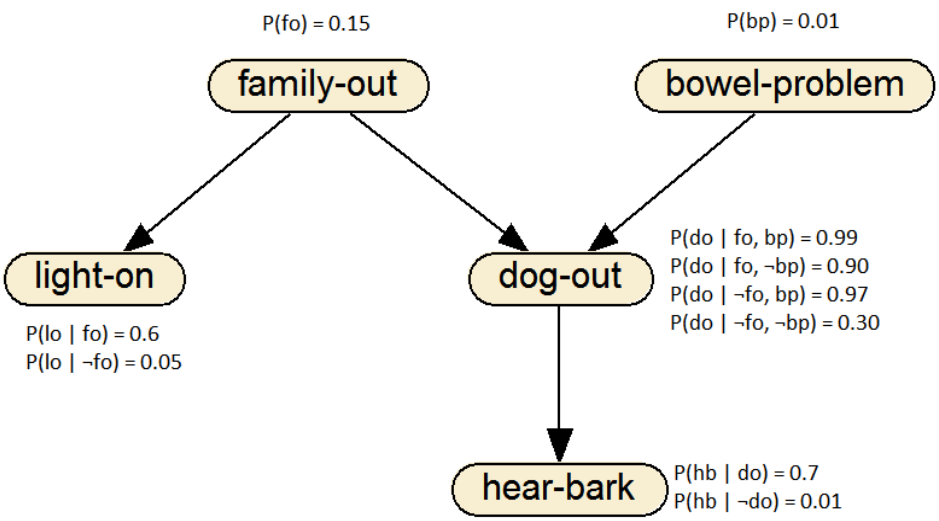


Figure 2: Bayesian network with probability distribution

It is important to emphasize that the causal relationships between variables do not need to be absolute. In our example, it is clear that the family being outside does not necessary imply the light to be switched on. In fact, according to the network, the light will be switched on the 60% of the cases when there is nobody at home but also the 5% of the cases in which the family is at home too.

Figure 2 represents a totally defined network because it is capable of explaining any possible scenario within the modeled world. The net shows, for instance, that if the dog is in the back yard, I will hear him bark the 70% of the cases. However I will also hear him the 1% of the cases in which he is inside the house either because I hear him when he is inside or because I mix him up with the neighbor's dog.

It is also possible to calculate the nodes' conditional probabilities instantiate the probabilities of some of the other nodes. Regarding our example, suppose we arrive home and we see the light switched on (light-on= yes) but we do not hear the dog barking (hear-bark = no), what is the probability of the family being out knowing those events? In other words, which is the value of $P(family-out \mid light-on = yes, hear-bark = yes)$? We illustrate this fact in figure 3.

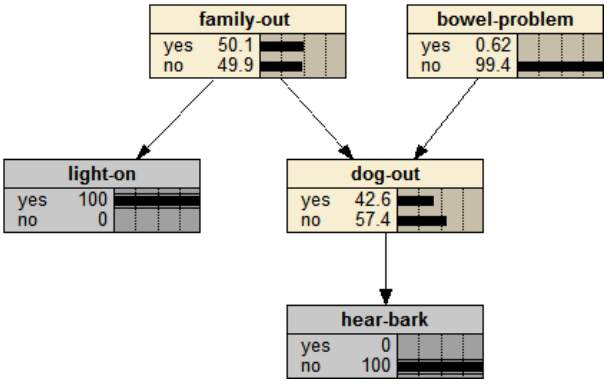


Figure 3: Bayesian network with evidences in nodes *light-on* and *hear-bark*

In our case, given the evidence, there is a 50,1% of chances that the family is away.

2.1.3. Independence assumptions

The main problem of using probabilistic theory is that, in order to have a full description of the probabilistic distribution, a huge amount of values is required. If we have a net with *n* binary random variables at most $2^n - 1$ probability values are needed. In our example we have 5 variables, thus we need $2^5 - 1 = 31$ variables. However we have seen that we just need 10 values to have the net completely defined (see figure 2). How could it

be? That depends on the branching factor of the net and the independences between the variables that we find in the network.

Let define a **d-separation** situation. According to F. V. Jensen (1996) "*two distinct variables A and B in a causal network are **d-separated** if, for all path between A and B, there is an intermediate variable V (distinct from A and B) such that either:*

- *The connection is serial or diverging and V is instantiated*

Or

- *The connection is converging and neither V nor any of V's descendants have received evidence.*

*If A and B are not d-separated, we call them **d-connected**."*

We use our example again to illustrate this fact. We first explain the serial connection case, then the diverging one and finally the converging case. For the former we focus on the nodes *family-out* and *hear-bark*. Are these variables independent? At first there seem not to be independent because if we can hear the dog barking it means that the family is out. However, what happens if we know that the dog is inside? Hearing the dog barking is still meaning that there is nobody at home? Are these variables independent? Before answering this question we have to remark that there is a difference between unconditional independence and conditional independence. This variables are not independent if $P(family-out \mid hear-bark) = P(family-out)$ does not hold. They are conditionally independent on *dog-out*. In other words, $P(family-out \mid hear-bark, dog-out) = P(family-out \mid dog-out)$ holds. That is d-separation depends on the value taken by the common variables (in this case *dog-out*). In our example, if we know exactly that the dog is inside the house then, the *family-out* node and the *hear-bark* one do not depend one on each other, and they are conditionally independent given *dog-out* or d-separated. This example illustrates the so called direct causality. We can find the graphical explanation in figures from 4 to 7.

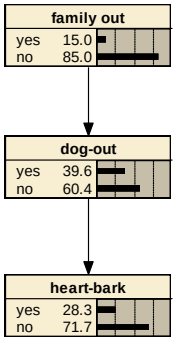


Figure 4: Serial connection without evidences

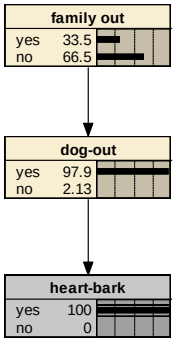


Figure 5: Serial connection with external evidence in node *hear-bark*

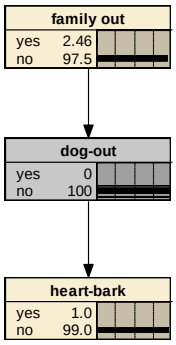


Figure 6: Serial connection with external evidence in node *dog-out*

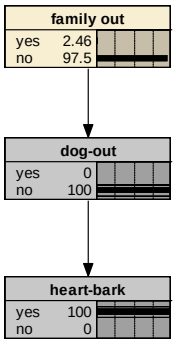


Figure 7: Serial connection with external evidences in nodes *dog-out* and *hear-bark*

To explain the divergence case we focus on the variables *light*-on and *dog*-out, divergently connected by means of the *family*-out node. The *light-on* and *dog-out* variables will be d-separated if and only if we know the state of the node connecting them. Thus, if we are certain that there is nobody at home the *dog-out* state will be conditionally independent given *family-out* from the *light-on* one. However, this will not be the case if we do not know the intermediate variable state (*family-out*). In this case, the *dog-out* value will depend on the *light-on* state, i.e. if we see the light on, there will be more chances of *family-out* being true and, as a consequence, the dog being in the back yard. The following figures (8 to 11) illustrate this particular situation:

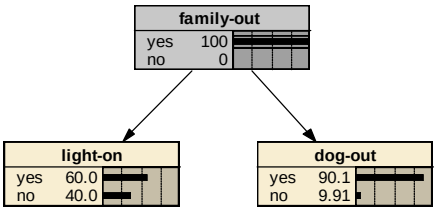


Figure 8: *Diverging connection with external evidence in the family-out node.*

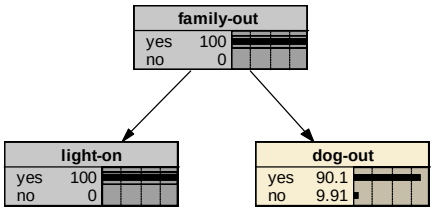


Figure 9*: Diverging connection with external evidence in family-out and light on.*

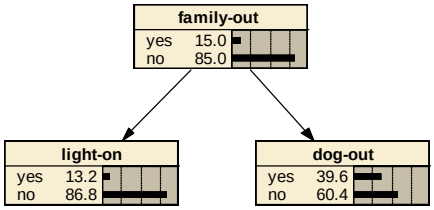


Figure 10*: Diverging connection without evidences.*

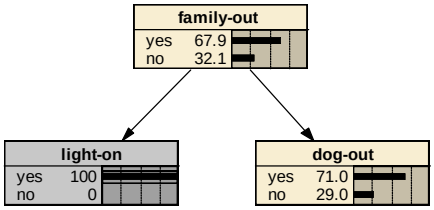


Figure 11*: Diverging connection with evidence in the light-on node.*

Finally, using the variables *family-out* and *bowel-problem* we explain the convergence connection case (both nodes converge into the *dog-out* node). We can see that both variables are only d-connected, and thus conditionally dependent given *dog-out*, when the state of the variable where they converge (*dog-out*) is known.

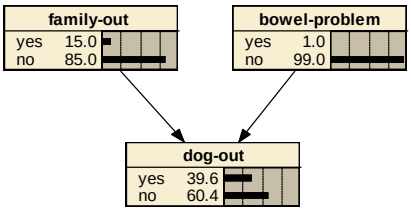


Figure 12*: Converging connection without evidences.*

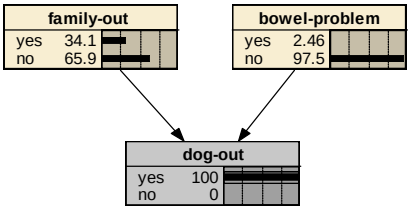


Figure 13*: Converging connection with external evidence in dog-out.*

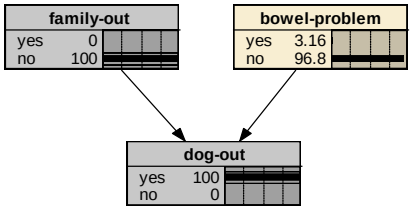


Figure 14*: Converging connection with external evidence in dog out and family out.*

2.1.4. Usage of Bayesian networks

In order to model a particular situation by means of Bayesian networks we need to know which are the necessary variables and their direct causal relationships. Thanks to their easy modeling, this type of causal nets can be applied to a huge number of different fields. Some examples can be found in medical diagnosis problems (Heckermann (1990); Spiegelhalter, Franklin and Bull (1989)), biology and bioinformatics, information retrieval, language understanding, decisions support systems, image processing, engineering, etc.

2.1.5. Learning algorithms

We understand learning in a Bayesian net as updating the process of a Bayesian network as soon as data is available in the form of cases. Each one of the cases represents an event on the modeled world. In our study, every case corresponds to an employee who answered the survey. The cases are based on a set of variables corresponding to the net nodes. Moreover, every node is equivalent to a question from the survey.

This initialized net is used to simulate new situations. There are situations where we usually only have data on one part of the variables, not all of them. These cases enter into the net as findings. Then, in order to determine the distributions of the variables, the probabilistic inference is done and we have the net ready for our analysis.

There are two kinds of learning in Bayesian nets: *structure learning* and *parameter learning*. In our project we are only interested in the latter, which determine the conditional probability tables of each node in the net, given its structure and data. Structure learning, on the other hand, determines the dependencies and independencies between the variables and suggests one causality direction (the place and orientation of the net links). There exist different methodologies to perform this type of learning, but alternatively, a net based on previous knowledge of an expert is equally valid.

Within this work we use NETICA to build up the Bayesian nets offers. This software offers three possible algorithms: *counting*, *expectation-maximization* and *gradient descent*. The order in which the cases are introduced is indifferent for the three algorithms. In other words, four our study, we get the same final net probability distributions no matter the row's order (each row corresponds to a respondent). The most recommended algorithm is the *counting* one because it is by far the quickest and simplest. It can always be used as long as the Bayesian network does not contain latent variables for the learning nodes or its parents.

If it is not possible to use the *counting* algorithm we can use either the *expectation-maximization (EM)* or the *gradient descent* one. The most recommendable way to choose the most appropriated one is to test both of them to see which one produces better results. However, the most commonly used is the EM algorithm due to its robustness, although some times the gradient descent algorithm is quicker.

We now turn to describe how parameter learning is done. The aim of this phase is to find the Bayesian network with maximum likelihood. In other words, we need to find, the most probable net, given the data. Let D be the data and N the Bayesian network. We look for the nets N whose P(N|D) value is the greatest. According to the Bayesian law it is:

$$P(N \mid D) = \frac{P(D \mid N) \cdot P(N)}{P(D)} = \propto P(D \mid N) \cdot P(N)$$

Moreover, to maximize the value $P(N \mid D)$ is equivalent to the maximization of $P(D \mid N) \cdot P(N)$. It is so because the value $P(D)$ is the same in all the networks and, as a consequence, it remains constant. Therefore, we only take into account this $P(D \mid N) \cdot P(N)$ expression. In order to make it simple, we express it in logs so as to have an additive expression.

$$\log(P(D \mid N) \cdot P(N)) = \log P(D \mid N) + \log P(N)$$

Therefore, the aim is now to maximize on one hand $P(D \mid N)$ and on the other hand the value corresponding to $P(N)$, so we analyze them separately.

The first term is what literature has called *log likelihood* and its interpretation is the following. Let the data consist of five independent cases: d_1, d_2, d_3, d_4, d_5. Then, the *log likelihood* is as follows:

$$\log P(D \mid N) = \log\left(P(d_1 \mid N) \cdot P(d_2 \mid N) \cdot P(d_3 \mid N) \cdot P(d_4 \mid N) \cdot P(d_5 \mid N)\right) = \\ = \log P(d_1 \mid N) + \log P(d_2 \mid N) + \log P(d_3 \mid N) + \log P(d_4 \mid N) + \log P(d_5 \mid N)$$

Each one of the terms is easily calculated, introducing every case in the net as a finding. Then, we need to infer the results to refresh the net (Netica will automatically calculate this step) in order to determine the findings' probabilities.

The second term, $\log P(N)$, represents the prior probability of each net (i.e. how probable is every net before entering any data). There are several alternatives to deal with this term. One of them consists on assuming that all the nets are equally probable. This is the simplest approach due to the fact that the value associated with every candidate net will be the same. And, as a consequence, this term $\log \mathrm{P}(\mathrm{N})$ can be ignored because it will sum up the same for every candidate net. Another approach is to penalize the most complex nets assuming them to be less probable than the others. However, this approach is more convenient for structure learning than for parameter learning.

Both, the EM algorithm and the gradient descent are iterative algorithms. The process begins with a candidate net. Then the net is calculated and its log likelihood is reported. The following step is to process the entire set of cases with the algorithm in order to find a better net. The algorithms nature makes the net to be better from iteration to the following. The process iterate until the log likelihood values are worse than the previous ones, given a specified tolerance. However, in case there are a maximum number of iterations prefixed, the process stops when this number is reached.

2.2. Customer Satisfaction

As mentioned in the introduction of this project, nowadays, there are many studies using Bayesian networks primarily as a tool to diagnose or predict. Despite this, it is not an easy task to find projects dealing with customer satisfaction using Bayesian networks. In most studies, this methodology is used to represent worlds where the explanatory variables are identified quickly and easily, such as the cases of medical diagnosis like the "visit Asia" example in Lauritzen and Spiegelhalter (1988).

One of the most similar studies to ours is the project carried out by Silvia Salini from the University of Milan in 2009: "Bayesian Networks of Customer Satisfaction Surveys Data". In that paper, Bayesian networks are used to analyze customer satisfaction surveys and to demonstrate the potential of the approach by means of two examples: one addressing a complex electronic product and a second dealing with the Eurobarometer public opinion surveys.

Our data is collected by a survey, which has been developed by an external person. This means that we have not been able to participate in the survey creation. In other words, sometimes, the design of the survey does not adapt easily to our framework.

3. Bayesian Model for Customer Satisfaction

First, it is necessary to present the scenario in which we perform our study as well as the survey used to collect the data. Using this information we build up the graph representing the world to model. Then we describe the possible nets to build, explaining their features and focusing on the differences between them. In the third step we describe how we transfer the data into the Bayesian network, taking into account whether the net is completely empty (i.e. without prior knowledge) or it has prior distributions before the learning process starts. Before introducing the data we explain the transformation process: from the original Excel file to its introduction in the network by means of the software Netica. Moreover, the algorithms used in the final nets are presented. It is also explained why those algorithms are the most suitable to these type of networks. Finally, we compare the final nets, describing their strengths and weaknesses. Furthermore we recommend the most promising one to be used in subsequent studies on customer satisfaction concerning administrative processes. We also present some future research questions.

3.1. Scenario description

As mentioned in section 1.1, the aim of this project is to analyze the customer satisfaction on administrative processes my means of a particular empirical example or scenario. The latter may be the basis to future research on this field.

Our scenario focuses on administrative processes which are associated with business trips. We are interested in all the procedures to be carried out: all the downstream and the upstream ones. A downstream activity is the one including the preparation details for the trip while an upstream activity consists on control duties related to travel expenses. Finally we understand as during-the-trip-activities those related to transportation facilities and accommodation.

3.1.1. Data collection

Once our scenario is described, we can focus our attention on the survey used to collect the data. We use this data to evaluate customer satisfaction in administrative processes.

The Survey consists of 47 questions divided into four parts:

1) Global satisfaction concerning the process and its sub-processes.

2) Satisfaction regarding the contact person.

3) Communication during the process.

4) Precise satisfaction concerning the sub-processes.

The first part of the survey, *satisfaction with the process*, consists of five questions which allow us to evaluate the general satisfaction of the respondent as well as the satisfaction concerning the necessary sub-processes to carry out a business travel. These questions regard travel request (application form), travel organization (accommodation and transportation booking, document delivery), business travel execution (transportation facilities in the destination, advanced cost payments, rebooking) and travel expenses accounting (fill in the application, accounting throughout the travel management system).

		1 Sehr zufrieden	2	3	4 Überhaupt nicht zufrieden
a) Wie zufrieden sind Sie insgesamt mit dem Reiseprozess am Fraunhofer IPA?	*ZU*	☐	☐	☐	☐
b) Wie zufrieden sind Sie mit dem Ablauf / der Unterstützung bei den folgenden Teilprozessen?					
Reise beantragen (Reiseantrag stellen, Genehmigung einholen)	*BEA*	☐	☐	☐	☐
Reise organisieren (Beförderungsmittel/Unterkunft buchen, Reiseunterlagen übergeben)	*ORG*	☐	☐	☐	☐
Reise durchführen (Beförderungsmittel am Zielort benutzen, Kosten vorstrecken, Umbuchungen durchführen)	*DUR*	☐	☐	☐	☐
Reise abrechnen (Formular ausfüllen, Abrechnung durch Reisemanagement)	*ABR*	☐	☐	☐	☐

Table 1: General questions of the survey

The second section of the survey, *satisfaction regarding the contact person*, consists of just two questions: one concerns the global satisfaction with the contact person, and another evaluates the satisfaction with the contact person's expertise.

		1 Sehr zufrieden	2	3	4 Überhaupt nicht zufrieden
Wie zufrieden sind Sie insgesamt mit Ihren Ansprechpartnern?	ANS	☐	☐	☐	☐
Wie zufrieden sind Sie mit der fachlichen Kompetenz Ihrer Ansprechpartner?	ANS1	☐	☐	☐	☐

Table 2: Survey's area: *contact person*

The third section, *communication throughout the process*, evaluates the degree of satisfaction concerning the communication with the contact person. This evaluation is done, firstly, through a global question and afterwards answering five specific questions regarding the person's performance. These questions also contain information on whether the contact person is easily reachable. Moreover, sometimes the contact person is not available. Then its response (via telephone or mail) is also evaluated in this section.

		1 Trifft voll-kommen zu	2	3	4 trifft über-haupt nicht zu
a) Wie zufrieden sind Sie insgesamt mit der Kommunikation mit den Ansprechpartnern während des Reiseprozesses?	*KOM*	☐	☐	☐	☐
b) Bitte bewerten Sie mit den folgenden Kriterien detailliert Ihre Zufriedenheit mit der Kommunikation.					
Die richtigen Ansprechpartner für den Reiseprozess sind eindeutig bzw. einfach zu identifizieren.	*KOM1*	☐	☐	☐	☐
Die Zeiträume, zu denen ich die Ansprechpartner erreichen kann, sind ausreichend.	*KOM2*	☐	☐	☐	☐
Die verantwortlichen Ansprechpartner im Reiseprozess sind während der täglichen Arbeitszeit ohne Wartezeit telefonisch erreichbar.	*KOM3*	☐	☐	☐	☐
Das Rückrufverhalten der Ansprechpartner ist gut.	*KOM4*	☐	☐	☐	☐
Auf Emails erhalte ich schnell eine Antwort.	*KOM5*	☐	☐	☐	☐

Table 3: Survey's area: *communication*

The last section, *sub-processes satisfaction*, is the most extensive. It consists of 34 questions (70,8% of the survey) and it is divided in four subsections coinciding with the four questions asked in the first section of the mentioned survey: travel application, travel organization, travel execution and settlement of travel expenses.

Travel application: travel request (first two questions and the fourth one), authorization (third, fifth and eighth questions), authorization notification to the employee (sixth and seventh questions).

		1 Trifft voll-kommen zu	2	3	4 trifft über-haupt nicht zu
Zur Erklärung des Ablaufs der Reisebeantragung werden alle notwendigen Informationen kommuniziert.	*BEA1*	☐	☐	☐	☐
Der Unterschied zwischen dem Vorgehen bei der Beantragung von Inland/Ausland-Reisen ist mir bekannt.	*BEA2*	☐	☐	☐	☐
Die richtigen Ansprechpartner zur Genehmigung der Reise sind mir bekannt.	*BEA3*	☐	☐	☐	☐
Die notwendigen Angaben (z.B. Projektnummer) die ich zur Reisebeantragung tätigen muss, sind mir zu diesem Zeitpunkt bekannt.	*BEA4*	☐	☐	☐	☐
Der Ablauf der Reisegenehmigung ist transparent und nachvollziehbar.	*BEA5*	☐	☐	☐	☐
Wird die Reise nicht genehmigt, werde ich als Mitarbeiter umgehend informiert.	*BEA6*	☐	☐	☐	☐
Bei Verzögerung der Reisegenehmigung durch administrative Probleme (z.B. Abwesenheit der zur Reisegenehmigung befugten Person) werde ich als Mitarbeiter umgehend über die Verzögerung informiert.	*BEA7*	☐	☐	☐	☐
Nach der Beantragung der Reise erfolgt die Genehmigung und Benachrichtigung des Mitarbeiters in kurzer Zeit.	*BEA8*	☐	☐	☐	☐

Table 4: Survey's area: *travel application*

Travel organization: This sub-process is divided in three subsections. The first consists on questions about the procedures and development of the travel arrangement process. The second deals with the choice and booking of travel facilities and accommodation. Finally, the third subsection focuses on four different questions evaluating issues related to the required documentation for the trip to take place.

		1 Trifft vollkommen zu	2	3	4 trifft überhaupt nicht zu
ABLAUF UND VORGEHENSWEISE					
Die Kompetenzen und Aufgaben zur Organisation einer Dienstreise zwischen Sekretariaten, Reisemanagement und Mitarbeiter sind klar verteilt („Wer macht was“).	*Abl1*	☐	☐	☐	☐
Vor der Organisation der Dienstreise findet eine klare Kommunikation bezüglich der Anforderungen und Wünsche an die Rahmenbedingungen der Beförderungsmittel- und Hotelbuchung statt. (z.B. Hotel in unmittelbarer Nähe des Kunden).	*Abl2*	☐	☐	☐	☐
Nach der Beantragung der Reise erfolgt die Reiseorganisation in kurzer Zeit.	*Abl3*	☐	☐	☐	☐
AUSWAHL UND BUCHUNG DER BEFÖRDERUNGSMITTEL					
Bei der Buchung von Flug- und Zugreisen wird auf meine Wünsche (z.B. Uhrzeit, Sitzplatz) Rücksicht genommen.	*BF1*	☐	☐	☐	☐
AUSWAHL UND BUCHUNG DER UNTERKUNFT					
Bei der Auswahl und Buchung der Unterkunft wird auf meine Wünsche Rücksicht genommen.	*UK1*	☐	☐	☐	☐
Die Hotelauswahl durch die Sekretariate und das Reisemanagement erfolgt transparent. Es ist ersichtlich warum bestimmte Hotels vor-/ausgewählt wurden.	*UK2*	☐	☐	☐	☐
BEFÖRDERUNGSMITTEL- UND UNTERKUNFTBUCHUNG					
Mit der Buchung der Beförderungsmittel bin ich stets zufrieden.	*BF*	☐	☐	☐	☐
Mit der Auswahl und Buchung der Unterkunft bin ich stets zufrieden.	*UK*	☐	☐	☐	☐
ABSCHLUSS DER REISEORGANISATION					
Die Reiseunterlagen werden rechtzeitig vor der Reise erstellt und an mich übergeben.	*ABS1*	☐	☐	☐	☐
Die Reiseunterlagen sind stets vollständig und fehlerfrei.	*ABS2*	☐	☐	☐	☐
Die Reiseunterlagen müssen nicht auf Vollständigkeit und Fehlerfreiheit kontrolliert werden.	*ABS3*	☐	☐	☐	☐
Bei fehlerhaften oder unvollständigen Unterlagen wird umgehend auf Reklamationen reagiert und der Fehler bzw. das Problem behoben.	*ABS4*	☐	☐	☐	☐

Table 5: Survey’s area: *organization*

Travel execution: it is the sub-process which, chronologically, takes place the third, after applying for the trip and organizing it. There are only three questions within this section focused on possible situations to address during the trip. These are for instance, knowing which form of transportation must be used or, if the employee has to pay out of pocket for something, how much money is reasonable to pay.

		1 Trifft voll-kommen zu	2	3	4 trifft über-haupt nicht zu
Bei Flug- und Bahnreisen ist die weitere Nutzung von Beförderungs-mitteln am Zielort (z.B. Taxi, öffentliche Verkehrsmittel) klar geregelt (Ich weiß, in welchen Fällen ich welches Beförderungsmitteln nutzen darf).	*DUR1*	☐	☐	☐	☐
Ich weiss an wen ich mich bei Problemen während der Reise wenden muss (z.B. Umbuchungen).	*DUR2*	☐	☐	☐	☐
Die Verauslagung der anfallenden Reisekosten durch den Mitarbeiter ist sinnvoll.	*DUR3*	☐	☐	☐	☐

Table 6: Survey's area: *travel execution*

Settlement of travel expenses: this sub-process, as the second one (travel organization) is divided in three subsections. The first one is focused on evaluating the accounting procedure. It consists of three questions trying to find out the customer satisfaction with the procedure in general, as well as the suitability and appropriateness of the red tape. The second subsection tries to figure out if the form is easy and intuitive, if the amount of data is appropriate and if the reason why the data is required is comprehensible. Finally, and to finish with the survey, the last subsection focuses on employee satisfaction with the accounting verification. This evaluation is done through five questions regarding specific items such as the processing time, the communication of the transfer date to the employee, the management precision and the responses when failures occur.

		1 Trifft vollkommen zu	2	3	4 trifft überhaupt nicht zu
VORGEHEN DER REISEKOSTENABRECHNUNG					
Der Ablauf zur korrekten Dienstreiseabrechnung ist klar definiert.	*VOR1*	☐	☐	☐	☐
Der bürokratische Aufwand zur Abrechnung der Reise ist angemessen.	*VOR2*	☐	☐	☐	☐
Die Art und Weise der Durchführung der Dienstreiseabrechnung bspw. mit Ausfüllen des Dokuments „Erklärungen zum Reiseablauf“ ist zweckmäßig.	*VOR3*	☐	☐	☐	☐
FORMULAR “ERKLÄRUNGEN ZUM REISEABLAUF” (FORMULAR ZUR REISEKOSTENABRECHNUNG)					
Das Formular ist einfach und intuitiv auszufüllen.	*FOR1*	☐	☐	☐	☐
Der Umfang der geforderten Angaben ist angemessen.	*FOR2*	☐	☐	☐	☐
Es ist nachvollziehbar, warum die jeweiligen Angaben gefordert werden.	*FOR3*	☐	☐	☐	☐
PRÜFUNG UND ABSCHLUSS DER REISEKOSTENABRECHNUNG					
Nach dem Einreichen der Unterlagen erfolgt die Abwicklung der Dienstreiseabrechnung in kurzer Zeit.	*PRU1*	☐	☐	☐	☐
Nach Bearbeitung der Dienstreiseabrechnung durch das Reisemanagement wird mir die Vollständigkeit und Korrektheit meiner Angaben sowie der Zeitpunkt der Überweisung mitgeteilt.	*PRU2*	☐	☐	☐	☐
Die Reisekostenabrechnung durch das Reisemanagement ist stets vollständig und fehlerfrei.	*PRU3*	☐	☐	☐	☐
Die Korrektheit der Dienstreiseabrechnung muss von mir nicht kontrolliert werden.	*PRU4*	☐	☐	☐	☐
Bei unvollständiger bzw. fehlerhafter Dienstreiseabrechnung wird nach Reklamation unmittelbar gehandelt.	*PRU5*	☐	☐	☐	☐

Table 7: Survey’s area: *settlement of travel expenses*

3.2. Network alternatives

We now turn to describe the nets built up to model our business trip scenario. In order to be as accurate as possible, each one of the 47 questions from the survey corresponds to the 47 different variables in the model. Moreover, these variables have four possible, exhaustive and mutually exclusive, states. We name each of the variables by an abbreviation corresponding to the subsection or group of questions they belong to. We can find these abbreviations in the second column of tables 1 to 7. Therefore, the variables are:

ZU												
BEA	BEA1	BEA2	BEA3	BEA4	BEA5	BEA6	BEA7	BEA8				
ORG	Abl1	Abl2	Abl3	BF	BF1	UK	UK1	UK2	ABS1	ABS2	ABS3	ABS4
DUR	DUR1	DUR2	DUR3									
ABR	VOR1	VOR2	VOR3	FOR1	FOR2	FOR3	PRU1	PRU2	PRU3	PRU4	PRU5	
ANS	ANS1											
KOM	KOM1	KOM2	KOM3	KOM4	KOM5							

Table 8: Summary of abbreviations

For any node we define four possible states: *sehr zufrieden* (very satisfied), *zufrieden* (satisfied), *nicht zufrieden* (not satisfied) or *ganz unzufrieden* (not satisfied at all). With each state, a numerical value is associated; corresponding to a scale, where 1 is the best value and 4 is the worst. As mentioned before, in order to apply the learning algorithms to Bayesian nets, the nodes must be discrete. In case they are continuous, the algorithm will first discretize the variables by defining intervals for any possible state of a variable. The latter is our case because there are some nodes with an associate equation calculating the state of this node. Therefore, in order to use these equations, we need to define first the intervals as follows:

$$sehr\ zufrieden = [0{,}5 - 1{,}5)$$ (very satisfied)

$$zufrieden = [1{,}5 - 2{,}5)$$ (satisfied)

$$unzufrieden = [2{,}5 - 3{,}5)$$ (not satisfied)

$$ganz\ unzufrieden = [3{,}5 - 4{,}5]$$ (not satisfied at all)

It is also important to mention the case of *latent variables* or *hidden nodes*. Those are variables deduced from other observable ones and thus they do not contain information deduced from the collected data (in our case collected directly from the survey). The applicable algorithms to Bayesian networks with latent variables differ from those applicable to networks without them. For this reason, it is important to know which net we are facing. In this project we work simultaneously with two nets, one with latent variables and another without them, presenting only the nodes corresponding directly to the survey questions.

What we learn about our variables is illustrated in the further example. If we take one of our variables as, for instance, FOR1, corresponding to the question *"Das Formular ist einfach und intuitiv auszufüllen"* (The application form can be fulled out easily and intuitively), the information we find in the net is collected in the followings table and figure:

Name of the variable:	FOR1	discrete case	continuous case
Possible states:	sehr zufrieden	1	0,5 - 1,5
	Zufrieden	2	1,5 - 2,5
	Unzufrieden	3	2,5 - 3,5
	ganz unzufrieden	4	3,5 - 4,5

Table 9: Nodes‘ states

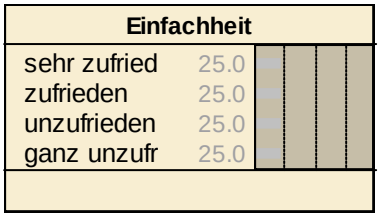


Figure 15: Node FOR1

The links, as mentioned in section 2.1.1, are the ones containing the information about the relationship between the variables they connect. The information is transmitted though the CPT matrices. Their size is important because the bigger the table, the harder to obtain data for all probabilities in the table. Therefore, it is better to have a lot of tables with few entries rather than few tables but very large. We now explain how to calculate the tables' size. If we take a subnet from the total network as, for example, *travel execution* we find the following graph:

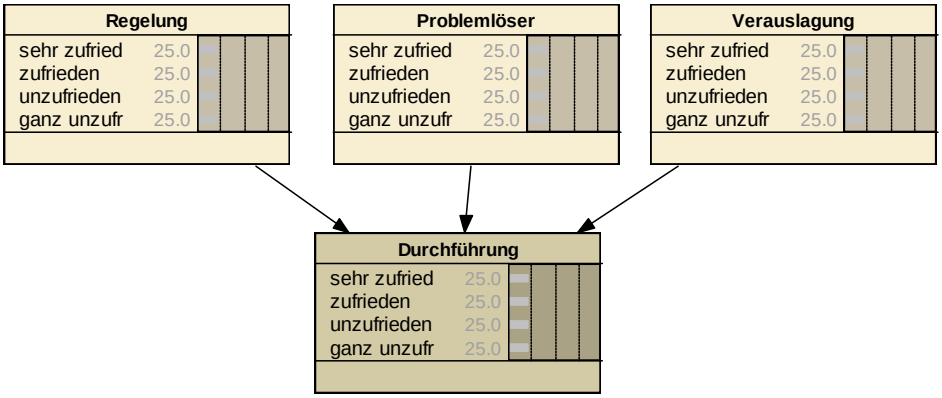


Figure 16: Subnet *travel execution*

We can see that it is a four-node net, three of them converging to one. The variables are the followings:

- ✓ Durchführung → *DUR*
- ✓ Regelung → *DUR1*
- ✓ Problemlöser → *DUR2*
- ✓ Verauslagung → *DUR3*

As every parent node (*DUR1, DUR2 and DUR3)* has four possible states, the number of rows of our conditional probabilities table in the son node (*DUR*) is $4^3 = 64$. If we take into account that the son node has four states too, i.e. that for every row *DUR* can take four possible values, the total number of values in the CPT will be $64 \cdot 4 = 256$. We can check the previous calculation looking at the CPT (table 10) that arises from the example:

DUR1	DUR2	DUR3	DUR			
			1	2	3	4
1	1	1				
1	1	2				
1	1	3				
1	1	4				
1	2	1				
1	2	2				
1	2	3				
1	2	4				
1	3	1				
1	3	2				
1	3	3				
1	3	4				
1	4	1				
1	4	2				
1	4	3				
1	4	4				

DUR1	DUR2	DUR3	DUR			
			1	2	3	4
2	1	1				
2	1	2				
2	1	3				
2	1	4				
2	2	1				
2	2	2				
2	2	3				
2	2	4				
2	3	1				
2	3	2				
2	3	3				
2	3	4				
2	4	1				
2	4	2				
2	4	3				
2	4	4				

DUR1	DUR2	DUR3	DUR			
			1	2	3	4
3	1	1				
3	1	2				
3	1	3				
3	1	4				
3	2	1				
3	2	2				
1	2	3				
3	2	4				
3	3	1				
3	3	2				
3	3	3				
3	3	4				
3	4	1				
3	4	2				
3	4	3				
3	4	4				

DUR1	DUR2	DUR3	DUR			
			1	2	3	4
4	1	1				
4	1	2				
4	1	3				
4	1	4				
4	2	1				
4	2	2				
4	2	3				
4	2	4				
4	3	1				
4	3	2				
4	3	3				
4	3	4				
4	4	1				
4	4	2				
4	4	3				
4	4	4				

Table 10: Conditional probability table for travel execution subnet

We can observe that, even in a simple case such as this one, the CPT's size is quite large. The reasons for this to happen are, on the one hand the fact that each variable has four possible states and, on the other hand that all the variables converge to the same node. It is for that reason that we present two nets simultaneously. As we will see in section 3.2.1, if we study the subnets belonging to the main network, we find two branches where: 8 nodes converge in one in the travel application case, 9 nodes converge in one in the organization case, and, in the worst-case scenario, 11 nodes converge in one in the settlement of account case. Therefore, this implies the tables to have $4^8 = 65.536$; $4^9 = 262.144$ and $4^{11} = 4.194.304$ rows, respectively. We can see that this fact can cause real problems for our research because it is impossible to find a company with more than 4 million employees traveling by business and, as a consequence, being able to answer the survey on a reliable way (the biggest corporate in the world is the supermarket chain Wal-Mart stores with approximately 2.100.000 employees). Thus, we need to find a solution to this problem.

Another important issue is the following: What happens if a CPT row does not have any data on it? Or how does the software behave in case the tables are only partially filled

with data? The most logical answer to those questions is that if I do not know the prior probabilities of a node, I simply allocate them uniformly. In other words, if any respondent answered 3, 4, 1 to the nodes DUR1, DUR2 and DUR3 respectively, the table does not receive any data. Then, Netica will assign a 25% of probability to each one of the four possible states of the corresponding row in our node DUR. Table 11 illustrates this example.

DUR1	DUR2	DUR3	DUR			
			1	2	3	4
	…					
3	4	1	25%	25%	25%	25%
	…					

Table 11: Uniform distribution in one arbitrary case for node DUR.

We now turn to find a solution to the tables' size problem. We present two possible solutions. The first one consists on writing an equation in the destination node or son node. In other words, if all the parent variables present the same value, the son variable should take this same value as well. In case the three variables present different values, their weight will determine the value of the son node. Such weights depend on the importance of the parent variable with respect to the others. Fortunately, in our survey we do not only get the level of satisfaction, but also the degree of importance of every variable. This importance is captured by an ordinal scale from one to six (1 is very important and 6 is completely insignificant). Thanks to that scale, we can weigh the importance of the variables to build up the equations in the destination nodes.

We illustrate now how to calculate these equations with an example. We use, once again, the Bayesian subnet *travel execution*. Let suppose that we first receive the degree of importance of the variables *DUR1*, *DUR2* and *DUR3*. Then, we calculate the average importance of each of the parent variables which are 2,33; 1,98 and 1,96 for *DUR1*, *DUR2* and *DUR3*, respectively (see table 12). The sum of these means is 6,27. Therefore, if we want to know the weighted value of *DUR1* with respect to the other two variables, we just need to divide its mean by the sum of averages (i.e. $\frac{2,33}{6,27} = 0,37131$) and so on.

WDUR1	WDUR2	WDUR3	∑
2.33	1.98	1.96	6.27
37.131%	31.601%	31.268%	100%

Table 12: Calculation values of equation for node DUR

At the end, the resulting equation for DUR node is the following.

$$DUR = 0{,}37131 \cdot DUR1 + 0{,}31601 \cdot DUR2 + 0{,}31268 \cdot DUR3$$

In order to use equations in a net, the variables might be either continuous or discrete if we want to calculate them with the software Netica. However, in case of them being continuous, they must be discretized before turn them into conditional probability tables, as mentioned before. When converting an equation in a CPT, we must choose a number of points to evaluate them in each interval for every variable. Netica uses the average of the values obtained from the equation. Obviously, the greater the number, the better the precision of the result.

As all the points within an interval cannot be chosen (we choose the number of points to evaluate per interval), there is a chance to add additional uncertainty due to sampling (on the contrary the sampling is supposed to be representative enough). When we add additional uncertainty no value from the table is null because we cannot afford to ignore any significant value in the interval.

Another solution to reduce the size of the tables is to incorporate latent variables or hidden nodes. Those nodes do not correspond to any of the entry variables and, therefore, they do not have any observations but they might be useful to model the scenario. So, this variable is created to express, in an easy way, the relations between the observed variables. It might look weird to do the learning of a net containing this type of variables but, their introduction may lead to easier networks (with less CPT entries). A numerical and empirical case is explained in the next section.

3.2.1. Subnets description

According to the previous sections, every node in this network corresponds to a question in the survey too. Therefore, the network follows a hierarchical structure and comprises six subnets. Those are: *contact person*, *execution*, *communication*, *travel application*, *travel organization* and *settlement of account*.

Contact person: in this subnet, dedicated to the contact person, we only find two nodes, corresponding to ANS and ANS1. Netica allows us to entitle the node differently from the variables' name. We do so in order to facilitate the interpretation and diagnosis of the network. The relation between the titles and the variables within this subsection is the following: "fachliche Kompetenz" corresponds to ANS1 and Ansprechpartner corresponds to ANS.

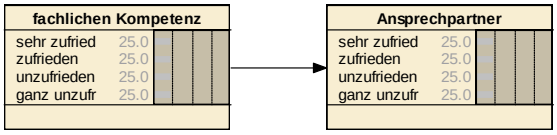


Figure 17: Subnet *contact person*

Travel execution: in this subnet we find four questions from the survey belonging to the subsection with the same name (travel execution), whose variables are DUR, DUR1, DUR2 and DUR3. The relationship between the titles and the variables are the followings: *DUR*=„Durchführung"; *DUR1*=„Regelung"; *DUR2*=„Problemlöser"; *DUR3*=„Verauslagung".

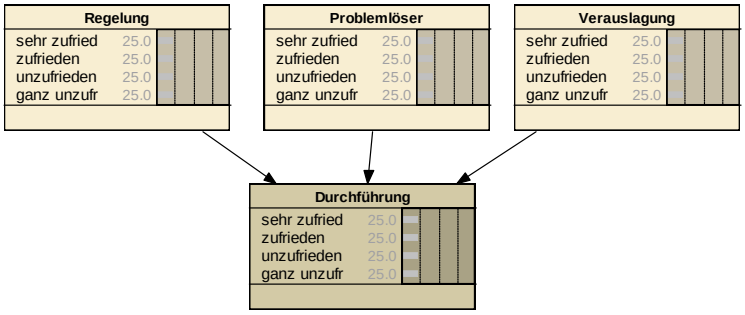


Figure 18: Subnet *travel execution*

Communication: We find six nodes in this subnets corresponding to the variables *KOM*, *KOM1*, *KOM2*, *KOM3*, *KOM4* y *KOM5*. The relation variable- title in this node is the following: KOM=„Kommunikation“; *KOM1*=„Richtig Ansprechpart.“; *KOM2*=„Bürozeiten“; *KOM3*=„Erreichbarkeit“; *KOM4*=„Rückrufverhalten“; *KOM5*=„E-mail Antwort“.

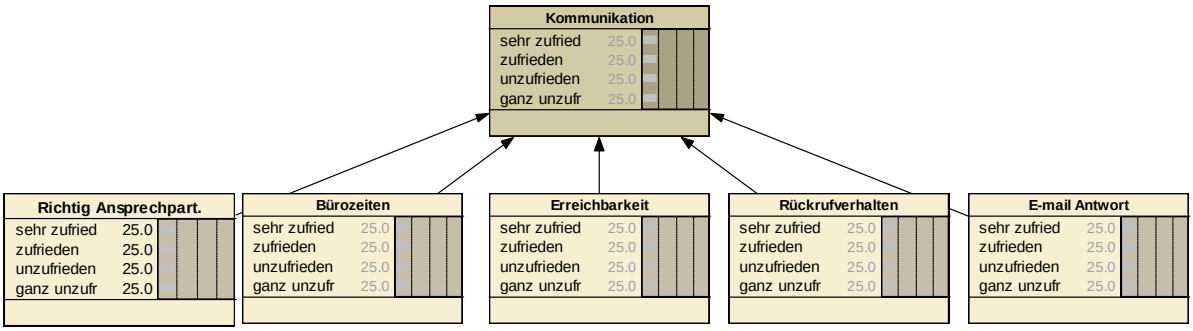


Figure 19: Subnet *communication*

Travel application: in this fourth subnet, corresponding to the sub-process previously described and named equally we find nine nodes. Here the relation variable-node is the following: *BEA*=„Beantragung“; *BEA1*=„Kommu notw Infos“; *BEA2*=„In- vs Auland-Reisen“; *BEA3*=„Richtig. Ansprechp.“; *BEA4*=„Notw. Angaben“; *BEA5*=„Ablauf Transparent“; *BEA6*=„nicht-Genehmigung“; *BEA7*=„Verzörgerng“; *BEA8*=„Zeitverluste“.

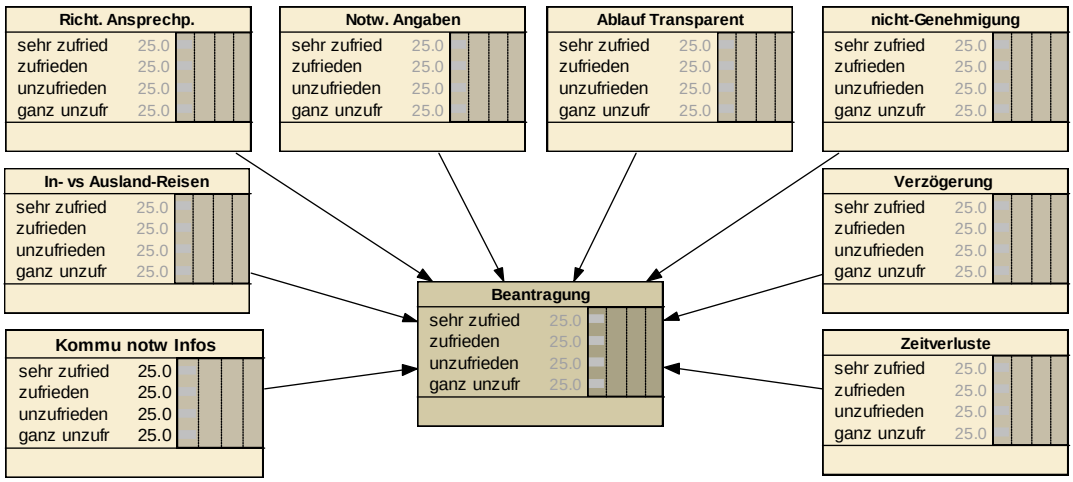


Figure 20: Subnet *travel application*

Travel organization: this subnet is one of the most complete. It consists of thirteen nodes corresponding to the questions related to the travel organization. The relation with the variables is the following: *ORG*=„Organisation“; *Abl1*=„Kompetenzenverteilung“; *Abl2*=„Klare Organisation“; *Abl3*=„Zeitabstand“; *BF*=„Beförderungsmittel“; *UK*=„Unterkunft“; *UK1*=„Wünscherücksichtigung“; *UK2*=„Hotelauswahl“; *ABS1*=„Unterlagen rechtzeitig“; *ABS2*=„Unterlagen fehlerfrei“; *ABS3*=„Keine Kontrolle“; *ABS4*=„Reklamationsreaktion“. In this subnet is where we find for the first time a difference between the two nets that we propose. We find this difference in the number of nodes of the networks. The net containing latent variables comprises two extra nodes (hidden nodes) that gather, on the one hand, the nodes *Abl1*, *Abl2* y *Abl3* (that converge in the new node *Abl*) and, on the other hand, *ABS1*, *ABS2*, *ABS3* and *ABS4* (that converge in the new node ABS).

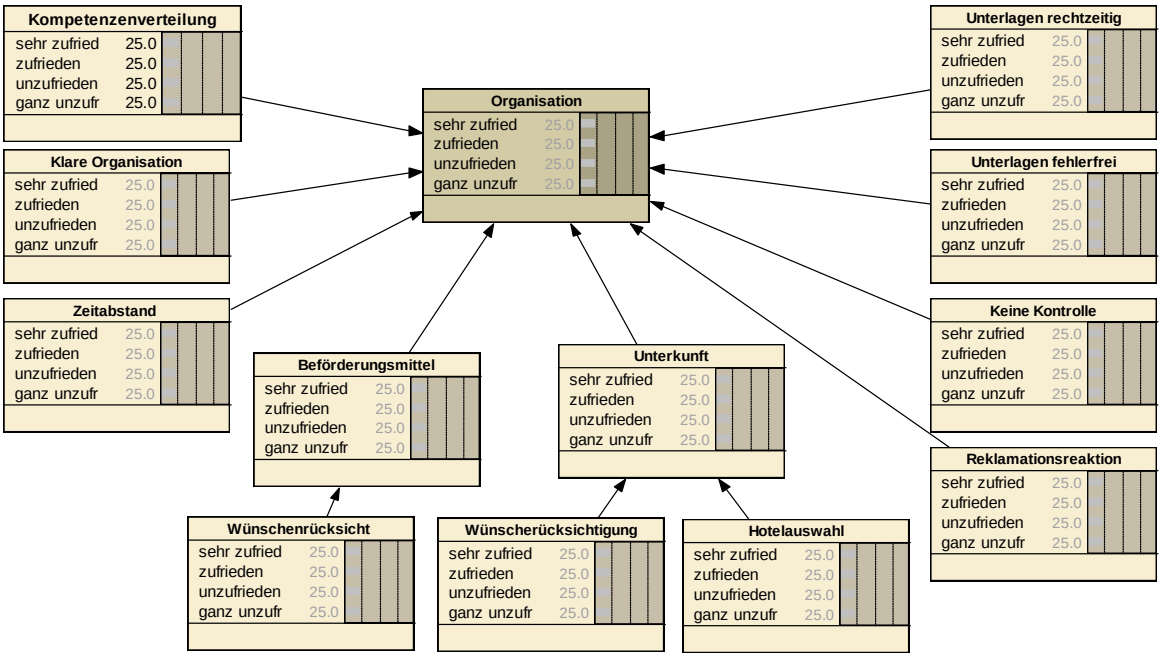


Figure 21: Subnet *organization*

Settlement of accounts: In this section we find twelve nodes. It is the second biggest in terms of the amount of nodes but it is also the one with the greatest CPT. It is so because 11 nodes converge into one and that implies a table with $4^{11} = 4.194.304$ rows and 4 columns, i.e., $4.194.304 \cdot 4 = 16.777.216$ values. When talking about the nets nomenclature we find the following names for the different nodes: *VOR1*=„Def. der Doku-Ablauf“; *VOR2*=„bürokratische Aufwand“; *VOR3*=„Durchführung“; *FOR1*=„Einfachheit“;

FOR2=„nachgefragt. Angaben“; *FOR3*=„Nachvollziehbarkeit“; *PRU1*=„Abwicklungszeit“; *PRU2*=„Mitteilungen“; *PRU3*=„Fehlerfreiheit“; *PRU4*=„Korrektheitskontrolle“; *PRU5*=„Reklamationen“.

In this subnet is where we find the second difference between the two final nets. We add three extra nodes in the net containing hidden nodes that classify the subnet in three different categories. *VOR1*, *VOR2* and *VOR3* converge to node *VOR*. *FOR1*, *FOR2* and *FOR3* do so in node *FOR* and *PRU1*, *PRU2*, *PRU3*, *PRU4* and *PRU5* converge in node *PRU*. At the same time, *VOR*, *FOR* and *PRU* converge into the node *ABR* so that the CPT from this node changes from 4.194.304 rows to four simple tables, three in the new nodes and one in the *ABR* node. The size of the new tables is 64 rows for the ones in the nodes *VOR*, *FOR* and *ABR* and 1.024 rows for the *PRU* node. Therefore we now just need values for 1024 cases instead of 4.000.000 (a 99.975% reduction of entries).

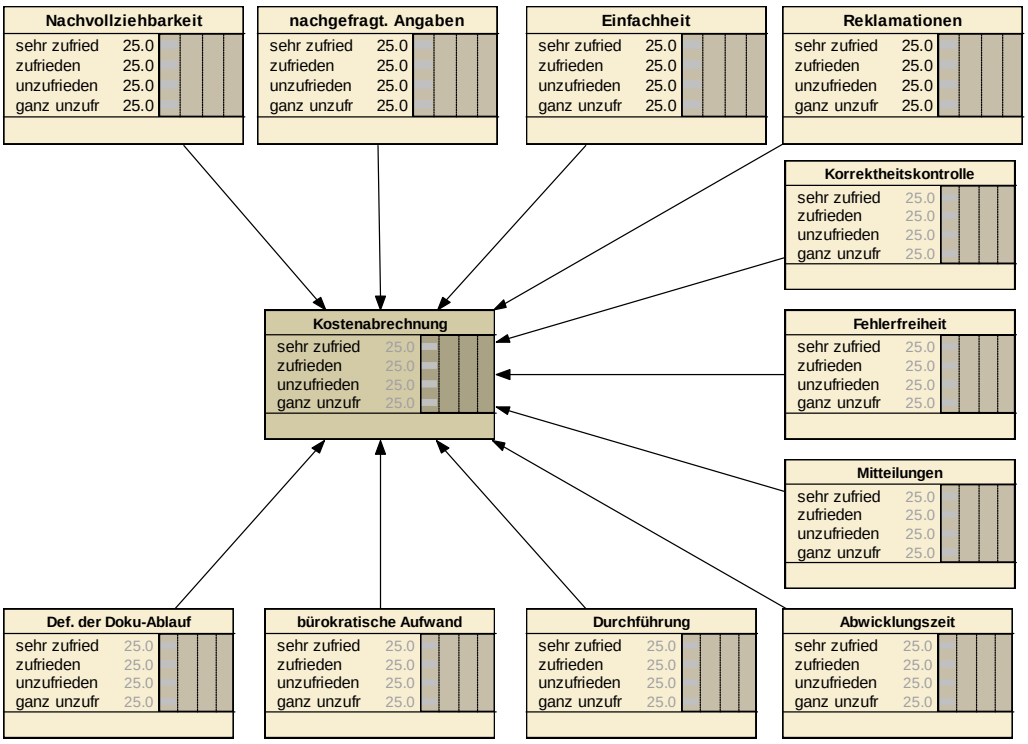


Figure 22: Subnet *settlement of accounts*

3.2.2. Subnet equations

Before getting into the final networks, we describe the equations that accompany the shadowed nodes from the previously mentioned subnets.

Travel execution: in this subnet, we find an equation in the DUR node. It is described in subsection 3.2 and it is the following:

$$DUR = 0{,}37131 \cdot DUR1 + 0{,}31601 \cdot DUR2 + 0{,}31268 \cdot DUR3$$

Communication: in this section, the average values concerning the degree of importance for the questions corresponding to the variables KOM1, KOM2, KOM3, KOM4 and KOM5 are 1,98; 1,98; 2,19; 1,95 and 1,98, respectively. Therefore, the weights for each node are: 19,666%, 19,666%, 21,688%, 19,309% and 19,670%, respectively too. Given this data, the equation for the node KOM is the following:

$$\begin{aligned} KOM = {} & 0{,}19666 \cdot KOM1 + 0{,}19666 \cdot KOM2 + 0{,}21688 \cdot KOM3 + \\ & + 0{,}19309 \cdot KOM4 + 0{,}19670 \cdot KOM5 \end{aligned}$$

Travel application: the data regarding the degree of importance valuations are 2,08; 2,94; 2,10; 2,35; 2,51; 2,18; 2,00 and 1,74 for the nodes *BEA1*, *BEA2*, *BEA3*, *BEA4*, *BEA5*, *BEA6*, *BEA7* and *BEA8*. Therefore, the resulting equation for the node *BEA* is the following:

$$\begin{aligned} BEA = {} & 0{,}11621 \cdot BEA1 + 0{,}16412 \cdot BEA2 + 0{,}11733 \cdot BEA3 + 0{,}13112 \cdot BEA4 + \\ & + 0{,}14024 \cdot BEA5 + 0{,}12177 \cdot BEA6 + 0{,}11174 \cdot BEA7 + 0{,}09747 \cdot BEA8 \end{aligned}$$

Travel organization: in this section we find two nodes with an equation (*ORG* and *UK*), i.e., two equations in total. We first present the equation corresponding to the accommodation node, *UK*. The parent nodes average, *UK1* and *UK2*, are 1,50 and 1,97, respectively. Therefore the equation from the *UK* is the following:

$$UK = 0{,}43182 \cdot UK1 + 0{,}56818 \cdot UK2$$

The equation concerning to the *ORG* is calculated with the following values for the nodes *Abl1*, *Abl2*, *Abl3*, *BF*, *UK*, *ABS1*, *ABS2*, *ABS3* and *ABS4*: 2,10; 1,67; 1,50; 1,63; 3,47; 1,48; 1,53; 2,04 and 1,51. Given this values, the resulting equation is the following:

$$\begin{aligned} ORG = {} & 0{,}12413 \cdot Abl1 + 0{,}09840 \cdot Abl2 + 0{,}08856 \cdot Abl3 + 0{,}09633 \cdot BF + 0{,}20508 \cdot UK \\ & + 0{,}08738 \cdot ABS1 + 0{,}09037 \cdot ABS2 + 0{,}12054 \cdot ABS3 + 0{,}08921 \cdot ABS4 \end{aligned}$$

Settlement of accounting: finally, the equation from the *ABR* node is calculated the same way tan the previous ones but taking the following values for the nodes *VOR1*, *VOR2*, *VOR3*, *FOR1*, *FOR2*, *FOR3*, *PRU1*, *PRU2*, *PRU3*, *PRU4* and *PRU5*: 2,18; 1,72; 1,96; 1,76; 1,74; 2,36; 1,78; 2,12; 1,48; 1,61 and 1,40. Therefore, the equation is:

$$\begin{aligned} ABR = {} & 0{,}10840 \cdot VOR1 + 0{,}08552 \cdot VOR2 + 0{,}09742 \cdot VOR3 + \\ & + 0{,}08751 \cdot FOR1 + 0{,}08652 \cdot FOR2 + 0{,}11735 \cdot FOR3 + \\ & + 0{,}08851 \cdot PRU1 + 0{,}10541 \cdot PRU2 + 0{,}07359 \cdot PRU3 + 0{,}08017 \cdot PRU4 + 0{,}06961 \cdot PRU5 \end{aligned}$$

3.2.3. The final networks

We have previously described all the parts of our Bayesian nets, so we now turn to explain the two final networks. It basically consists on the union of the six subnets described in the previous section by means of a common son node. This new node is entitled in the graph as “Allgemeine Zufriedenheit” and corresponds to the variable *ZU*.

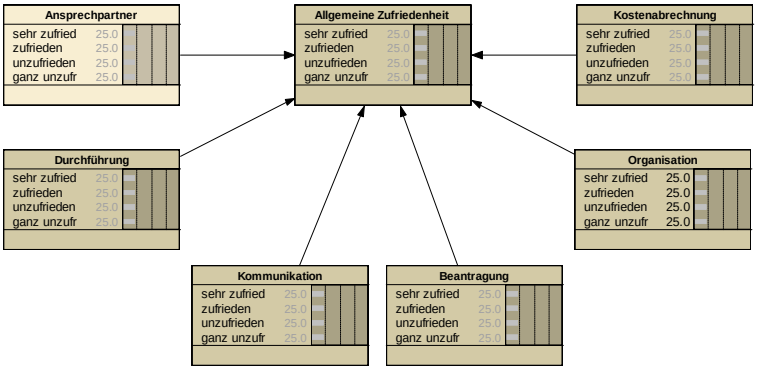


Figure 23: Part of the net with nodes *ZU*, *ANS*, *KOM*, *BEA*, *ORG*, *DUR* and *ABR*

Therefore, the first net, the one without hidden nodes, is the following:

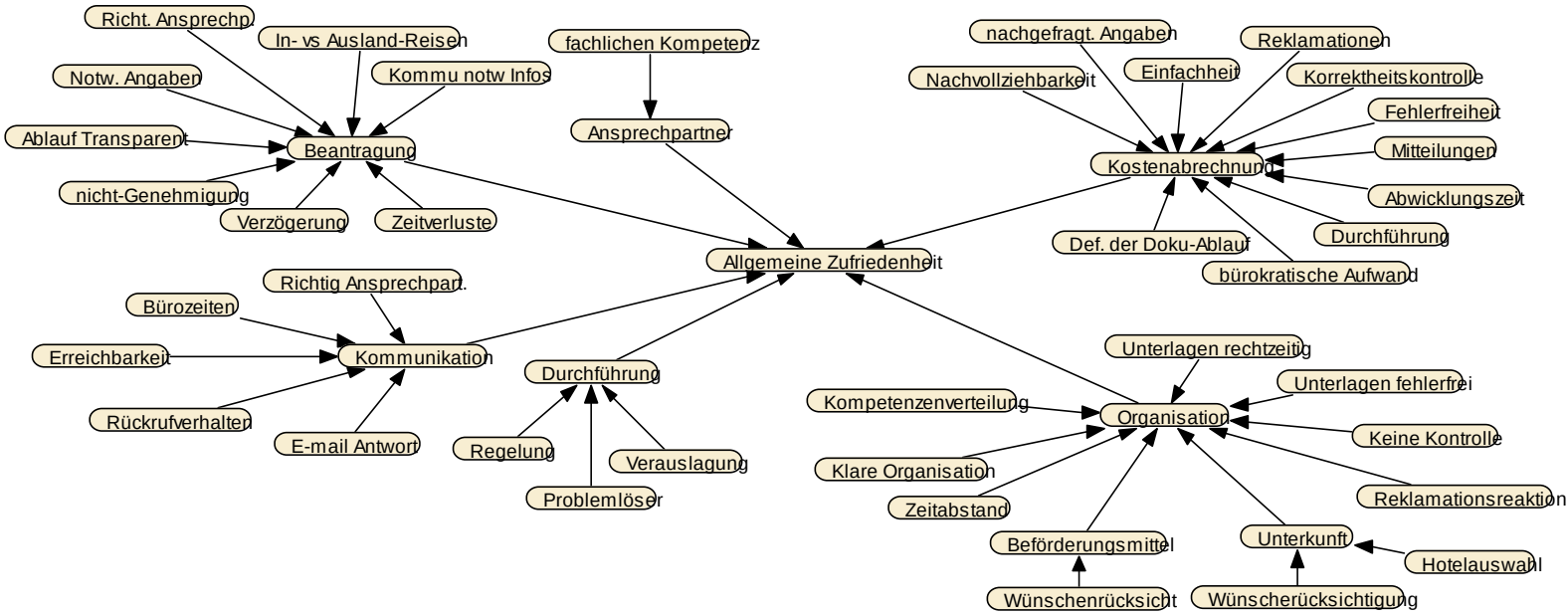


Figure 24: Net without latent variables or hidden nodes

And the alternative network, with hidden nodes, is:

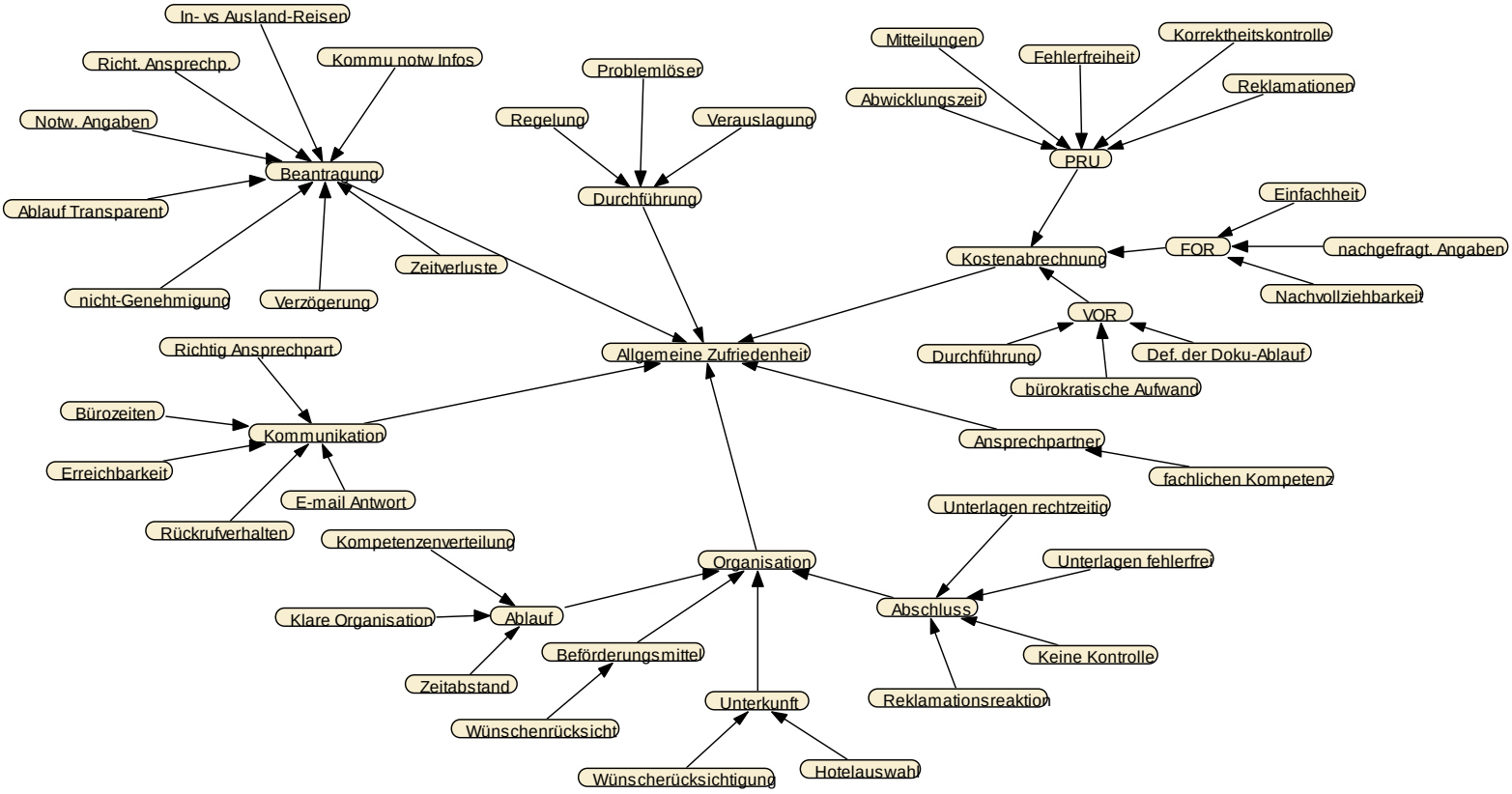


Figure 25: Net with hidden nodes or latent variables

3.2.4. Prior Knowledge

Once the Bayesian network has been described, we need to know whether we have prior knowledge or not. Besides, in case we have it, we must justify its origin and shape. Netica allows us to do the learning from data with empty tables. Therefore we compare our net's behavior both in case of data without prior knowledge and data with it. In the latter case, we explore different hypothesis. The first one consists on assuming a completely lack of information about worker's satisfaction in all nodes. As a consequence, we allocate the probabilities tending to one (i.e. "sehr zufrieden"). Finally, the last hypothesis is the opposite case in which the workers are not satisfied in any of the root nodes.

In order to compare the net's behavior between the different cases, we only apply prior knowledge into the "travel execution" subnet. We do so because the results are much more interpretable in a small Bayesian network rather than in a net with 47 or 47+5=52 nodes (we would barely see the probabilities from each node).

3.2.4.1. Acquisition

Our database consists on a survey answered by 301 workers from Fraunhofer IPA institute. Unfortunately, the sample size is small because it comprises only 66 observations (21,9% of the respondents). Taking into account this small size, it seems important to study and compare the different behaviors after the learning process with or without prior knowledge.

Table 13 presents part of our data set. An initial inspection shows a huge number of missing values. Therefore, the question is the following: Are these missing values randomly distributed or do they follow any pattern? To answer this question we examine the data set both horizontally and vertically. The horizontal analysis counts, for any observation or row, the number of answered questions. We can see 13 rows where the respondent answered less than the 30% of the 47 questions (33 or more missing values per row). In table 13 we observe how in observations 16, 34, 62 and 63 no question was answered. Moreover there are other cases where just a few questions were answered such as observation 10 in which

only the question corresponding to the ZU variables was answered and observation 38 with information only on the first section of the survey is available (ZU, BEA, ORG, DUR and ABR).

Furthermore observations 3, 15, 39 and 40 where only the 9 first questions are answered (ZU, BEA, ORG, DUR, ABR, ANS, ANS1, KOM y KOM1) and observations 36 and 48 where the information for the 13 first questions is available (ZU, BEA, ORG, DUR, ABR, ANS, ANS1, KOM, KOM1, KOM2, KOM3, KOM4 y KOM5) are shown.

Idnum	ZU	BEA	ORG	DUR	ABR	ANS	ANS1	KOM	KOM1	KOM2	KOM3	KOM4	KOM5	BEA1	BEA2	BEA3	BEA4	BEA5	BEA6	BEA7	...
16	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	...
34	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	...
62	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	...
63	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	...
10	2	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	...
38	3	1	1	2	4	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	...
3	2	1	2	3	2	1	1	*	*	*	*	*	*	*	*	*	*	*	*	*	...
15	2	1	3	4	1	1	3	*	*	*	*	*	*	*	*	*	*	*	*	*	...
39	4	1	1	2	4	2	3	*	*	*	*	*	*	*	*	*	*	*	*	*	...
40	2	2	2	2	4	2	2	*	*	*	*	*	*	*	*	*	*	*	*	*	...
17	3	3	1	1	2	1	1	2	2	*	*	*	*	*	*	*	*	*	*	*	...
36	4	2	2	3	4	1	1	2	3	3	2	1	1	*	*	*	*	*	*	*	...
48	2	1	1	3	2	1	1	1	1	1	1	1	1	*	*	*	*	*	*	*	...

Table 13: observations with missing variables

If we vertically analyze the data set, we observe that the nodes BEA6 and BEA7 are the ones with the highest number of missing values (47% and 39% respectively). These nodes contain questions about the application rejection or delay. From the previous facts, we can extract the conclusion that the worker may either have not experienced any of these situations or considered himself not enough able to answer these questions.

3.2.4.2. Formalization

In this section we explain how the data file must be in order to the software Netica to read it. There is only one requirement for Netica to read data files containing a set of cases compatible with the network. It is that the name of the columns in the input file (in our case this name is in the first row of our Excel file) must be identical to the variables name. In other words let DUR, DUR1, DUR2 and DUR3 be the only nodes in our net. Then, the values of the columns in our Excel table, where we have the 66 cases from the survey, must be DUR, DUR1, DUR2 and DUR3. If this criterion is fulfilled, we just need not to have any value being different from the possible states of the network nodes, and we will have our dataset ready to be applied to Netica.

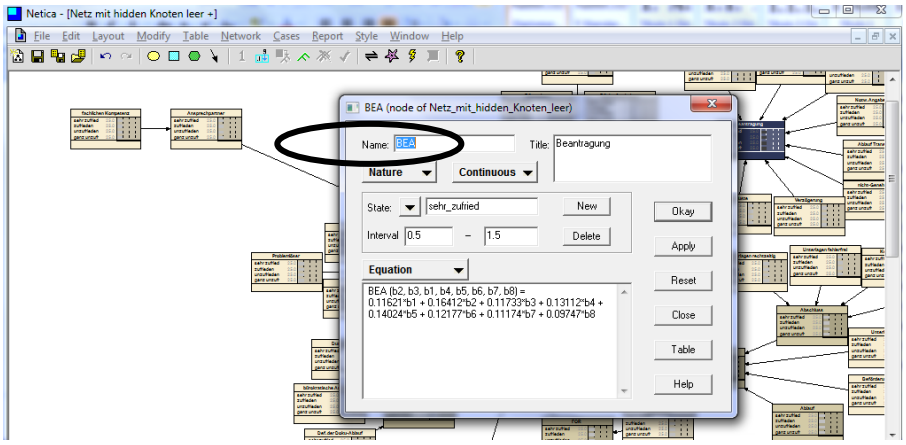


Figure 26: Node's settings

Moreover, Netica also supports data with missing values. It deals with them by estimating a possible value for that missing value when it has the necessary information to do so. In our networks, there exist equations in order to perform this step. In case there are not equations, Netica uniformly distributes probabilities through the possible states of this particular case. When there are cases in which something is unknown, the procedure is the same.

3.2.4.3. Net without priors

In this section, using the subnet DUR (*travel application*) we illustrate the learning behavior without prior probabilities. Figure 27 shows the results:

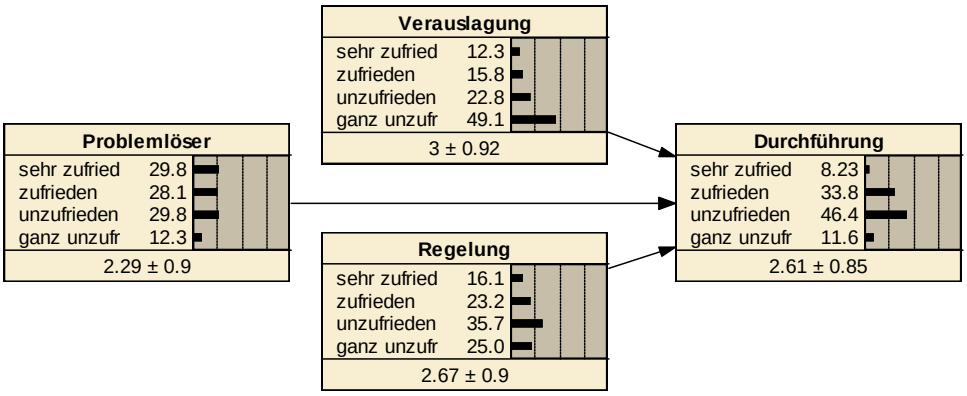


Figure 27: Learned subnet DUR without prior probabilities

3.2.4.4. Net with priors

We apply three different prior probabilities in order to do the learning for the subnet DUR (*travel application*):

- Uniform probabilities.
- Distribution tending to “sehr zufrieden” (very satisfied).
- Distribution tending to “überhaupt nicht zufrieden” (not satisfied at all)

Before showing the graphs corresponding to each of the three cases, we show the shares of each state for the nodes DUR1, DUR2 and DUR3 for every prior in Table 14.

NODE DUR1			
States	Uniform distribution	Distribution tending to very satisfied	Distribution tending to not satisfied at all
1 (very satisfied)	25%	50%	2,275%
2 (satisfied)	25%	34,134%	13,591%
3 (not satisfied)	25%	13,591%	34,134%
4 (not satisfied at all)	25%	2,275%	50%

Table 14: Prior distributions

These distributions are applied to every root node from the network, associating them an experience of 10 cases. Figures 28, 29 and 30 present the updated subnet once the learning procedure has been done:

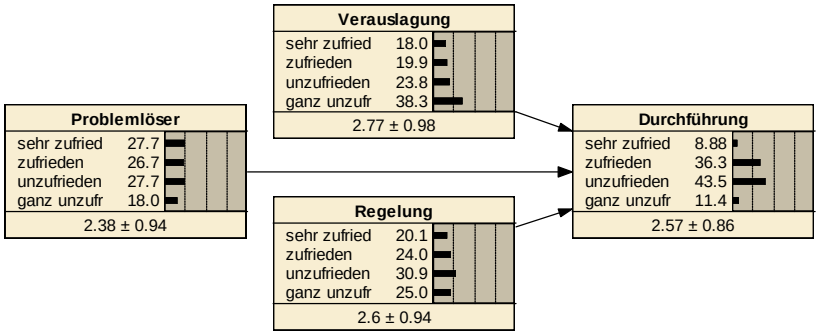


Figure 28:Learned subnet DUR with uniform priors

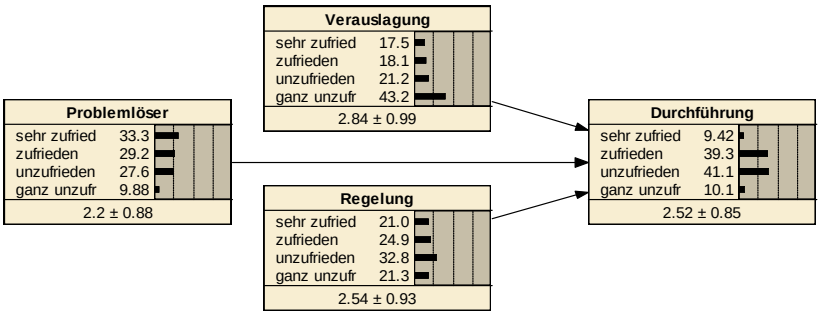


Figure 29: Learned subnet DUR with prior probabilities tending to "very satisfied"

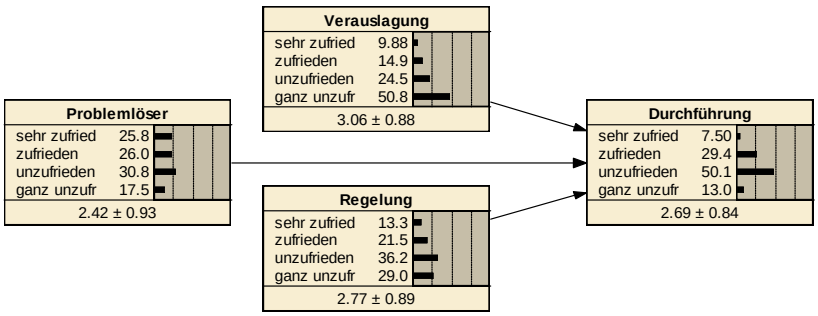


Figure 30: Learned subnet DUR with prior probabilities tending to "not satisfied at all"

At first glance we can see that our target node's behavior does not significantly vary across any of the four presented cases. These results are robust when changing the subnet of study into any other subnet or the total network. Therefore, as there is not any significant change when using prior probabilities, from now on we work with the net without priors. Moreover the priors are based in assumptions and not in real facts so the net without prior probabilities represent a more accurate version of the real world. Table 15 shows the variation between the case without priors and the cases with them.

DUR subnet without priors		DUR subnet with uniform priors			DUR subnet with priors tending to "very satisfied"			DUR subnet with priors tending to "not satisfied at all"		
State	**Belief**	**State**	**belief**	**Variation**	**State**	**belief**	**Variation**	**State**	**belief**	**Variation**
1	8%	**1**	9%	0.7%	**1**	9%	1.2%	**1**	8%	-0.7%
2	34%	**2**	36%	2.5%	**2**	39%	5.5%	**2**	29%	-4.4%
3	46%	**3**	44%	-2.9%	**3**	41%	-5.3%	**3**	50%	3.7%
4	12%	**4**	11%	-0.2%	**4**	10%	-1.5%	**4**	13%	1.4%
Satisfied	42%	**Satisfied**	45%	3%	**Satisfied**	49%	7%	**Satisfied**	37%	-5%
Not satisfied	58%	**Not satisfied**	55%	-3%	**Not satisfied**	51%	-7%	**Not satisfied**	63%	5%

Table 15: Distribution with priors vs without priors for node DUR

3.2.5. Learning phase

In this section we work simultaneously with the two final nets (the one with latent variables and the one without them). According to section 2.1.5, we apply the counting algorithm to the network without latent variables and the expectation-maximization (from now on EM) one to the net with latent variables.

3.2.5.1. Integration of data.

In this section we analyze the network once the data collected from the survey has been learnt. The analysis has several phases. The first stage is to know how to introduce the data into the net. In our case, in order to be consistent, we use the software Netica again but there are many others such as "HUGIN EXPERT" or "WinBUGS" that can do the same as well. In the second stage we justify the implementation of some equations in the nodes ZU, KOM,

BEA, ORG, DUR, ABR and UK as well as in the five hidden nodes from the second network (Abl, ABS, VOR, FOR and PRU). We do so using some examples from our principal networks. The next step consists on a brief analysis of the differences in the final distributions between our two main nets after the implementation of their particular learning algorithms. Finally we introduce some new possible cases and we interpret how the net behaves when we do it.

The integration of data into Netica differs from one net to the other. In the first network data must be introduced by clicking on "Case→Incorporate Case File" and selecting the Excel file that contains the answers from the survey. Then, the learning from data can be done by clicking in "Network→Compile". For our second network, the one with latent variables, we use the EM algorithm. Therefore the path chosen is "Case→Learn Using EM". For both procedures, Netica asks us if we want to delete the probabilities from the CPT of each node. We do not delete them because we study the network's behavior both with these prior probabilities and without them. Moreover, in the nodes containing equations, their CPTs are calculated with the methodology explained in section 2.1.2. Figures 31 to 36 illustrate these steps explained a few lines above.

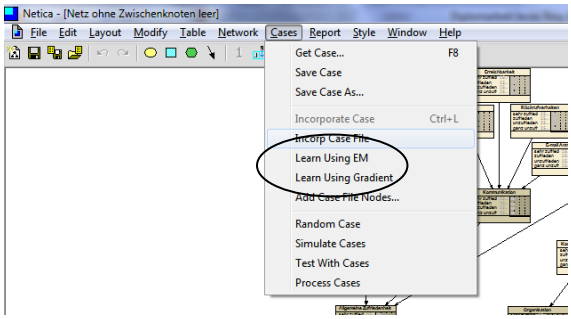


Figure 31: Incorporate Case File Option

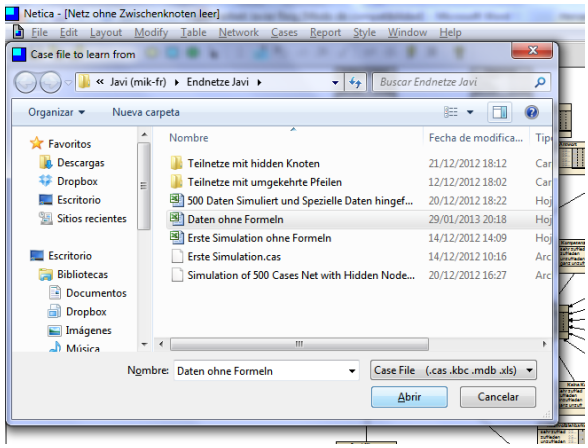


Figure 32: Selection of the file with the data

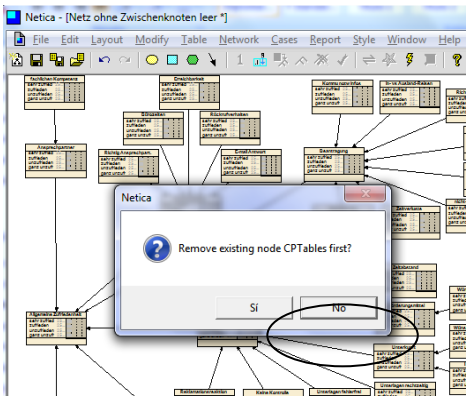


Figure 33: Rebuild CPT option

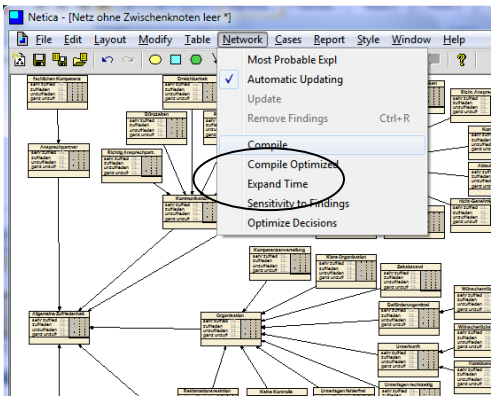


Figure 34: Compile bottom for learning from data

More specifically, figures 35 and 36 show the procedure followed when dealing with the net containing latent variables. In figure 36 we can see that the number of iterations of the EM algorithm rises up to 53 but the network's improve from the sixth iteration until the end is never more than 0,005%.

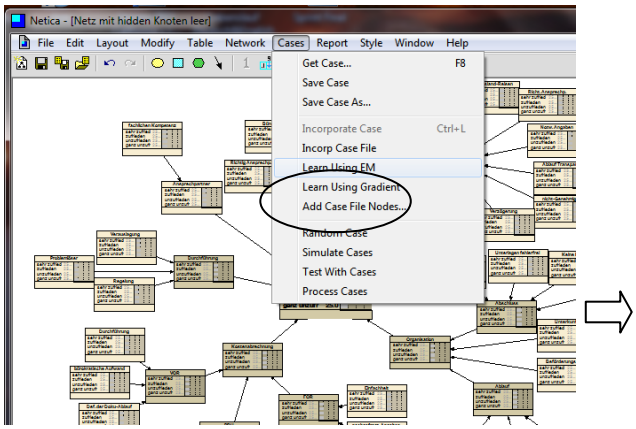


Figure 35: Learning using EM algorithm

Iteration	Log Likelihood	Change %
0	54.5602	
1	44.865	17.7698
2	44.2093	1.4613
3	44.1109	0.2226
4	44.0897	0.0480
5	44.0835	0.0141
6	44.081	0.0057
7	44.0796	0.0031
8	44.0787	0.0022
9	44.0779	0.0017
10	44.0773	0.0014
11	44.0768	0.0012
12	44.0764	0.0010
13	44.076	0.0009
14	44.0756	0.0008
15	44.0753	0.0007
16	44.0751	0.0006
17	44.0748	0.0005
18	44.0747	0.0004
19	44.0745	0.0004
20	44.0743	0.0003
21	44.0742	0.0003
22	44.0741	0.0002
23	44.074	0.0002
24	44.0739	0.0002
25	44.0739	0.0002
26	44.0738	0.0001
27	44.0738	0.0001
28	44.0737	0.0001
29	44.0737	0.0001
30	44.0737	0.0001
31	44.0736	0.0001
32	44.0736	0.0001
33	44.0736	0.0000
34	44.0736	0.0000
35	44.0736	0.0000
36	44.0736	0.0000
37	44.0735	0.0000
38	44.0735	0.0000
39	44.0735	0.0000
40	44.0735	0.0000
41	44.0735	0.0000
42	44.0735	0.0000
43	44.0735	0.0000
44	44.0735	0.0000
45	44.0735	0.0000
46	44.0735	0.0000
47	44.0735	0.0000
48	44.0735	0.0000
49	44.0735	0.0000
50	44.0735	0.0000
51	44.0735	0.0000
52	44.0735	0.0000
53	44.0735	0.0000

Figure 36: Iterations for EM algorithm

The second phase on the analysis of the network consists on justifying the use of equations in the son nodes from every subnet. In order to do so we focus on the net without hidden nodes. We compare the network without any equation with the one containing the previously described equations. Figures 37 and 38 show the results of the learning for both cases respectively. We can observe that our target node distribution (ZU) is totally different from one figure to the other. In case the net containing equations, the states “satisfied” and “not satisfied” gather the 85,5% of the probability distribution while only the 14,5% is concentrated in the extreme cases (“very satisfied” and “not satisfied at all”). If we split the sample into being satisfied (“satisfied” and “very satisfied”) and not satisfied (“not satisfied” and “not satisfied at all”), in the 71,57% of the cases the worker is satisfied. In the net without equations the same node is almost uniformly distributed: 24,9% “very satisfied”; 25,1% “satisfied”; 25,0% “not satisfied”; 24,9% “not satisfied at all”.

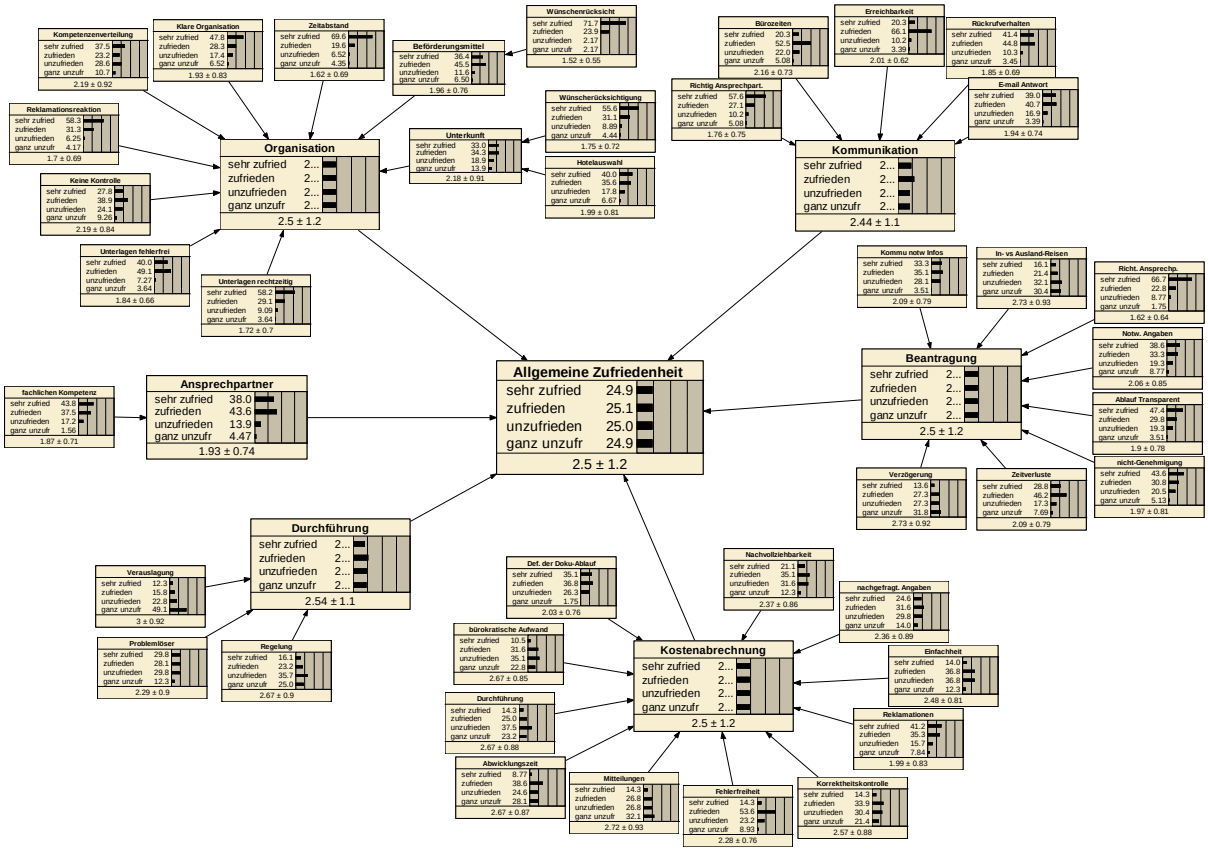


Figure 37: Learned net without equations

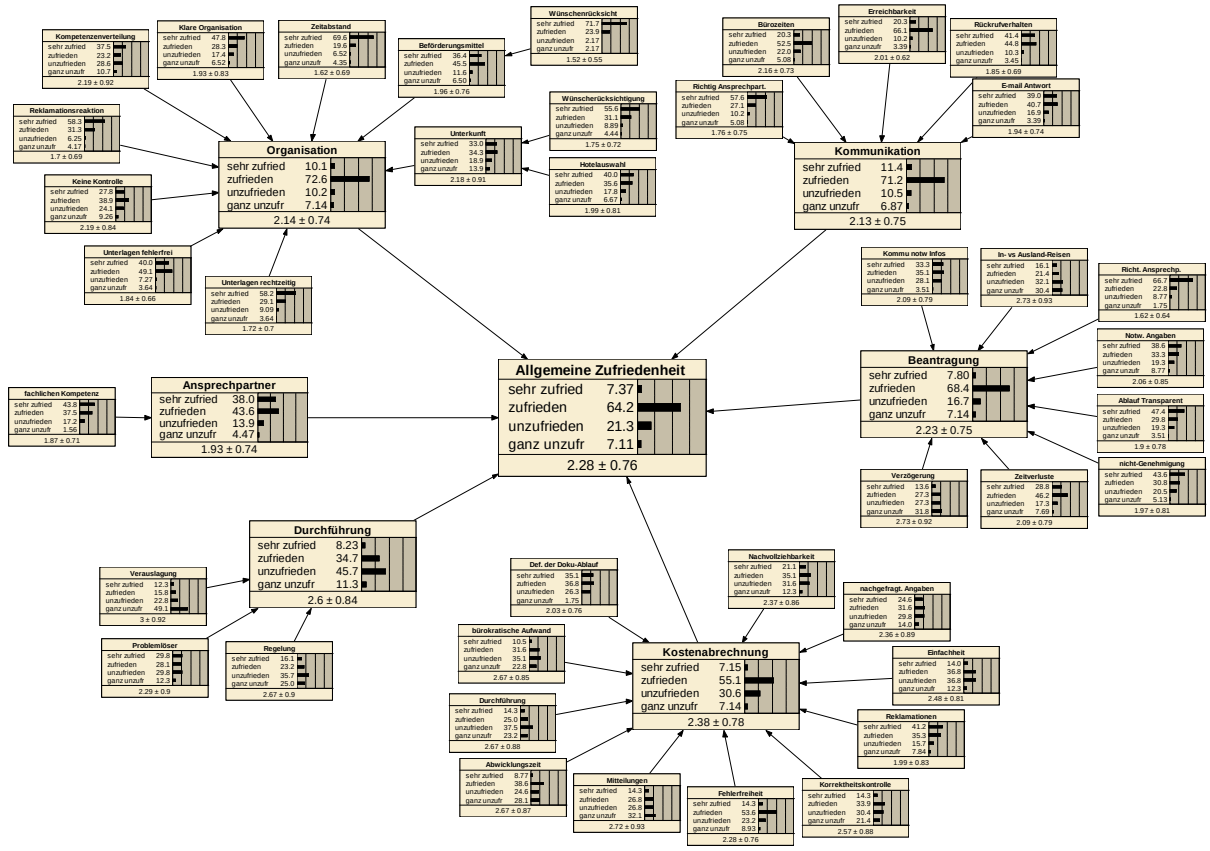


Figure 38: Learned net with equations

Furthermore, if we perform an analysis what if with both nets we can see that, for the case where we know if a certain person is satisfied with all the sub-processes from the survey (i.e. she answers either 1 or 2), the probabilities distribution of the target node ZU is still uniform. However, when doing the same in the net with equations, the probabilities distribution of ZU slightly vary from a satisfied believe of 71,57% to one of 84,68%. Figures 39 and 40 illustrate both situations.

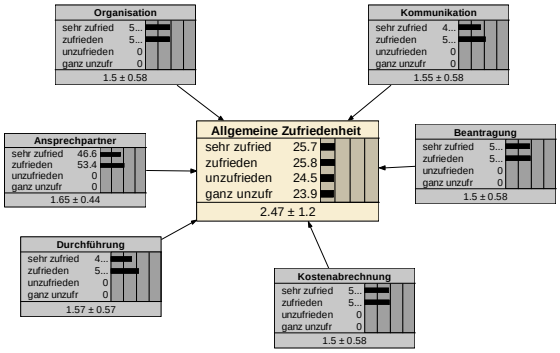


Figure 39: ZU node's distributions with findings on its father nodes and without equations.

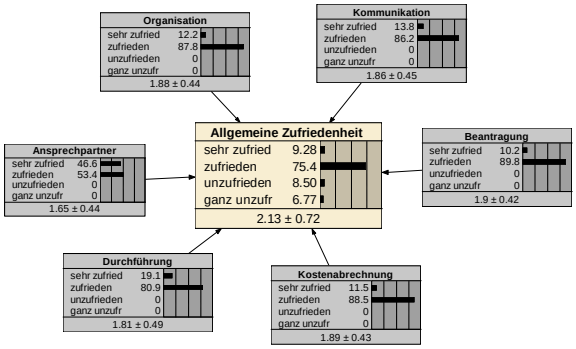


Figure 40: ZU node's distributions with findings on its father nodes and with equations.

Obviously, the equations are very useful for us because we have CPTs in the nodes of the biggest subnets (BEA, ABR and ORG). Therefore, in absence of equations, the probabilities would be allocated uniformly through the four possible states due to the small sample size. However, when introducing a simple equation, we help Netica to get closer to a logical reasoning. For instance, let suppose that a respondent assigns a 2 ("satisfied") to all the questions of a particular subsection or subnet, i.e., DUR1, DUR2 and DUR3 present a

value equal to 2. Therefore it would seem reasonable for the node DUR to present a value also equal to 2. However, the net without equations make the node DUR to present a uniform distribution of conditional probabilities for the four states because this example is not part of our excel file. Figure 41 illustrates this example.

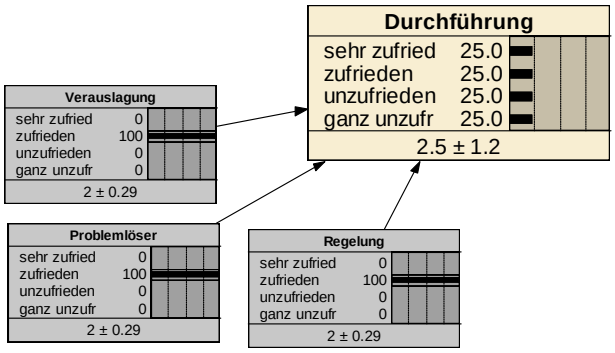
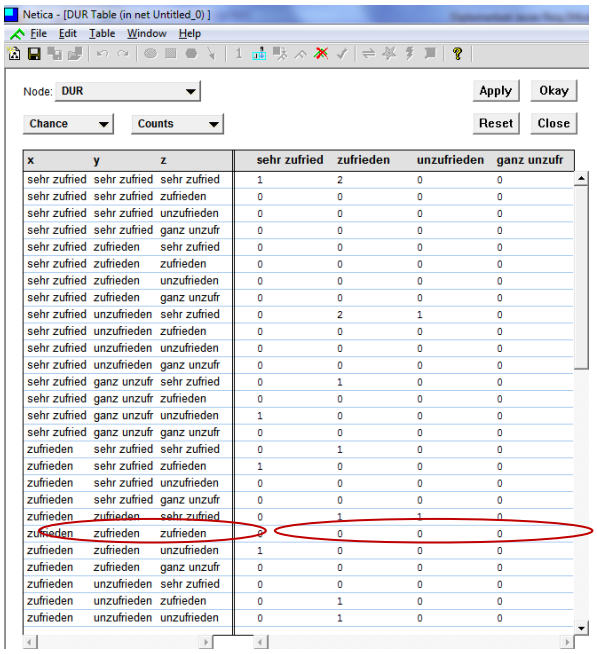


Figure 41: Subnet DUR without equation and with findings in roots nodes



x	y	z	sehr zufried	zufrieden	unzufrieden	ganz unzufr
sehr zufried	sehr zufried	sehr zufried	1	2	0	0
sehr zufried	sehr zufried	zufrieden	0	0	0	0
sehr zufried	sehr zufried	unzufrieden	0	0	0	0
sehr zufried	sehr zufried	ganz unzufr	0	0	0	0
sehr zufried	zufrieden	sehr zufried	0	0	0	0
sehr zufried	zufrieden	zufrieden	0	0	0	0
sehr zufried	zufrieden	unzufrieden	0	0	0	0
sehr zufried	zufrieden	ganz unzufr	0	0	0	0
sehr zufried	unzufrieden	sehr zufried	0	2	1	0
sehr zufried	unzufrieden	zufrieden	0	0	0	0
sehr zufried	unzufrieden	unzufrieden	0	0	0	0
sehr zufried	unzufrieden	ganz unzufr	0	0	0	0
sehr zufried	ganz unzufr	sehr zufried	0	1	0	0
sehr zufried	ganz unzufr	zufrieden	0	0	0	0
sehr zufried	ganz unzufr	unzufrieden	1	0	0	0
sehr zufried	ganz unzufr	ganz unzufr	0	0	0	0
zufrieden	sehr zufried	sehr zufried	0	1	0	0
zufrieden	sehr zufried	zufrieden	1	0	0	0
zufrieden	sehr zufried	unzufrieden	0	0	0	0
zufrieden	sehr zufried	ganz unzufr	0	0	0	0
zufrieden	zufrieden	sehr zufried	0	1	1	0
zufrieden	zufrieden	zufrieden	0	0	0	0
zufrieden	zufrieden	unzufrieden	1	0	0	0
zufrieden	zufrieden	ganz unzufr	0	0	0	0
zufrieden	unzufrieden	sehr zufried	0	0	0	0
zufrieden	unzufrieden	zufrieden	0	1	0	0
zufrieden	unzufrieden	unzufrieden	0	1	0	0

Figure 42: Counts table of node DUR

As mentioned before, equations help us to get closer to reality (or what it would be). Figure 43, corresponding to the travel execution subnet, shows the probabilities distribution when applying equations. We observe that the 78,6% of the probabilities correspond to the state 2 (“satisfied”) while the probability of being in any other state is 7,14%for each of

them. Due to the lack of data, Netica adds additional uncertainty and this is the explanation of the 7,14%.

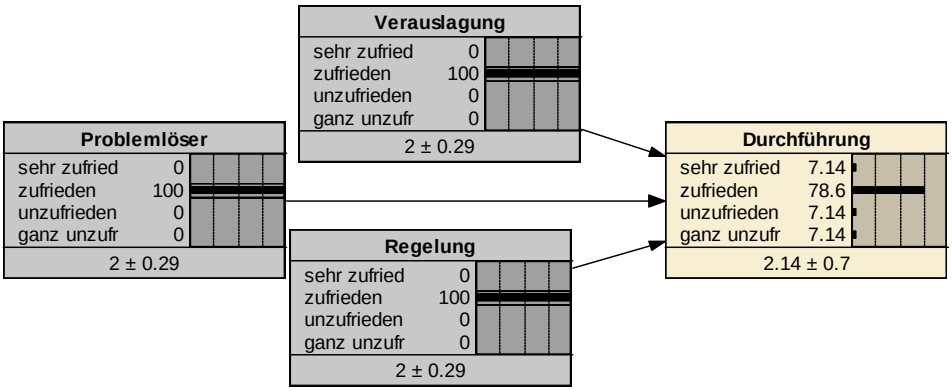


Figure 43: Subnet DUR with equation and findings in root nodes

There is an additional problem which is the tables' size, as we have explained a few lines above. The solution to this problem lies on the inclusion of latent variables. Figure 44 shows the net with latent variables and the corresponding learning. Figure 38 illustrates the differences between this network and the one without latent variables.

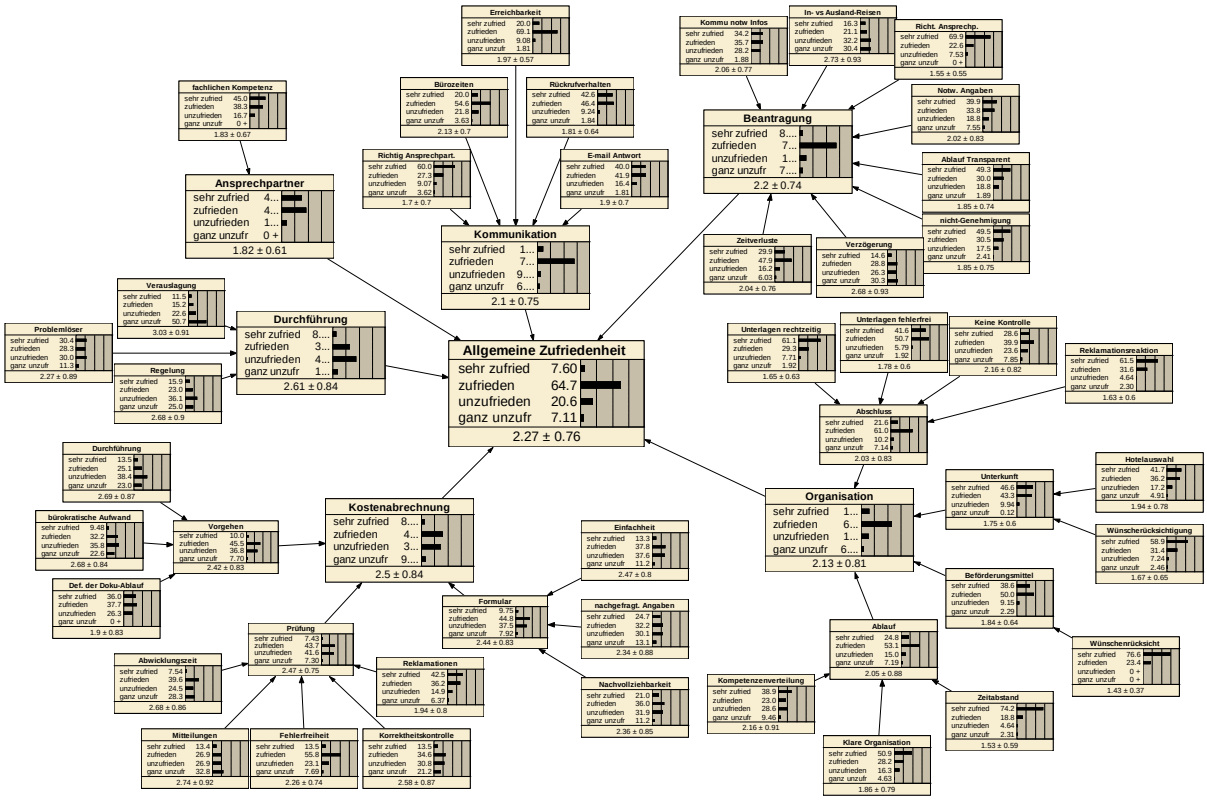


Figure 44: Learned net with equations and hidden nodes

At first sight, it seems not to be significant changes because the distributions in the nodes are still approximately the same (as long as no finding is introduced into the net). If we analyze how both nodes behave when implementing new cases, this behavior is again the same. For instance, let all the root nodes to take the values 1 or 2 (satisfied or very satisfied). Then, the distributions of both nets are almost identical in those nodes without findings. This example is illustrated in figures 45 and 46.

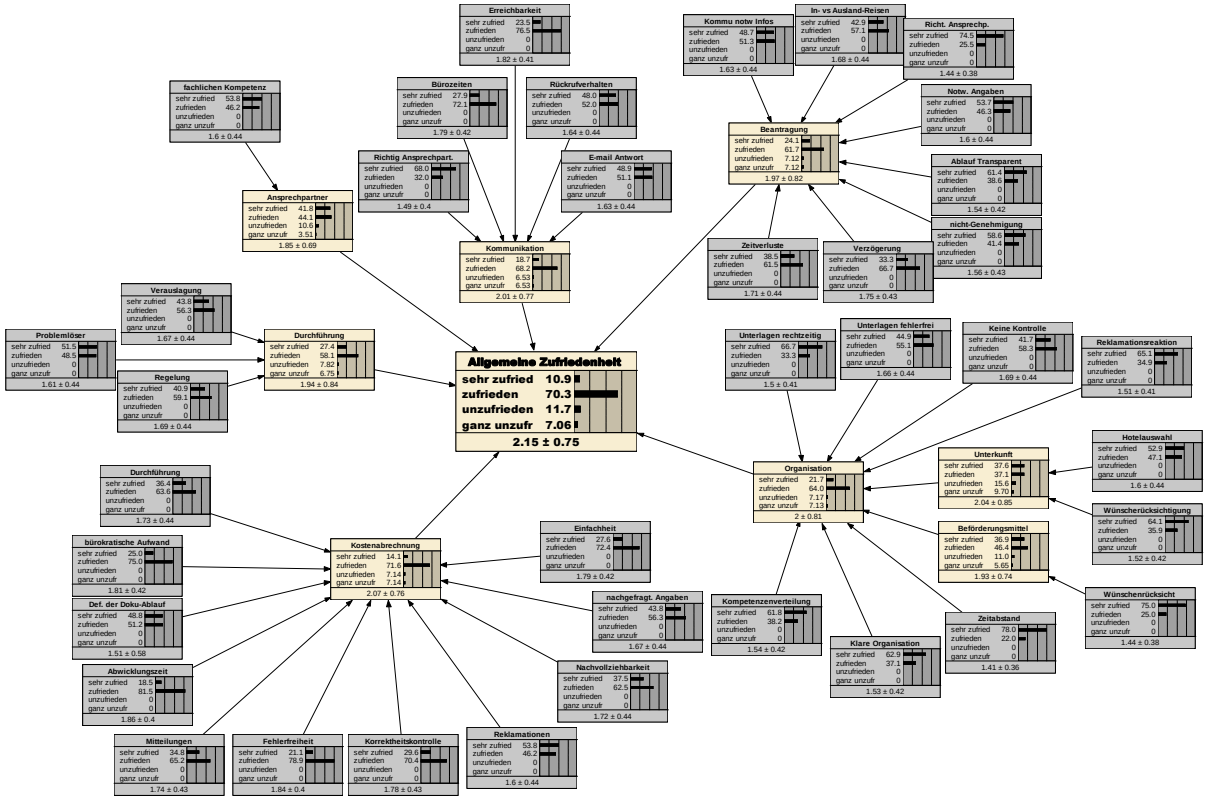


Figure 45: Net without latent variables and findings in root nodes.

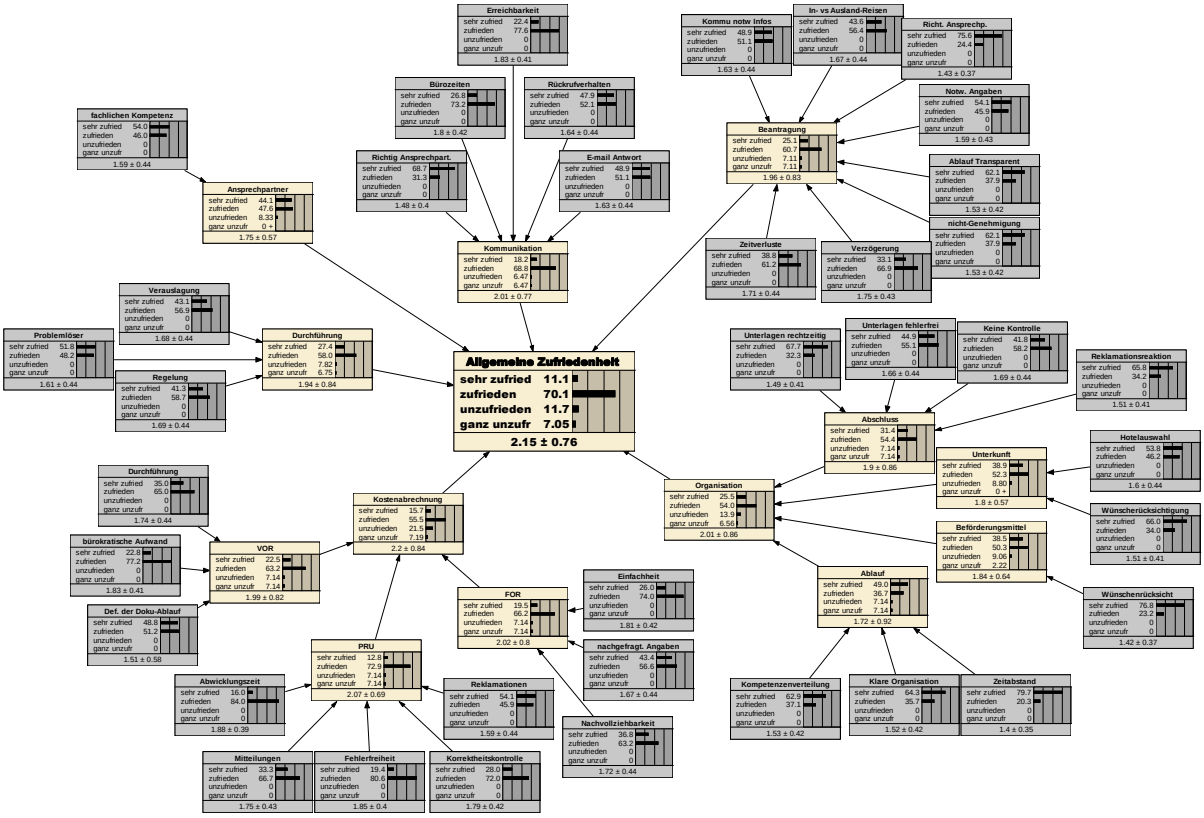


Figure 46: Net with latent variables and findings in root nodes

Thus, the difference between both nets does not lie on the net's behavior but in the necessary amount of data for constructing the tables. We can easily assume that both networks are equally valid in terms of results. However, if we try to work with them, we would rapidly see that the net without latent variables needs 5,5 seconds to update the probabilities when entering a finding in any of the nodes. Moreover, in case of entering the findings for all the root nodes, the inference lasts 59 seconds. These times are reduced to fractions of a second in the first case and to 3 seconds in the latter when introducing the 5 hidden nodes into the net. These reductions are explained by the amount of data these networks work with: introducing hidden nodes the number of tables increases an 11% but the number of entries within them decreases from 17.322.488 values to 290.296. It represents a 98% reduction.

3.2.6. The most promising approach

It seems very intuitive after reading the project that the most promising approach corresponds to the second network which is the one with latent variables or hidden nodes. We present it as the most promising approach because, although both networks work well in terms of probabilistic inference and learning parameters, the net with latent variables needs a smaller amount of memory (the network without latent variables occupies 3.443 KB while the other one occupies only 363 KB, an 89% less). In addition, as mentioned in the previous section, the actualization times are also smaller.

Besides, if we introduced five more questions, each one corresponding to one of the hidden nodes, these would stop being latent variables and we could be able to apply the counting algorithm to do the learning from parameters. This learning procedure is more precise and simple than the EM one because it does not have to go through the expectation phase where an expectative value is calculated for every latent variable (this phase is performed for every case within the data file).

However, there exists a problem regarding the election of this network, which is the increase in the number of questions of the survey from 47 to 52. Therefore, it may be difficult to find employees willing to answer it in an accurate way.

4. Conclusions

4.1. Further steps and open research questions

This project tries to build the foundations for further projects. However, new studies on customer satisfaction on administrative processes should correct some come of our problems such as the sample size, the survey structure or extension or the size of the conditional probabilities tables. Throughout the project, we try to address this last issue and we finally reduce the tables' size but only for the organization and settlement of accounting subnets. Besides, one of the presented solutions, the inclusion of latent variables, implies using a slower and more complicated algorithm which assumes some estimated values for the nodes. Another possible solution to the sample and tables size would be to reduce the amount of question in the survey. As the survey design and the subsequent data analysis were performed separately, the survey presents some characteristic difficult to deal with in our empirical analysis. Therefore, a joint design of the survey and network could be helpful for further studies.

It is also important to observe that, when analyzing vertically the survey, the questions associated to nodes BEA6 and BEA7 were not answered for a huge amount of people. Therefore, it would be interesting to know the reason why this happen. It is because the respondents do not understand the question? It is because the worker does not have experience enough to answer it? Thus, we need to know the answers to these questions in order to remove that part from the survey or not.

Finally, in future research, it would be interesting to study whether there are cases in which the overall satisfaction in a sub-process does not correspond to the satisfaction showed in the individual questions of that particular sub-process. In other words, it can be the case that a person answer is equal to 1 for the node DUR but its answer is 4 for the nodes DUR1, DUR2 and DUR3. Therefore, we can interpret that if the answers are that different, it might be the case that we are missing something important regarding the questions of the sub-process DUR.

5. Annex

Idnum	ZU	BEA	ORG	DUR	ABR
1	2	2	1	1	3
2	2	2	2	3	2
3	2	1	2	3	2
4	3	1	2	1	1
5	2	2	2	3	2
6	4	3	3	4	4
7	2	2	2	3	3
8	2	2	2	2	2
9	1	1	1	2	1
10	2	*	*	*	*
11	2	2	3	2	3
12	2	2	1	2	2
13	3	3	3	4	4
14	1	1	1	1	1
15	2	1	3	4	1
16	*	*	*	*	*
17	3	3	1	1	2
18	3	2	2	2	2
19	2	1	1	3	2
20	2	2	2	3	3
21	1	1	1	2	1
22	4	2	3	2	4
23	3	2	*	2	1
24	2	1	1	2	3
25	2	1	2	2	3
26	3	1	2	2	4
27	2	2	3	2	3
28	2	2	1	2	3
29	3	1	2	2	4
30	2	1	1	2	1
31	1	1	1	2	2
32	1	1	1	1	1
33	3	2	2	3	3
34	*	*	*	*	*
35	2	1	1	2	3
36	4	2	2	3	4
37	2	1	2	3	2
38	3	1	1	2	4
39	4	1	1	2	4
40	2	2	2	2	4
41	2	2	1	2	3
42	3	2	*	4	3
43	3	1	1	4	4
44	2	2	*	2	2
45	4	2	*	3	4
46	2	1	2	3	2
47	2	1	1	2	1
48	2	1	1	3	2
49	3	1	3	3	2
50	3	2	2	3	3
51	3	2	3	4	2
52	1	1	1	2	1
53	3	2	2	3	3
54	4	2	2	3	4
55	2	2	3	2	3
56	1	1	1	2	2
57	2	2	3	2	3
58	3	2	2	3	4
59	3	3	2	2	3
60	3	1	2	4	3
61	2	2	2	2	2
62	*	*	*	*	*
63	*	*	*	*	*
64	2	3	3	4	4
65	2	1	3	3	3
66	3	1	1	1	4

Table 16: Overall satisfaction and principal subsections dataset

Idnum	ANS	ANS1
1	1	2
2	*	2
3	1	1
4	1	1
5	2	1
6	2	2
7	2	2
8	2	2
9	1	1
10	*	*
11	2	2
12	2	2
13	3	3
14	1	1
15	1	3
16	*	*
17	1	1
18	1	1
19	1	1
20	2	2
21	2	2
22	3	2
23	2	3
24	1	1
25	2	3
26	2	2
27	2	2
28	2	1
29	1	1
30	1	1
31	1	1
32	1	1
33	1	3
34	*	*
35	1	1
36	1	1
37	2	1
38	*	*
39	2	3
40	2	2
41	2	2
42	3	2
43	3	3
44	1	1
45	3	2
46	2	2
47	2	1
48	1	1
49	1	1
50	2	2
51	2	1
52	1	1
53	2	3
54	2	2
55	3	2
56	1	1
57	3	3
58	2	2
59	2	2
60	1	1
61	2	1
62	*	*
63	*	*
64	2	3
65	2	2
66	1	1

Table 17: *Contact person* dataset

Idnum	KOM	KOM1	KOM2	KOM3	KOM4	KOM5
1	*	1	2	3	2	3
2	*	2	2	2	1	3
3	*	*	*	*	*	*
4	1	1	3	2	2	3
5	2	1	1	1	1	1
6	3	4	3	2	2	2
7	2	2	2	2	2	2
8	2	1	2	2	2	2
9	1	2	1	1	2	1
10	*	*	*	*	*	*
11	2	2	2	1	2	1
12	2	1	2	2	1	2
13	2	2	2	2	2	2
14	1	1	1	2	1	1
15	*	*	*	*	*	*
16	*	*	*	*	*	*
17	2	2	*	*	*	*
18	2	1	2	2	1	1
19	2	1	2	2	3	3
20	*	1	3	2	1	2
21	1	2	2	2	2	2
22	*	*	3	2	2	2
23	2	3	3	2	3	2
24	1	1	2	2	1	1
25	2	2	1	2	2	2
26	2	1	3	2	1	1
27	2	1	2	2	*	2
28	2	1	2	2	2	2
29	3	1	4	4	2	2
30	2	1	2	2	2	1
31	1	1	2	1	1	1
32	1	1	2	2	1	1
33	3	3	2	2	3	4
34	*	*	*	*	*	*
35	1	1	1	2	1	1
36	2	3	3	2	1	1
37	2	1	2	1	1	1
38	*	*	*	*	*	*
39	*	*	*	*	*	*
40	*	*	*	*	*	*
41	2	2	2	3	3	3
42	2	2	2	2	2	2
43	2	1	3	2	1	1
44	2	1	2	3	2	3
45	3	1	3	3	4	2
46	2	2	3	2	2	2
47	2	1	1	2	1	1
48	1	1	1	1	1	1
49	2	1	2	2	1	3
50	1	1	1	2	2	2
51	2	1	2	2	2	3
52	1	1	1	1	1	1
53	3	3	3	2	2	2
54	2	2	2	2	2	2
55	3	3	2	2	2	2
56	1	1	1	1	1	1
57	3	4	4	3	3	3
58	2	1	2	2	1	1
59	2	2	2	1	2	2
60	2	1	2	2	1	1
61	2	2	2	2	2	2
62	*	*	*	*	*	*
63	*	*	*	*	*	*
64	2	1	2	2	2	2
65	2	2	3	1	1	1
66	1	1	1	1	1	1

Table 18:*Communication* dataset

Idnum	BEA1	BEA2	BEA3	BEA4	BEA5	BEA6	BEA7	BEA8
1	3	3	2	1	1	2	2	3
2	3	4	2	2	3	2	2	1
3	*	*	*	*	*	*	*	*
4	1	2	1	2	1	*	*	*
5	3	4	1	4	2	3	3	3
6	3	2	3	2	3	3	3	2
7	3	4	2	2	2	2	3	2
8	1	3	1	1	1	*	*	*
9	2	2	1	2	1	1	1	1
10	*	*	*	*	*	*	*	*
11	1	3	1	1	1	1	3	2
12	2	2	1	1	1	1	1	1
13	2	4	2	3	2	2	2	1
14	1	1	1	1	1	*	*	*
15	*	*	*	*	*	*	*	*
16	*	*	*	*	*	*	*	*
17	*	*	*	*	*	*	*	*
18	1	2	1	3	1	1	2	3
19	1	3	2	3	2	2	4	2
20	3	3	1	1	3	*	*	2
21	1	1	1	2	1	*	*	*
22	2	1	1	1	2	2	*	*
23	1	3	1	2	1	1	4	2
24	2	3	1	2	2	2	2	2
25	4	3	2	2	1	1	3	2
26	2	1	1	1	1	1	2	1
27	3	4	1	1	1	*	4	3
28	2	4	1	2	2	*	*	1
29	3	4	1	4	1	1	1	1
30	1	2	1	1	1	1	2	1
31	1	2	1	2	1	*	4	2
32	2	3	2	3	2	*	3	2
33	2	4	1	2	1	3	3	2
34	*	*	*	*	*	*	*	*
35	2	3	1	3	2	1	2	1
36	*	*	*	*	*	*	*	*
37	2	4	1	1	3	3	3	3
38	*	*	*	*	*	*	*	*
39	*	*	*	*	*	*	*	*
40	*	*	*	*	*	*	*	*
41	2	3	1	1	3	*	4	4
42	2	2	2	3	3	2	3	2
43	1	4	1	1	2	*	*	1
44	2	3	1	1	1	*	*	2
45	1	3	1	2	2	1	4	2
46	1	2	1	2	1	*	*	3
47	1	*	1	2	1	*	*	1
48	*	*	*	*	*	*	*	*
49	3	4	3	4	2	4	4	3
50	1	4	1	2	2	3	4	2
51	1	1	1	1	2	3	4	3
52	1	1	1	1	1	1	2	2
53	3	4	2	3	3	2	4	4
54	2	2	1	1	1	1	3	2
55	3	3	3	3	3	1	2	2
56	1	1	1	1	1	1	1	1
57	2	3	1	1	2	2	4	2
58	2	4	2	1	1	1	1	1
59	3	3	1	2	1	*	3	2
60	2	3	1	1	2	*	*	2
61	3	4	2	3	1	*	*	2
62	*	*	*	*	*	*	*	*
63	*	*	*	*	*	*	*	*
64	3	4	3	4	3	3	4	4
65	2	1	2	2	3	2	2	2
66	3	2	1	3	4	*	4	1

Table 19: *Travel application* dataset

Idnum	Abl1	Abl2	Abl3	BF	BF1	UK	UK1	UK2
1	2	*	*	*	*	*	*	*
2	*	2	1	2	1	2	2	2
3	*	*	*	*	*	*	*	*
4	1	*	*	*	*	*	*	*
5	3	1	1	1	1	1	1	2
6	1	2	1	2	1	2	1	1
7	1	1	1	2	1	2	1	1
8	3	2	1	1	1	1	1	1
9	1	1	1	2	2	3	1	2
10	*	*	*	*	*	*	*	*
11	2	2	1	2	1	2	2	2
12	2	1	1	1	1	1	1	1
13	4	3	2	2	1	2	3	3
14	1	*	*	*	*	*	*	*
15	*	*	*	*	*	*	*	*
16	*	*	*	*	*	*	*	*
17	*	*	*	*	*	*	*	*
18	1	2	1	2	1	2	1	1
19	2	1	1	2	1	2	2	3
20	1	1	1	1	1	1	1	1
21	1	*	*	*	*	*	*	*
22	4	*	*	*	*	*	*	*
23	1	3	4	3	2	3	3	2
24	2	2	1	2	1	2	1	1
25	3	1	2	1	1	1	1	1
26	4	1	1	2	1	2	1	2
27	2	3	2	2	2	3	2	2
28	2	1	1	1	1	2	2	2
29	3	2	1	1	1	1	3	3
30	2	2	1	2	1	1	1	1
31	1	1	1	1	1	1	1	1
32	1	1	1	1	1	1	1	1
33	3	*	*	*	*	*	*	*
34	*	*	*	*	*	*	*	*
35	1	1	1	1	1	2	1	2
36	*	*	*	*	*	*	*	*
37	3	3	2	2	1	1	2	3
38	*	*	*	*	*	*	*	*
39	*	*	*	*	*	*	*	*
40	*	*	*	*	*	*	*	*
41	2	*	*	*	*	*	*	*
42	3	4	2	2	2	4	4	4
43	1	1	1	1	1	1	1	1
44	3	*	*	*	*	*	*	*
45	4	*	*	*	*	*	*	*
46	1	1	1	2	1	3	2	2
47	1	*	*	*	*	*	*	*
48	*	*	*	*	*	*	*	*
49	3	4	1	3	2	3	2	3
50	3	1	1	2	1	1	1	2
51	2	2	3	2	1	2	1	2
52	1	1	2	1	2	1	1	1
53	2	3	3	3	2	2	2	3
54	2	2	2	2	1	2	2	3
55	4	3	1	2	1	2	1	2
56	1	1	1	1	1	1	1	1
57	1	1	1	4	2	2	2	4
58	1	1	1	1	1	1	1	1
59	3	2	1	2	2	1	1	2
60	3	1	1	2	1	*	*	*
61	3	2	2	1	1	2	2	2
62	*	*	*	*	*	*	*	*
63	*	*	*	*	*	*	*	*
64	3	*	*	*	*	*	*	*
65	3	3	1	3	2	2	2	1
66	1	1	1	1	1	1	1	1

Table 20: *Organization* dataset 1

Idnum	ABS1	ABS2	ABS3	ABS4
1	2	2	2	2
2	2	2	3	*
3	*	*	*	*
4	1	1	1	*
5	1	1	2	1
6	1	2	2	1
7	1	1	1	1
8	*	*	*	*
9	1	1	1	1
10	*	*	*	*
11	2	2	2	2
12	1	1	2	1
13	2	2	2	2
14	*	*	*	*
15	*	*	*	*
16	*	*	*	*
17	*	*	*	*
18	3	2	1	1
19	1	1	2	2
20	1	1	1	*
21	1	2	2	1
22	2	2	3	2
23	2	2	2	1
24	1	2	2	1
25	2	2	3	2
26	1	3	4	1
27	2	3	3	2
28	1	1	2	1
29	1	1	1	1
30	1	1	3	1
31	1	1	2	1
32	1	2	3	1
33	4	2	3	2
34	*	*	*	*
35	1	1	4	*
36	*	*	*	*
37	1	1	1	1
38	*	*	*	*
39	*	*	*	*
40	*	*	*	*
41	2	2	3	2
42	2	2	2	2
43	1	1	1	1
44	2	2	*	*
45	1	2	2	3
46	1	2	3	1
47	1	1	1	*
48	*	*	*	*
49	1	1	1	*
50	1	2	4	1
51	3	2	3	1
52	1	1	2	2
53	3	3	3	3
54	2	1	2	2
55	1	2	2	1
56	1	1	1	1
57	2	2	2	4
58	1	2	3	1
59	2	2	1	1
60	1	1	2	1
61	1	2	2	1
62	*	*	*	*
63	*	*	*	*
64	3	4	4	2
65	2	2	1	2
66	1	1	1	1

Table 21: *Organization* dataset 2

Idnum	DUR1	DUR2	DUR3
1	3	2	2
2	3	1	4
3	*	*	*
4	2	3	2
5	4	4	4
6	3	3	4
7	2	3	4
8	3	1	2
9	2	2	4
10	*	*	*
11	1	1	3
12	1	1	1
13	4	3	4
14	1	1	1
15	*	*	*
16	*	*	*
17	*	*	*
18	2	1	4
19	2	3	4
20	4	2	4
21	2	1	2
22	4	3	4
23	3	4	4
24	2	3	3
25	1	1	1
26	3	1	4
27	4	2	4
28	3	2	2
29	4	4	1
30	1	1	3
31	3	2	3
32	2	2	1
33	4	2	4
34	*	*	*
35	2	2	3
36	*	*	*
37	4	1	3
38	*	*	*
39	*	*	*
40	*	*	*
41	3	3	3
42	3	2	4
43	3	1	4
44	3	3	3
45	4	2	4
46	3	2	3
47	*	3	4
48	*	*	*
49	4	2	4
50	3	3	4
51	3	3	4
52	2	1	1
53	3	3	4
54	2	2	4
55	3	4	2
56	1	1	4
57	4	3	3
58	1	1	3
59	4	2	2
60	3	4	4
61	3	3	3
62	*	*	*
63	*	*	*
64	4	4	4
65	2	1	2
66	1	3	4

Table 22: *Travel execution* dataset

Idnum	VOR1	VOR2	VOR3	FOR1	FOR2	FOR3
1	3	4	3	3	4	3
2	3	4	3	2	3	1
3	*	*	*	*	*	*
4	1	3	4	2	3	3
5	2	4	4	3	3	4
6	3	4	3	4	4	4
7	2	3	3	3	2	2
8	1	1	1	2	1	1
9	1	2	2	2	2	2
10	*	*	*	*	*	*
11	1	2	3	2	2	2
12	2	2	2	2	1	1
13	2	3	4	2	4	3
14	1	2	1	1	1	1
15	*	*	*	*	*	*
16	*	*	*	*	*	*
17	*	*	*	*	*	*
18	1	3	3	2	1	2
19	2	3	2	1	1	3
20	1	3	4	3	3	3
21	1	2	2	2	2	2
22	2	4	4	3	3	2
23	1	1	2	2	3	3
24	3	4	3	3	3	3
25	2	4	4	4	2	2
26	1	2	2	1	1	1
27	2	3	4	3	3	3
28	2	3	3	3	3	3
29	2	4	4	4	4	4
30	1	2	2	1	1	1
31	2	2	1	3	2	2
32	1	1	1	2	1	1
33	2	4	2	3	3	2
34	*	*	*	*	*	*
35	1	3	3	3	3	3
36	*	*	*	*	*	*
37	3	2	3	2	2	3
38	*	*	*	*	*	*
39	*	*	*	*	*	*
40	*	*	*	*	*	*
41	3	2	3	3	2	2
42	2	4	4	4	4	4
43	3	3	3	4	4	2
44	2	2	2	2	2	2
45	1	3	2	2	1	2
46	2	2	2	2	2	2
47	1	2	1	1	1	1
48	*	*	*	*	*	*
49	3	3	*	3	3	3
50	1	3	3	2	3	1
51	1	3	4	3	3	3
52	2	1	1	1	1	1
53	3	3	4	3	2	2
54	2	2	2	3	2	2
55	2	2	3	3	2	3
56	1	1	1	1	1	1
57	3	3	3	2	2	3
58	3	3	3	2	2	2
59	2	3	3	3	2	3
60	3	4	3	3	3	2
61	3	2	3	3	3	4
62	*	*	*	*	*	*
63	*	*	*	*	*	*
64	2	3	3	2	1	3
65	1	2	2	2	2	2
66	3	4	4	4	4	4

Table 23: *Settlement of account* dataset 1

Idnum	PRU1	PRU2	PRU3	PRU4	PRU5
1	2	2	3	2	2
2	2	3	2	3	1
3	*	*	*	*	*
4	2	4	2	4	1
5	4	4	3	3	3
6	2	2	2	3	2
7	4	4	3	3	2
8	2	*	1	2	*
9	1	2	2	2	2
10	*	*	*	*	*
11	2	2	1	1	2
12	2	2	2	2	1
13	4	2	3	3	2
14	1	1	2	1	1
15	*	*	*	*	*
16	*	*	*	*	*
17	*	*	*	*	*
18	2	2	2	2	1
19	3	3	4	4	3
20	2	1	2	4	1
21	2	3	2	3	1
22	4	4	2	2	2
23	2	4	1	1	1
24	3	2	2	3	3
25	4	2	4	4	1
26	4	2	3	4	1
27	2	3	2	2	2
28	3	2	2	3	2
29	2	4	4	4	3
30	1	4	1	1	1
31	2	1	2	3	1
32	2	1	2	2	1
33	3	3	1	1	2
34	*	*	*	*	*
35	3	4	3	4	1
36	*	*	*	*	*
37	4	4	2	1	2
38	*	*	*	*	*
39	*	*	*	*	*
40	*	*	*	*	*
41	3	3	2	2	2
42	2	4	2	2	2
43	4	4	4	4	4
44	2	3	2	*	*
45	4	4	2	2	4
46	3	3	3	3	1
47	2	2	*	2	*
48	*	*	*	*	*
49	3	4	3	3	2
50	2	2	1	2	1
51	3	4	2	3	1
52	2	4	1	3	1
53	3	3	2	4	3
54	4	3	3	3	3
55	3	3	2	2	2
56	3	3	2	2	1
57	4	4	3	2	3
58	4	4	2	3	2
59	4	3	2	1	*
60	3	3	2	2	*
61	2	2	2	3	*
62	*	*	*	*	*
63	*	*	*	*	*
64	4	1	3	4	4
65	1	1	3	2	2
66	4	1	2	4	1

Table 24: *Settlement of account* dataset 2

6. References

[Jensen (1996)] **An Introduction to Bayesian Networks**, Finn V. Jensen, 1996

[Jensen (2001)] **Bayesian Networks and Decision Graphs**, Finn V. Jensen, 2001

[Neapolitan (2004)] **Learning Bayesian Networks**, Richard E. Neapolitan, 2004

[Morawski (Mai 1989)] **Understanding Bayesian Networks**, Paul Morawsky, Mai 1989

[Charniak (Winter 1991)] **Bayesian Networks without Tears**, Eugene Charniak, Winter 1991

[Kipper, Brants, Plach and Schäfer (1995)] **Bayessche Netze: Ein einführendes Beispiel**, Kipper, Bernhard and Brants, Torsten and Plach, Marcus and Schäfer, Ralf, 1995

[Heckerman (March 1995)] **A Tutorial on Learning with Bayesian Networks**, David Heckerman, March 1995 (Revised 1996)

[Salini and Kennet (October 2009)] **Bayesian Network of Customer Satisfaction Survey Data**, Salini, Silvia and Kenett, Ron S., October 2009