# Exploring Bayesian belief network for risky behavior modelling: discretization and latent variables

**A Suvorova**[1]

[1] TICS Lab, SPIIRAS, 39 14th Line, St. Petersburg, Russia

**Abstract.** Decision making in many areas is based on data about individual behavior often measured using different surveys. The study investigates the proposed approach for behavior modelling on the base of Bayesian belief networks that allows predicting behavior characteristics using small and incomplete data from surveys about behavior episodes. We explored the characteristics of the models using the automatically generated dataset that included 44350 cases. During our experiment, we considered three different model structures and compared three different discretization strategies. We found that simpler structures showed better prediction quality for all measures (average accuracy, precision, recall, F1 score). The observed difference was statistically significant but did not exceed 1% that can be considered unimportant if error price is low. Our findings suggested that ways of data transformation, particularly discretization strategies for input data, had a significant impact on prediction quality: background knowledge about distributions, theoretical assumptions about behavior led to higher prediction quality.

## 1. Introduction

Decision making in many areas is based on data about individual behavior: personnel behavior [1–3], customer behavior [4], user behavior [5, 6], patient behavior [7, 8]. Most of the studies measures behavior frequency or behavior rate [8, 9]: using special devices, diaries or surveys. Since the diary method (recording of episodes) is extremely time-consuming, resource-intensive and even hardly possible, surveys are especially popular in psychology, sociology and public health research (for example, see [10]). But surveys cannot be very detailed and very long: respondents become tired, less attentive and not willing to continue [11]. As a result, researchers have to deal with small and sometimes incomplete data about behavior. In [12] authors provided the method based on data about several behavior episodes.

The idea of analysis of such kind of data is based on Bayesian Belief Network (BBN) theory that allows complementing empirical data with inputs from other models and expert knowledge [13]. Due to its features, BBNs are widely used in decision making in many areas (for example, see [14]).

Bayesian Belief Network is a type of probabilistic graphical models that represents a set of random variables and their conditional dependencies [13]. It consists of two components: structure and parameters. A network structure is presented in the form of a directed acyclic graph where nodes correspond to the random variables and directed edges represent dependencies among variables. Parameters are represented as a set of conditional probability distributions, one for each variable, characterizing the dependencies represented by the edges [15].

Previous studies showed high prediction quality of BBN models for estimating risky behavior characteristics [16–18]. This study explores the structure of the models and influence of structure changes on prediction quality.

## 2. Model description

The input data for the model [16] include the lengths of intervals between three last episodes of risky behavior and the lengths of minimum and maximum intervals between episodes during a given period of interest $T$ . The data about episodes in most applications is obtained from respondents' self-reports [12]. In addition, the model includes the latent variable that corresponds to the number of episodes during $T$ and the behavior rate, that is the key variable, the one we want to estimate. We assume that for each respondent occurrence of episodes follows Poisson random process.

Adding data about minimum and maximum intervals decreases the influence of recent behavior represented by the last episodes. However, combining all the data about episodes leads to very complicated joint distribution [19] that is impossible to represent as an elementary function.

On the contrary, Bayesian belief networks allow determining complex relationships in terms of simpler dependencies between small parts. Modelling risky behaviour as BBN gives a way to add all available data into the model as well as include expert assumptions about relationships between them and their distributions. Moreover, the existed software tools for BBNs representation, visualization, structure and parameter learning, inference and analysis, for example [20] or [21], allow researchers focus on description of the model while calculations are performed automatically.

The structure of BBN model is a directed acyclic graph $G(V,L)$ with vertices $V=\{\mathrm{le}_1,\mathrm{le}_2,\mathrm{le}_3,t_{\min},t_{\max},\mathrm{rate},n\}$ and edges (or links) $L=\{(u,v):u,v\in V\}$, where rate is random variable for behavior rate; $\mathrm{le}_i$ is random variable for the length of the interval between ($i$-1)-th and $i$-th episodes from the end (0 corresponds to interview moment); $t_{\min}$ and $t_{\max}$ (min and max in figures) are random variables for the length of minimum and maximum intervals; $n$ is random variable for the number of episodes during period of interest.

Previous studies [16, 17] proposed a BBN risky behavior model where edges were defined theoretically under the assumptions of Poisson random process. The corresponding structure is presented on figure 1. The theoretical assumptions require the inclusion of latent variable $n$ into the model while it is not observed in input data.

Further studies [22] explored other structures including those that were data-based and used structure-learning algorithms. The Hill-Climbing algorithm and other score-based methods in many experiments produced a simplified structure close to naive Bayes classifier (figure 2). In this structure the behavior rate is related to number of episodes and all other variables depended on the number of episodes only. This structure has simple interpretation because the number of episodes can be directly calculated on the base of the rate.

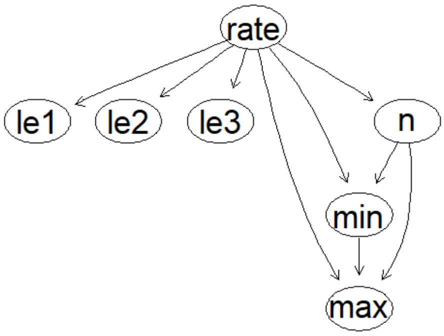

**Figure 1.** Initial model structure.

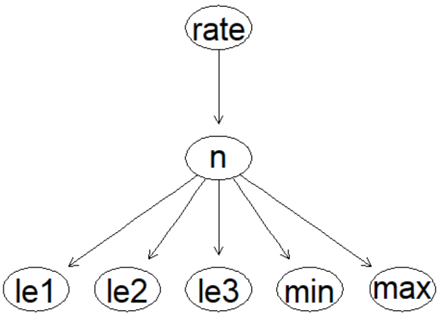

**Figure 2.** Simplified structure.

The aim of this study is to explore the next step: if the number of episodes can be directly estimated on the base of the rate and vice versa when $T$ is given, let's exclude one of these variables and explore a new, even more simplified model (figure 3). Thus, we plan to estimate the influence of latent variable $n$ .

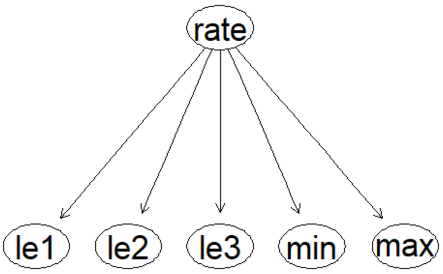

**Figure 3.** Simplified structure without latent variable *n*.

## 3. Methods

Since collecting the real data about behavior episodes is extremely time-consuming and requires financial and organizational resources, we explored the characteristics of the models using automatically generated dataset. The generation process assumes that the behaviour follows Poisson random process: the occurrence of the next episode is independent from the previous ones, length of interval between concurrent episodes follows exponential distribution. This assumption corresponds to the features of real-life risky behavior and it is widely used in previous studies [23]. The detailed description of the theoretical background and previously designed software is provided in [24]. The use of the automatically generated dataset provides an important advantage: we can compare the theoretical (the a priori given rate) and the estimated rate.

### *3.1. Data description*

On the first step we generated 1500 values for behavior rate from Gamma distribution (shape $k = 1.2$, scale $\theta = 0.3$). The choice of parameters' value was aimed to get dataset that corresponded to "real-life" risky behavior: behavior rate in most cases (in 94% of cases in our dataset) was less than 1 and generated about one episode in 3–4 days.

Next step included generation of 30 "respondents" (sequences of behavior episodes) for each rate value for period of 365 days in total, that summed up to 45000 sequences of episodes. Then we calculated initial data for the model: lengths of minimum, maximum intervals between episodes and lengths of intervals between the last three episodes. After deletion of incomplete cases (e.g. cases with only one episode during 365 days) the final dataset included 44350 cases (or "respondents").

To estimate prediction quality of the models we randomly select 5000 cases to the test dataset and do not use them in model learning.

### *3.2. Discretization strategies*

As mentioned in Section 2 we conducted the series of experiments to explore the influence of the latent variable $n$ (the number of behavior episodes during period of interest $T$) on the prediction quality of the model. Hence, we considered three different model structures (Figure 1–3). Since we explored discrete Bayesian networks, the discretization strategy can influence the prediction quality measures. So, in this study, we compared three different discretization strategies. The detailed description of each strategy is provided below. We used the same number of intervals for all strategies for further comparison.

The first discretization is based on breaks provided by experts. The general idea is to make smaller intervals for more frequent values. In addition, this discretisation uses more interpretable breaks. For example, $\mathsf{rate} \in [0.5,1)$ means that there was one behavior episode in 1–2 days. Moreover, interpretability becomes more important for variables corresponded to interval lengths: $\mathsf{le}_1 \in [7,14)$ means that the last behaviour episode was 1–2 weeks ago. The usage of weeks / months / year notation can simplify the questionnaire design for real-world applications.

The second discretization strategy is quite simple: we divided possible range into intervals with equal lengths and then added the last interval with infinity as the upper limit. We included background knowledge about behavior characteristics in this strategy too: the infinite interval started at 1 for rate variable (since we assumed that most of the cases had rate less than 1) and it started at 180 for time-length variables (since we assumed at least two behavior episodes).

The last strategy uses the idea of equal interval probabilities not lengths and is based on quantile calculation. At the experiment we used sample quantiles but for real-world problems when the rate distribution is unknown it is possible to use theoretical quantiles based on distribution assumptions.

The example of the variable discretization according to the strategies is shown on figure 4.

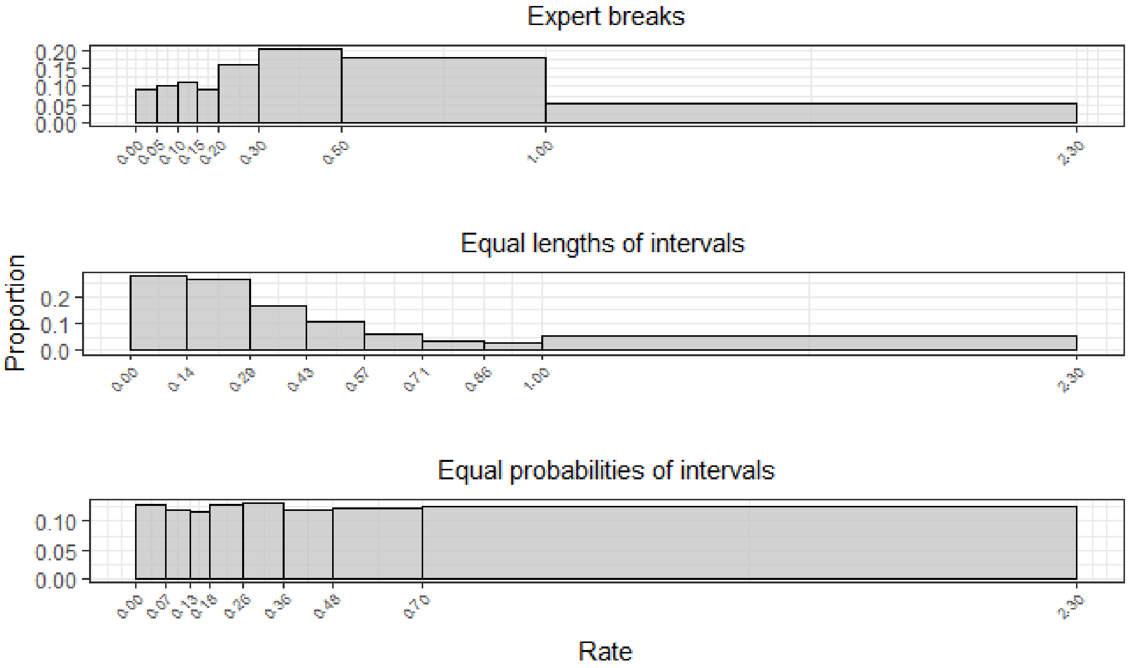

**Figure 4.** Discretization strategies: rate variable example.

### *3.3. Experiment design*

First, we applied the described discretization strategies to the train dataset that consisted of 39350 cases in total. On each iteration we randomly selected 10000 cases from our train dataset. Then we learned the parameters of Bayesian belief network for all three structure models described in Section 2: 1) the initial model, 2) the simplified model with fewer links and 3) the last model without $n$ variable. Thus, we had nine different settings at each iteration (all possible combinations of three model structures and three discretization strategies). Finally, we used test dataset with corresponding discretization to estimate prediction quality (accuracy, precision, recall, F1 score). All measures were calculated according to multi-class classification metrics (average accuracy, macro precision, macro recall, macro F1 score) [25]. To summarize the results we repeated the experiment 50 times.

The calculations and statistical analysis were performed using R [26]; in particular, for Bayesian network analysis we used *bnlearn* [27] and *gRain* [28] packages.

## 4. Results

The average accuracy through 50 iterations is shown on figure 5. The discretization with equal probabilities presented the highest results for all model structures, while the discrete intervals with equal length were the worst approach.

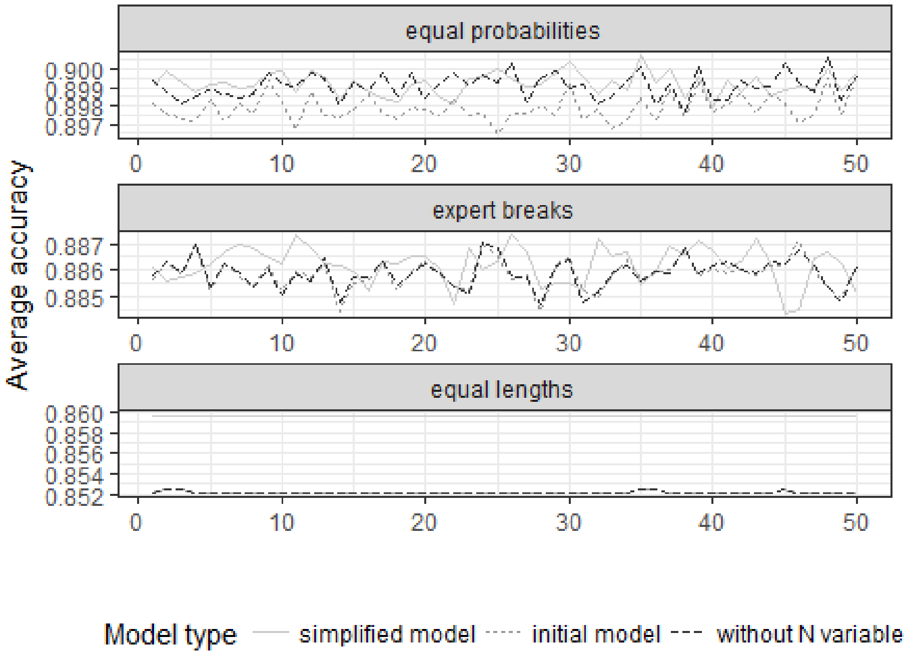

**Figure 5.** Average accuracy measures.

The F1 score presented the same pattern: the highest measures for discretization with equal probabilities, the lowest for the one with equal interval lengths (figure 6). F1 score was even more skewed: it was about 0.6 for equal probability strategy, about 0.54 for expert-defined breaks and 0.16–0.22 only for equal length strategy. The latter was explained with further analysis of precision and recall measures. The discretization with equal lengths had a poor prediction quality for all classes except the first two of them. Due to extremely skewed distribution (see figure 4), neither model, regardless of its structure, predicted rates higher than 0.3, so predictions were all at the first two classes.

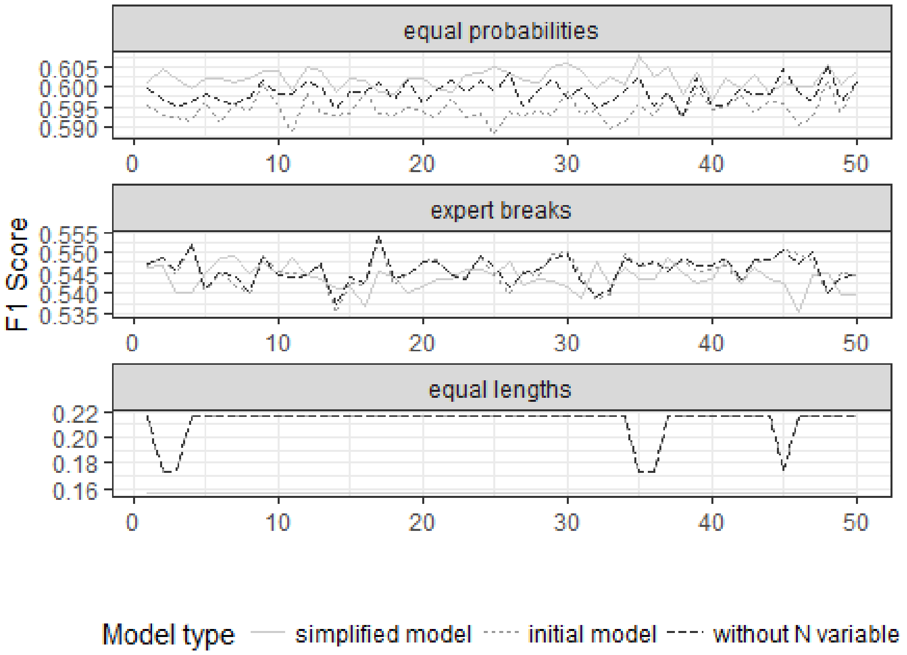

**Figure 6.** F1 measures.

Mean quality measures and the standard deviations (SD) for all model structures and discretization strategies are summarized in table 1. Since the discretization with equal lengths showed the similar results for the initial structure and the structure without $n$ variable and slightly better but still poor results for the simplified model (structure on figure 2) we focused on other discretization strategies.

**Table 1.** Prediction quality on the test dataset.

| Model type | Accuracy, mean (SD) | Average accuracy, mean (SD) | Precision, mean (SD) | Recall, mean (SD) | F1Score, mean (SD) |
|---|---|---|---|---|---|
| Expert discretization | | | | | |
| initial model | 0.5433 (0.0023) | 0.8858 (0.0006) | 0.5450 (0.0057) | 0.5458 (0.0024) | 0.5454 (0.0037) |
| simplified model | 0.5448 (0.0028) | 0.8862 (0.0007) | 0.5517 (0.0026) | 0.5361 (0.0056) | 0.5438 (0.0031) |
| without N variable | 0.5435 (0.0022) | 0.8859 (0.0006) | 0.5456 (0.0052) | 0.5459 (0.0024) | 0.5457 (0.0034) |
| Equal length intervals | | | | | |
| initial model | 0.4085 (0.0004) | 0.8521 (0.0001) | 0.1979 (0.0143) | 0.2306 (0.0122) | 0.2130 (0.0135) |
| simplified model | 0.4390 (0.0000) | 0.8598 (0.0000) | 0.1277 (0.0000) | 0.2021 (0.0000) | 0.1565 (0.0000) |
| without N variable | 0.4085 (0.0004) | 0.8521 (0.0001) | 0.1979 (0.0143) | 0.2306 (0.0122) | 0.2130 (0.0135) |
| Equal probabilities (quantiles) | | | | | |
| initial model | 0.5915 (0.0029) | 0.8979 (0.0007) | 0.5953 (0.0033) | 0.5934 (0.0029) | 0.5944 (0.0031) |
| simplified model | 0.5969 (0.0024) | 0.8992 (0.0006) | 0.6055 (0.0027) | 0.5986 (0.0023) | 0.6020 (0.0024) |
| without N variable | 0.5964 (0.0028) | 0.8991 (0.0007) | 0.5989 (0.0031) | 0.5984 (0.0027) | 0.5986 (0.0029) |

To compare all quality measures among the proposed models we run pairwise t-test for multiple comparisons with Bonferroni correction. The simplified model had statistically significant higher prediction quality measures comparing to both initial model and model without $n$ variable in case of expert-defined discrete intervals (table 1 (mean values) and table 2 (p-values)). There was no statistically significant difference in quality measures between initial model and model without latent $n$ variable.

In case of discretization with equal probabilities of intervals, both the model with simplified structure and the model with structure without $n$ variable outperformed the initial model. The simplified model also had significantly higher precision comparing to the model without $n$ variable with the same levels of other measures.

**Table 2.** Prediction quality measures: model comparison (p-values according to pairwise t-test with Bonferroni correction).

| Comparison | Accuracy | Average accuracy | Precision | Recall | F1Score |
|---|---|---|---|---|---|
| Expert discretization | | | | | |
| initial vs simplified | 0.014 | 0.014 | <0.001 | <0.001 | 0.060 |
| initial vs without N | 1 | 1 | 1 | 1 | 1 |
| without N vs simplified | 0.033 | 0.33 | <0.001 | <0.001 | 0.016 |
| Equal length intervals | | | | | |
| initial vs simplified | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| initial vs without N | 1 | 1 | 1 | 1 | 1 |
| without N vs simplified | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| Equal probabilities (quantiles) | | | | | |
| initial vs simplified | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| initial vs without N | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| without N vs simplified | 1 | 1 | <0.001 | 1 | <0.001 |

## 5. Conclusion

Risky behavior modelling, behavior parameter estimating and prediction became the focus of studies in many research areas. An increased attention is given to investigation of online behavior and prediction of social media user's characteristics [29, 30], but traditional survey-based studies continue to be promising, helpful and efficient.

The current study explores the proposed behavior model on the base of Bayesian belief networks that allows processing incomplete data about behavior episodes and predicting behavior characteristics.

Our findings suggested that ways of data transformation, particularly discretization strategies for input data, had a significant impact on prediction quality: background knowledge about distributions, theoretical assumptions about behavior led to higher prediction quality.

We found that for the same period of interest the simpler structures showed better prediction quality but it is important to mention that the quality difference did not exceed 1% (for example, 89.7% vs 89.9% for average 8-class accuracy for equal probability discretization). The effect of this difference can be estimated in practical settings only: sometimes the error price is high; in other circumstances it is diminishing.

In general, proposed model showed high prediction quality and has a great potential for analyzing real-life behavior problems.

## 6. References

[1] Hasanbeigi A, Menke C and Du Pont P 2010 Barriers to energy efficiency improvement and decision-making behavior in Thai industry *Energy Efficiency* **3(1)** pp 33–52

[2] van Ryn M, Burgess DJ, Dovidio JF, Phelan SM, Saha S, Malat J, Griffin JM, Fu SS and Perry S 2011 The impact of racism on clinician cognition, behavior, and clinical decision making *Du Bois review: social science research on race* **8(1)** pp 199–218

[3] Rubin EV and Kellough JE 2011 Does civil service reform affect behavior? Linking alternative personnel systems, perceptions of procedural justice, and complaints *Journal of Public Administration Research and Theory* **22(1)** pp 121–141

[4] Mohan G, Sivakumaran B and Sharma P 2013 Impact of store environment on impulse buying behavior *European Journal of Marketing* **47(10)** pp 1711–1732

[5] Farzan R and Brusilovsky P 2011 Encouraging user participation in a course recommender system: An impact on user behavior *Computers in Human Behavior* **27(1)** pp 276–284.

[6] Beutel A, Akoglu L and Faloutsos C 2015 Fraud detection through graph-based user behavior modelling *Proc. of the 22nd ACM SIGSAC Conf. on Computer and Communications Security* ACM pp 1696–1697

[7] Amundsen A, Nordøy T, Lingen KE, Sørlie T and Bergvik S 2018 Is patient behavior during consultation associated with shared decision-making? A study of patients' questions, cues and concerns in relation to observed shared decision-making in a cancer outpatient clinic *Patient education and counseling* 2018 **101(3)** pp 399–405

[8] Hojilla JC, Koester KA, Cohen SE, Buchbinder S, Ladzekpo D, Matheson T and Liu AY 2016 Sexual behavior, risk compensation, and HIV prevention strategies among participants in the San Francisco PrEP demonstration project: a qualitative analysis of counseling notes *AIDS and Behavior* **20(7)** pp 1461–1469

[9] Chiauzzi E, Rodarte C and DasMahapatra P 2015 Patient-centered activity monitoring in the self-management of chronic health conditions *BMC medicine* **13(1)** p 77

[10] Ehlers AP, Drake FT, Kotagal M, Simianu VV, Achar C, Agrawal N, Joslyn SL and Flum DR 2017 Factors influencing delayed hospital presentation in patients with appendicitis: the APPE survey *Journal of Surgical Research* **207** pp 123–130

[11] Hardigan PC, Popovici I and Carvajal MJ 2016 Response rate, response time, and economic costs of survey research: a randomized trial of practicing pharmacists *Research in Social and Administrative Pharmacy* **12(1)** pp 141–148

[12] Tulupyeva T, Paschenko A, Tulupyev A, Krasnoselskikh T and Kazakova O 2008 H*IV risky behavior models in the context of psychological defense and other adaptive styles* Nauka, SPb

[13] Pearl J 2000 *Causality: Models, Reasoning, and Inference* Cambridge University Press, Cambridge

[14] Barton DN, Benjamin T, Cerdan CR, DeClerck F, Madsen AL, Rusch GM and Villanueva C 2016 Assessing ecosystem services from multifunctional trees in pastures using Bayesian belief networks *Ecosystem Services* **18** pp 165–174

[15] Darwiche A 2009 *Modelling and reasoning with Bayesian networks* Cambridge: Cambridge University Press

[16] Suvorova A 2013 Socially significant behavior modeling on the base of super-short incomplete set of observations *Information-measuring and Control System*s **9(11)** pp 34–38

[17] Suvorova, AV, Tulupyev AL and Sirotkin AV 2014 Bayesian belief networks for risky behavior rate estimates. *Nechetkie sistemy i myagkie vychisleniya [Fuzzy Systems and Soft Computing]* **9(2)** pp 115–129

[18] Suvorova A and Tulupyev AL 2016 Evaluation of the model for individual behavior rate estimate: Social network data *XIX IEEE Int. Conf. on Soft Computing and Measurements (SCM)* IEEE pp 18–20

[19] Stepanov DV, Musina VF, Suvorova A, Tulupyev AL, Sirotkin AV and Tulupyeva TV 2012 Risky behavior Poisson model identification: heterogeneous arguments in likelihood *Trudy SPIIRAN* **23** pp 157–184

[20] *GeNIe& SMILE*. Decisions systems laboratory. School of Information Sciences. University of Pittsburg http://genie.sis.pitt.edu/

[21] *AgenaRisk Bayesian network tool* http://www.agenarisk.com

[22] Suvorova A and Tulupyev A 2016 Learning Bayesian Network Structure for Risky Behavior Modelling *Proc. of the 3rd Int. Scientific Conf. on Intelligent Information Technologies for Industry (IITI'16)*. Springer International Publishing pp 95–102

[23] Spiegelman D and Hertzmark E 2005 Easy SAS calculations for risk or prevalence ratios and differences *American journal of epidemiology* **162(3)** pp 199–200

[24] Suvorova A 2015 Test data generator for risky behavior probabilistic graphical model *Proc. of the VIII Int. Scientific Conf. on Integrated models and soft computing in artificial intelligence* Fizmatlit 2 pp 799–805

[25] Sokolova M and Lapalme G 2009 A systematic analysis of performance measures for classification tasks *Information Processing & Management* **45(4)** pp 427–437

[26] R Core Team 2017 *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/

[27] Scutari M 2010 Learning Bayesian Networks with the bnlearn R Package *Journal of Statistical Software* **35(3)** pp 1–22

[28] Hojsgaard S 2012 Graphical Independence Networks with the gRain Package for R *Journal of Statistical Software* **46(10)** pp 1–26

[29] Preoţiuc-Pietro D, Volkova S, Lampos V, Bachrach Y and Aletras N 2015 Studying user income through language, behaviour and affect in social media *PloS one* **10(9)** e0138717.

[30] Ruths D and Pfeffer J 2014 Social media for large studies of behavior *Science* **346** pp 1063–1064

**Acknowledgments**

This work was partially supported by the by RFBR according to the research project No. 16-31-60063 and No. 18-01-00626