The Copula Bayesian Network with Mixed Discrete and Continuous Nodes to Forecast Railway Disruption Lengths

Aurelius A. Zilko [a,1], Dorota Kurowicka [a], Anca M. Hanea [b], Rob M.P. Goverde [c]
[a] Delft Institute of Applied Mathematics, Delft University of Technology
2628CD, Delft, The Netherlands
[1] E-mail: A.A.Zilko@tudelft.nl, Phone: +31 (0) 15 27 87261
[b] Centre of Excellence for Biosecurity Risk Analysis, The University of Melbourne
3010, Melbourne, Australia
[c] Department of Transport and Planning, Delft University of Technology
2628CN, Delft, The Netherlands

Abstract
The highly uncertain nature of a railway disruption complicates the tasks carried by the Dutch Operational Control Centre Rail (OCCR) in the Netherlands. A good prediction of disruption length is believed to help the decision making in dealing with the disruption. Zilko, et al. [Non-Parametric Bayesian Network to Forecast Railway Disruption Lengths, In: *The Second International Conference on Railway Technology: Research, Development and Maintenance (Railways 2014)*, Ajaccio, France (2014)] proposes the use of the Non-Parametric Bayesian Network (NPBN) method, a graphical model based on a probabilistic approach that represents the dependence between the variables of interest, to predict the disruption length. The model offers an attractive feature for the real-time decision making environment in the OCCR, which is its efficiency and fast inference. This paper extends the model construction. More variables are added into the NPBN model to increase the prediction power of the model. From the data analysis, it turns out that some of the influencing variables are discrete variables; thus resulting in a mixed discrete and continuous model. This raises some serious questions because the NPBN method is originally designed to work with continuous variables. This paper investigates how the presence of discrete variables affects the NPBN method and how one can proceed with the mixed discrete and continuous model. As an illustration, a model about the railway disruption in the Netherlands caused by track circuit failures is presented in this paper as well.

Keywords
Bayesian Networks, Copula Bayesian Networks, Railway Disruptions, Railway Traffic Management, Uncertainty Analysis.

1 Introduction

Zilko et al. (2014) proposes the use of the NPBN method in constructing a probabilistic model to predict the railway disruption lengths. The NPBN's fast computational nature makes this method very interesting from the point of view of the Operational Control Centre Rail (OCCR), which is working in a real-time decision making environment. Separate NPBN models for each type of (major) disruption in the Dutch railway network are going to be constructed. To predict the length of a disruption, the NPBN model is conditionalized on the variables describing the situation of interest. In this paper, as an illustration, the simple example presented in Zilko et al. (2014) is extended, which is a model for the GRS track circuit (TC) disruption in the Netherlands.

The NPBN algorithm has already been implemented in a user-friendly and computationally efficient software, UNINET, which has been developed at Delft University of Technology. The software is available at www.lighttwist.net/wp/uninet. Hence, the model can immediately be used in real-time as a decision support tool of the OCCR.

Before proceeding further, it is necessary to divide the disruption length into two mutually exclusive definitions of time: the latency time and the repair time. The latency time is the length of time the repair team needs to go to the disruption site. The repair time is the length of time they need to solve the problem. These two definitions of time are affected by different factors.

More factors and variables have been analyzed and included in the NPBN method. As it turns out, some of the influencing variables are discrete variables; resulting in a mixed discrete-continuous model. This raises some questions because the NPBN model is originally introduced to work with continuous variables (Kurowicka and Cooke (2005), Hanea et al. (2006), Hanea et al. (2010)).

As its name suggests, the NPBN method is a Bayesian Network (BN)-based approach. The model is constructed as a BN with nodes and (uncyclic) arcs which represent the variables and the flow of influence between the variables, respectively. In this paper, the structure of the BN is learned from the data, with a small adjustment performed to the structure to deal with the mixed discrete-continuous nature of the BN.

The NPBN method specifies the dependence between the variables with a set of (conditional) copulas that are parameterized by the (conditional) rank correlations. A copula is the joint distribution of n uniform variables in the $n-$dimensional unit cube. The heart of the copula application in dependence modelling lies on the Sklar's theorem which states that any joint cumulative distribution function of variables $(X_1, \ldots, X_n)$, denoted as $F_{1,\ldots,n}$, can be rewritten in terms of the corresponding copula C as (Sklar (1959)):

$$F_{1,\ldots,n}(x_1, \ldots, x_n) = C(F_1(X_1), \ldots, F_n(X_n)) \tag{1}$$

where $F_i(X_i)$ denotes the marginal distribution of the i-th variable. If the variables are continuous, the copula satisfying equation (1) is unique. For discrete variables, however, the copula satisfying equation (1) is no longer unique.

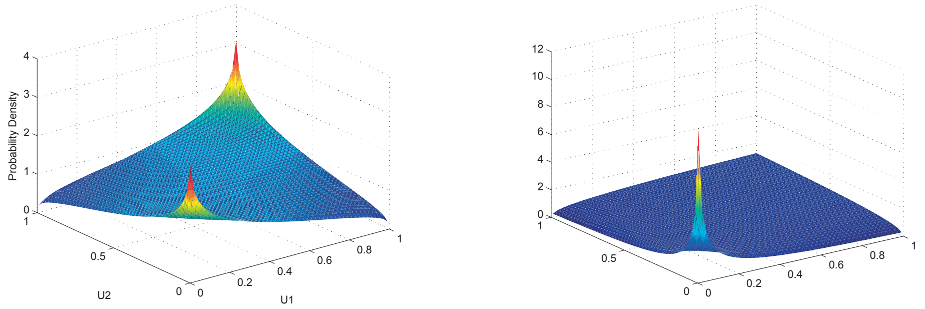
There are many different available copulas, one of them is the Normal, or Gaussian, copula. This copula is defined as:

$$C_R(u_1, \ldots, u_n) = \Phi_R(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)) \tag{2}$$

where Φ^{-1} denotes the inverse cumulative distribution of a univariate standard normal distribution and Φ_R denotes the joint cumulative distribution of a multivariate normal distribution with zero mean and correlation matrix R.

For the bivariate case, the density of the Normal copula with correlation ρ_{12} is:

$$c_{\rho_{12}}(u_1, u_2) = \frac{1}{\sqrt{1-\rho_{12}^2}} \exp\left(-\frac{\rho_{12}^2\Phi^{-1}(u_1)^2 - 2\rho_{12}\Phi^{-1}(u_1)\Phi^{-1}(u_2) + \rho_{12}^2\Phi^{-1}(u_2)^2}{2(1-\rho_{12}^2)}\right). \tag{3}$$



(a) The density of a Normal copula with $\rho_{12} = 0.2$.

(b) The density of a Clayton copula with $\theta = 0.3$.

Figure 1: The densities of two bivariate copulas.

Figure 1(a) shows the density of the bivariate Normal copula when $\rho_{12} = 0.2$.

Another copula is, for instance, the Clayton copula (Clayton (1978)). This copula is interesting for some applications because it captures the lower tail dependence between the variables. The formula for a bivariate Clayton copula is given by:

$$C_\theta(u_1, u_2) = \left(\max(u_1^{-\theta} + u_2^{-\theta} - 1, 0)\right)^{-1/\theta} \tag{4}$$

where θ is the copula's parameter with domain $[-1, \infty)\backslash\{0\}$. Figure 1(b) shows the density of a bivariate Clayton copula with $\theta = 0.3$.

In this project, the choice of Normal copula as the underlying copula of the NPBN method is very attractive because it allows conditionalization to be performed analytically. As a result, the computation requires very little amount of time and this goes hand in hand with the real-time decision making process of the OCCR. Therefore, later on, it is investigated how the Normal copula fits the data for the NPBN method when some of the variables are discrete.

The rest of the paper is organized as follows: In Section 2, the data analysis of the factors influencing the disruption length is presented. The data analysis reveals that several of the factors influencing the disruption length are discrete variables. Therefore, Section 3 follows where discussion about the presence of discrete variables in the NPBN method takes place. Section 4 contains the Normal copula validation, the BN structure learning, the BN model validation, and a case study of the application of the model. The paper is summarized in Section 5 where the conclusions and future work are presented.

2 Data Analysis

The analysis presented in this paper is performed based on the same data source as in Zilko et al. (2014). The data contains the historical TC problem in the entire Dutch railway network from 1 January 2011 up to 30 June 2013. Only high priority incidents, i.e. incidents which require urgent actions, are considered. This filtering results in 2113 sample points from this time period.

The two definitions of time, the latency time and the repair time, are affected by different factors. Unfortunately, the main factor believed to affect the length of the repair time, the technical cause of the disruption, e.g. what went wrong, what the repair team did to solve the problem, etc., is missing in the main data source that is used in this project. To tackle this problem, an expert judgment exercise is planned to be performed in the future. For the time being, the focus lies on modelling the factors affecting the latency time.

2.1 Factors Influencing the Latency Time

In general, the factors influencing the length of latency time can be divided into three different groups: time, location, and the weather. Later on, one extra variable which does not belong to any of these groups, the presence of an overlapping disruption, surfaces.

Time

In the current model, there are two variables that represent the time: whether the incident occurs during the repair team's contractual working hours or not and whether it is during the rush hour period or not.

Regarding the first variable, a different operation is performed by the repair team depending whether an incident occurs during their contractual working hours (on a weekday between 7 AM and 4 PM) or not. If an incident occurs within this period, they leave from their working station (post) to the disruption site. Outside of this time period, the repair team is not in their working station. Instead, they are available on call and, when an incident occurs, they leave from wherever they are to the disruption site.

The repair team travels to the disruption by cars. Therefore, it might be of importance to consider whether an incident occurs during rush hour time. During this time, it is more likely to encounter traffic jams, which can affect the latency time.

Figure 2 presents the difference in the distribution of the latency time during the repair team's contractual working hours or not. The latency time during non working hours appears to be longer than the latency time during working hours. The mean of latency time during the working hours is 40.52 minutes while the otherwise is 44.87 minutes. This conclusion is supported by performing the two-sample Kolmogorov-Smirnov (KS) test to see whether the two distributions are statistically different or not. Higher p-value of the test indicates that there is not enough evidence in the data to conclude that the two distributions are different. Normally, the threshold for the p-value is chosen to be 5%. Performing the test to the two distributions of the latency time yields a p-value of 0.00000907, confirming

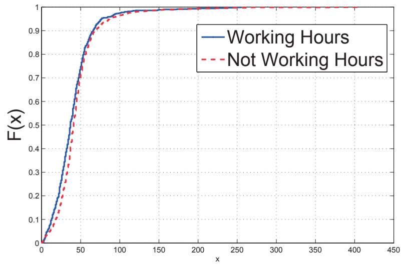


Figure 2: The cdf of the latency time whether an incident occurs during the repair team's contractual working hours or not.

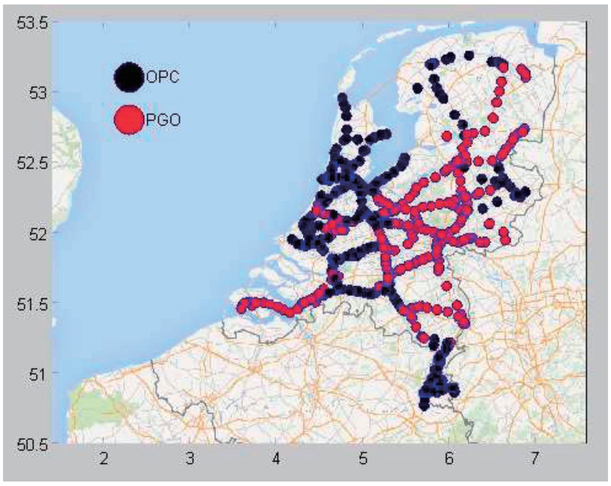


Figure 3: The spread of the OPC and PGO contracts in the Netherlands.

that the two distributions cannot be considered the same.

The influence of rush hour itself on the latency time, however, is small. While the latency time during rush hour is (slightly) longer, executing the KS test to the distribution of the latency time yields a p-value of 0.2333.

Location

In the model, the location needs to be modeled through some representative properties of the location. Three variables represent the location of the disruption in the current model: the distance to the nearest contractor's post, the distance to the nearest level crossing, and the type of contract.

The first variable has been discussed in Zilko et al. (2014). However, the data has been refined with more information about the contractors. Zilko et al. (2014) assumes only four main contractors are present which correspond to the four regions of the Dutch railway network. More detailed information has enabled more relevant analysis since then, where now it is known which exact contractor (and the location of their corresponding post) is responsible for each recorded incident. With this new information, the distance has been recalculated. However, even with the new information, the rank correlation between latency time and the distance to the nearest contractor's post still remains small at 0.1402 with a 95% confidence bound of $(0.0968, 0.1829)$. Because zero is not included in the confidence bound, this indicates a small positive dependence between the two variables.

Because the repair team drives a car to go to the disruption site, they need to park their car somewhere as close as possible to the disruption site. Usually, they do this by going to the nearest level crossing, park their vehicle there, and walk to the disruption site. This is the reason the second variable, the distance to the nearest level crossing, becomes interesting. The rank correlation of this variable with the latency time is 0.0909 with 95% confidence bound of $(0.0472, 0.1342)$; thus, again, indicating a small positive dependence.

In the Netherlands, currently there are two type of contracts between ProRail, the organization responsible for the railway infrastructure, and the contractors. The old OPC contract

(*Output-procescontracten* in Dutch, translated as the Output-based Contract) is based on the amount of work the contractors perform and the new PGO contract (*Prestatiegericht Onderhoud* in Dutch, translated as the Performance-based Maintenance) introduces a penalty if the work takes too much time. Plotting the geographical spread of these two contracts, as in Figure 3, indicates that certain areas of the Netherlands are still with the old OPC contract while the others already switch to the new PGO contract. Therefore, contract type can be seen as one property of location. This is especially interesting because the railway network in the most crowded area of the Netherlands, the Randstad, is actually still with the older OPC contract.

The OPC contract type appears to lead to longer latency time than the newer contract type. The mean of the latency time of the OPC contract is 45.1680 minutes while for the PGO contract, the mean is 41.9217 minutes. This conclusion is supported by the KS-test as well where the p-value is 0.0011.

The Weather and An Overlapping Disruption

A TC problem occurs more frequently during times when the temperature is high. For example, in the afternoon of 28 June 2011, the temperature in the Netherlands went above 30^oC. A closer look into the data recorded from this day show that there were 19 high priority TC problems occurring in the network while the average daily number of TC problems was 2.32. Some remarks indicated that extreme heat played a role in the incidents.

In this project, the temperature of 25^oC is chosen as the threshold to still have a reasonable amount of samples representing the warm weather. With this threshold, 123 incidents occur where the temperature is above or equal to 25^0C, or about 5.82% of all recorded TC incidents.

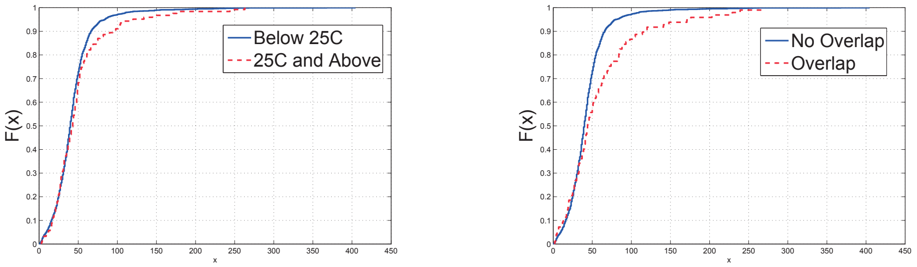
Table 1: Data division based on the temperature threshold and the presence of an overlapping incident.

	No Overlap	Overlap	
$< 25^o$C	1911	79	1990
$\geq 25^o$C	105	18	123
	2016	97	

Two incidents are said to be overlapping if they are handled by the same contractor, of the same technical field, and the closest contractor's posts are the same. Table 1 presents the division of the data based on whether or not the temperature is above 25^oC and whether or not there is an overlapping incident. It shows that when the temperature is below 25^oC, 3.97% of all incidents have an overlapping incident; while when the temperature is 25^oC or above, 14.63% of all incidents have an overlapping incident. Thus, whether or not the temperature is higher increases the chance for an overlapping disruption to occur.

Nonetheless, the more interesting question is how these two variables affect the latency time. Intuitively, when there is an overlapping incident, the latency time of the second incident should be longer because the repair team must finish their work on the first incident before being able to go to the second incident. Figure 4 presents the difference in the latency time based on these two variables.

One has to be careful in interpreting Figure 4 because of the very different number of samples in each of the categories. Performing the KS-test to both cases yields a p-value of 0.3888 for the warm/not warm variable and 0.000918 for the overlap/no overlap variable.



(a) The cdf of the latency time with respect to whether or not the temperature is above the 25^oC threshold.

(b) The cdf of the latency time with respect to the presence of an overlapping incident.

Figure 4: The effect of warm/not warm and overlap/no overlap variables on the latency time.

Therefore, the warm/not warm variable does not have direct effect on the latency time but it influences the presence of an overlapping incident which affects the latency time.

3 Discrete Variables in the NPBN

From the data analysis, it turns out that five of the eight variables involved in the model are discrete variables, resulting in a mixed discrete-continuous model. The NPBN is originally designed to work with BN with continuous variables (Kurowicka and Cooke (2005)). Since the core of the NPBN method lies in the use of copula to model the dependence between the variables, this section investigates how copula can be used for discrete variables.

3.1 Bivariate Discrete Distribution

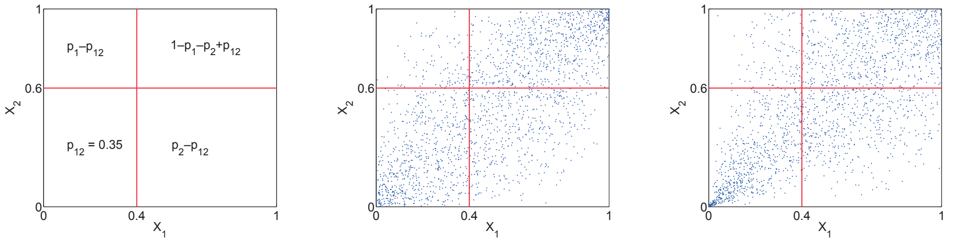
The discussion starts with the simplest possible case, the bivariate Bernoulli distribution, i.e. a discrete joint distribution of two variables, X_1 and X_2, with binary marginals.

Let $\mathbb{P}(X_1 = 0) = p_1$ and $\mathbb{P}(X_2 = 0) = p_2$ define the marginals of the joint bivariate Bernoulli distribution. A copula C is said to model the dependence of the variables X_1 and X_2 if it satisfies equation (1), which for this case takes the form of:

$$\mathbb{P}(X_1 = 0, X_2 = 0) = p_{12} = C(p_1, p_2). \tag{5}$$

When the variables are discrete, the copula satisfying equation 5 is no longer unique (Genest and Nešlehová (2007)). On a side note, it can also be shown that the parameter of a Normal copula that satisfies equation (5) exists for all possible choice of p_1, p_2, and p_{12}.

Example 1. *Let $p_1 = 0.4$, $p_2 = 0.6$, and $p_{12} = 0.35$. Figure 5 illustrates how the process of finding a copula that satisfies equation (5) works for this case. Figure 5(a) depicts the joint discrete distribution at hand in the unit square. The horizontal and vertical axis correspond to the probability of variable X_1 and X_2 respectively. The interval from 0 to $p_1 = 0.4$ in the horizontal axis corresponds to the value of $X_1 = 0$ while the otherwise corresponds to $X_1 = 1$; and similarly for the vertical axis. A copula needs to be fitted such that the mass, which corresponds to the joint probability, in the bottom left rectangle equals to $p_{12} = 0.35$, i.e. it satisfies equation (5).*



(a) The discrete joint distribution at hand.

(b) Fitting a Normal copula with parameter $\rho = 0.6990$.

(c) Fitting a Clayton copula with parameter $\theta = 1.8934$.

Figure 5: Fitting copulas to the bivariate discrete distribution in Example .

The information about p_1, p_2, and p_{12} are sufficient to fully represent the joint distribution. As shown in Figure 5(a), the mass in the two left rectangles must sum up to $p_1 = 0.4$, hence the mass in the top left rectangle must equal to $p_1 - p_{12} = 0.05$; and similarly for the two bottom rectangles. Because the mass in a rectangle is the probability of the corresponding pair of (X_1, X_2) the rectangle represents, the mass of the four rectangles must sum up to 1. Consequently, the top right rectangle's mass must equal to $1 - p_1 - p_2 + p_{12}$.

When fitting a Normal copula, the Normal copula with parameter $\rho = 0.6990$ yields a mass of 0.35 in the bottom left rectangle, as shown in Figure 5(b). This can be interpreted as 35% of the points in the scatter plot in Figure 5(b) falls inside this rectangle. However, a Clayton copula with parameter $\theta = 1.8934$ yields a mass of 0.35 in the bottom left rectangle too, as shown in Figure 5(c).

Furthermore, the range of possible copula satisfying equation (5) can be very wide. To show this, the bounds introduced by Carley (2002), which are the best pointwise bounds for all copula satisfying equation (5), can be used. For the case in Example 1, the lower Carley bound copula has a Spearman's rank correlation of -0.486 while the upper Carley bound's is 0.978. This means that the range of copula satisfying equation (5) for this particular example has correlation from as low as -0.486 to as high as 0.978.

However, the non-uniqueness problem of the copula is not a problem if one already knows the family of copula (s)he wants to work with. The only remaining question is whether or not the chosen copula solves equation (5).

In the continuous setting, a popular approach on how to estimate the parameter of a copula is by establishing a functional relationship between the parameter with some concordance measure that can be computed from the data, e.g. the Spearman's rank correlation or the Kendall's tau rank correlation. However, when applying the same technique to the discrete case, this results in a biased estimate of the parameter as noted by Genest and Nešlehová (2007). This means that another approach is needed for the parameter estimation. For this, the standard maximum likelihood estimation works. Therefore, later on in this paper, the parameter of the copula is estimated using the maximum likelihood.

Conditionalization

The prediction from the BN model is obtained through conditioning the BN on observed variables. For this reason, the choice of Normal copula is very attractive in a continuous BN model because the conditional probability function is known. Conditioning a multivariate normal distribution on any fixed value yields a multivariate normal distribution.

However, when the Normal copula is used to model the dependence between discrete variables, observation of a variable does not correspond to conditioning the Normal copula on a fixed value. Instead, the conditionalization is performed on an interval. Yet, the conditional distribution of a multivariate normal distribution given an interval is no longer multivariate normal (Swaroop et al. (1980)).

This problem, however, can be tackled by sampling the conditioning interval uniformly and then conditioning the multivariate distribution on each of these samples. Each of the conditionalized multivariate distribution is, then, sampled. The union of these samples is taken as the samples of the conditional distribution of the multivariate normal distribution given the corresponding interval. This, however, will certainly cost in longer computation time.

3.2 Moving On to the Higher Dimension

The same copula fitting procedure as in the bivariate Bernoulli case can be extended to a higher dimenstion to fit the multivariate Bernoulli distribution. For example, in the trivariate Bernoulli distribution of (X_1, X_2, X_3), instead of a unit square divided into four rectangles as in Figure 5(a), one has a unit cube divided into eight smaller rectangular cuboids each corresponding into one permutation of the possible values of (X_1, X_2, X_3). The mass in each rectangular cuboid is calculated with the trivariate joint distribution. Fitting a copula to the trivariate joint distribution means finding the parameter of the copula such that the mass in each rectangular cuboid implied by the copula matches the mass defined by the trivariate joint distribution, for instance, $\mathbb{P}(X_1 = 0, X_2 = 0, X_3 = 0) = C_R(p_1, p_2, p_3)$ where R denotes the parameter of the copula and p_i denotes the marginal of the i-th variable.

Fitting a copula to a joint Bernoulli distribution in a dimension higher than 2 is problematic. For example, not every trivariate Bernoulli distribution can be represented by the Normal copula. The following example illustrates this.

Example 2. *Let (X_1, X_2, X_3) be a trivariate Bernoulli distribution with the corresponding joint mass distribution:*

- $\mathbb{P}(X_1 = 0, X_2 = 0, X_3 = 0) = 0.02$
- $\mathbb{P}(X_1 = 1, X_2 = 0, X_3 = 0) = 0.35$
- $\mathbb{P}(X_1 = 0, X_2 = 1, X_3 = 0) = 0.11$
- $\mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 0) = 0.02$
- $\mathbb{P}(X_1 = 0, X_2 = 0, X_3 = 1) = 0.22$
- $\mathbb{P}(X_1 = 1, X_2 = 0, X_3 = 1) = 0.01$
- $\mathbb{P}(X_1 = 0, X_2 = 1, X_3 = 1) = 0.05$
- $\mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 1) = 0.22$

This trivariate Bernoulli distribution corresponds to the marginals of $\mathbb{P}(X_1 = 0) = 0.4$, $\mathbb{P}(X_2 = 0) = 0.6$, and $\mathbb{P}(X_3 = 0) = 0.5$. Intuitively, fitting a Normal copula into this trivariate distribution is likely to be problematic due to the very little mass of 0.02 in the rectangular cuboid corresponding to $\mathbb{P}(X_1 = 0, X_2 = 0, X_3 = 0)$.

The parameter of the Normal copula that best fits this trivariate Bernoulli distribution can be estimated via a minimization problem with the objective of minimizing the distance between the joint distribution of (X_1, X_2, X_3) and (Y_1, Y_2, Y_3), the joint discrete distribution implied by the Normal copula. The distance between the two is defined by the squared difference distance. For this example, the elements of the parameter of the Normal copula, the correlation matrix Σ, that minimizes the distance are $\rho_{12} = -0.0624$, $\rho_{13} = -0.4174$, and $\rho_{23} = 0.4174$. The trivariate discrete distribution of (Y_1, Y_2, Y_3) is as follows:

- $\mathbb{P}(Y_1 = 0, Y_2 = 0, Y_3 = 0) = 0.1016$
- $\mathbb{P}(Y_1 = 1, Y_2 = 0, Y_3 = 0) = 0.2646$
- $\mathbb{P}(Y_1 = 0, Y_2 = 1, Y_3 = 0) = 0.0322$
- $\mathbb{P}(Y_1 = 1, Y_2 = 1, Y_3 = 0) = 0.1016$
- $\mathbb{P}(Y_1 = 0, Y_2 = 0, Y_3 = 1) = 0.1290$
- $\mathbb{P}(Y_1 = 1, Y_2 = 0, Y_3 = 1) = 0.1048$
- $\mathbb{P}(Y_1 = 0, Y_2 = 1, Y_3 = 1) = 0.1372$
- $\mathbb{P}(Y_1 = 1, Y_2 = 1, Y_3 = 1) = 0.1290$

It is visible that the joint distribution of (X_1, X_2, X_3) *is different from* (Y_1, Y_2, Y_3)*'s. However,* (Y_1, Y_2, Y_3) *does recover some properties of* (X_1, X_2, X_3) *well. For example,* $\mathbb{P}(Y_1 = 0) = 0.4$*,* $\mathbb{P}(Y_2 = 0) = 0.6$*, and* $\mathbb{P}(Y_3 = 0) = 0.5$*. Furthermore, the bivariate marginals of* (X_1, X_2, X_3) *and* (Y_1, Y_2, Y_3) *are also similar:*

- $\mathbb{P}(X_1 = 0, X_2 = 0) = 0.24$
- $\mathbb{P}(X_1 = 0, X_3 = 0) = 0.13$
- $\mathbb{P}(X_2 = 0, X_3 = 0) = 0.37$
- $\mathbb{P}(Y_1 = 0, Y_2 = 0) = 0.2306$
- $\mathbb{P}(Y_1 = 0, Y_3 = 0) = 0.1338$
- $\mathbb{P}(Y_2 = 0, Y_3 = 0) = 0.3662$

where the difference is in the order of 10^{-3}*.*

With the information about the marginals, the problem of fitting a copula to a bivariate Bernoulli distribution in the previous section has 1 degree of freedom. Coincidentally, the bivariate Normal copula also has one parameter. Hence, the number of degree of freedom and the number of parameter to estimate matches. In the trivariate case, however, with information about the marginals, the problem has 4 degrees of freedom while a trivariate Normal copula has only three parameters to estimate: the pairwise correlations between each pair of the three marginals. In some cases, like in Example 2, this creates a problem.

3.3 Conditional Independence

The absence of an arc between two nodes in a BN implies (conditional) independence between the two variables the nodes represent. In the NPBN method, this corresponds to zero (conditional) rank correlation between the two variables.

However, the (conditional) independence inferred by a Normal copula does not always correspond to (conditional) independence of the corresponding discrete Bernoulli variables the Normal copula models. The following example illustrates this problem.

Example 3. *Let* (Y_1, Y_2, Y_3) *be a trivariate Bernoulli distribution with binary margins and* $\mathbb{P}(Y_1 = 0) = 0.6$*,* $\mathbb{P}(Y_2 = 0) = 0.45$*, and* $\mathbb{P}(Y_3 = 0) = 0.4$*. Let it be joined by the Normal copula with parameters:* $\rho_{12} = 0.7$*,* $\rho_{13} = -0.8$*, and* $\rho_{23} = \rho_{12} \cdot \rho_{13} = -0.56$*. This choice entails the conditional correlation of variable* Y_2 *and* Y_3 *given* Y_1*,* $\rho_{23|1}$*, to equal zero which means the Normal copula implies that the variable* Y_2 *and* Y_3 *are independent given variable* Y_1*.*

However, the corresponding conditional on variable Y_1 *discrete distributions implied by this Normal copula are not independent as follows:*
$\mathbb{P}(Y_2 = 0, Y_3 = 0|Y_1 = 0) = 0.0806 \neq 0.1061 = \mathbb{P}(Y_2 = 0|Y_1 = 0)\mathbb{P}(Y_3 = 0|Y_1 = 0)$
$\mathbb{P}(Y_2 = 0, Y_3 = 1|Y_1 = 0) = 0.5618 \neq 0.5363 = \mathbb{P}(Y_2 = 0|Y_1 = 0)\mathbb{P}(Y_3 = 1|Y_1 = 0)$
$\mathbb{P}(Y_2 = 1, Y_3 = 0|Y_1 = 0) = 0.0845 \neq 0.0590 = \mathbb{P}(Y_2 = 1|Y_1 = 0)\mathbb{P}(Y_3 = 0|Y_1 = 0)$
$\mathbb{P}(Y_2 = 1, Y_3 = 1|Y_1 = 0) = 0.2731 \neq 0.2986 = \mathbb{P}(Y_2 = 1|Y_1 = 0)\mathbb{P}(Y_3 = 1|Y_1 = 0)$

and

$$\mathbb{P}(Y_2 = 0, Y_3 = 0|Y_1 = 1) = 0.1056 \neq 0.1214 = \mathbb{P}(Y_2 = 0|Y_1 = 1)\mathbb{P}(Y_3 = 0|Y_1 = 1)$$
$$\mathbb{P}(Y_2 = 0, Y_3 = 1|Y_1 = 1) = 0.0558 \neq 0.0400 = \mathbb{P}(Y_2 = 0|Y_1 = 1)\mathbb{P}(Y_3 = 1|Y_1 = 1)$$
$$\mathbb{P}(Y_2 = 1, Y_3 = 0|Y_1 = 1) = 0.6467 \neq 0.6309 = \mathbb{P}(Y_2 = 1|Y_1 = 1)\mathbb{P}(Y_3 = 0|Y_1 = 1)$$
$$\mathbb{P}(Y_2 = 1, Y_3 = 1|Y_1 = 1) = 0.1919 \neq 0.2077 = \mathbb{P}(Y_2 = 1|Y_1 = 1)\mathbb{P}(Y_3 = 1|Y_1 = 1)$$

4 The Latency Time Model Construction

4.1 Fitting Parametric Distributions to the Continuous Variables

Three of the eight variables in the model are continuous variables. One interesting question regarding these continuous variables is whether or not it is possible to fit a known parametric distribution into each of these three variables. The advantage of fitting a parametric distribution to a continuous variable is the ability, in the future, to conditionalize the variable on values which have not been observed in the data.

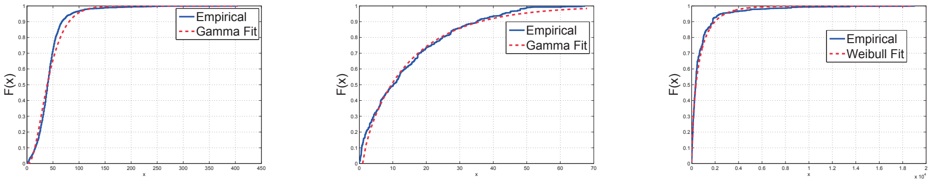
The latency time and the distance to the contractor's post variables are, each, fitted with a Gamma distribution and the distance to the nearest level crossing is fitted with a Weibull distribution. These distributions are the best parametric distributions to model their corresponding continuous variables that are currently supported by the software UNINET.

Figure 6 shows the empirical distribution (solid blue) and the parametric estimates (dashed red). Three goodness of fits test, the KS-test, the Anderson-Darling (AD) test (Anderson and Darling (1952)), and the Chi-Square (CS) test, are performed to see whether each parametric estimation can be used to represent its corresponding continuous variable or not. However, all three tests do not accept any of the hypothesis that these three variables can be represented with each of their corresponding parametric distribution.

Consequently, in this paper, none of the above parametric continuous distribution is used in the model. In practice, if an unobserved conditioning value is encountered, an adjustment needs to be made to the model. For example, if the conditioning value is larger than the maximum value of the empirical distribution, the variable will be conditionalized on its maximum value.

4.2 The Normal Copula Validation

The discussion in Section 3 should not be interpreted as that the NPBN method cannot be used in the presence of discrete variables. Instead, it accentuates the importance of the Normal copula validation before applying the method.



(a) Fitting the latency time distribution.

(b) Fitting the distance to contractor's post distribution.

(c) Fitting the distance to level crossing distribution.

Figure 6: Fitting the three continuous variables with parametric distributions.

First of all, the parameter of the Normal copula, the correlation matrix, is computed using the maximum likelihood estimation from data. At the moment, this functionality is not available in UNINET. Therefore, one needs to manually compute the parameter value in another software and then inputs the result into UNINET.

Several algorithms have been developed to validate whether a copula can be used to model a set of data for purely continuous and purely discrete models. For this reason, the validation step in this paper is performed on two groups: the continuous variables group and the discrete variables group. The validation step for the continuous and discrete variables is for the future work.

Continuous Variables

To validate the Normal copula for the continuous variables, a goodness of fit test introduced by Breymann, et al. (2003) which is based on the Rosenblatt's probability integral transform (Rosenblatt (1952)) is used. The test is as follows:

Let **U** $= (\tilde{U}_1, \tilde{U}_2, \tilde{U}_3)$ be random vector with uniform margins. If a normal copula C_R represents the joint distribution of **U**, then U_1, U_2, and U_3 defined as:

$$\begin{aligned} U_1 &= \tilde{U}_1 \\ U_2 &= C_{\rho_{12}}(\tilde{U}_2|U_1) \\ U_3 &= C_R(\tilde{U}_3|U_1, U_2), \end{aligned}$$

where ρ_{12} is an element in the correlation matrix R representing the dependence between the first and the second variable, are uniform and independent. This means that the statistics:

$$V = \left(\Phi^{-1}(U_1)\right)^2 + \left(\Phi^{-1}(U_2)\right)^2 + \left(\Phi^{-1}(U_3)\right)^2, \tag{6}$$

where Φ^{-1} denotes the inverse standard normal distribution, has a Chi-square distribution with 3 degrees of freedom.

The vector **U** can be obtained through transforming every continuous variable in the data into a uniform using their empirical distributions or their fitted parametric distributions discussed in Section 4.1.

When the empirical distributions are used, the test does not reject the hypothesis that the Normal copula fits the data. The KS-Test and AD-Test do not reject the uniformity hypothesis of U_1, U_2, and U_3 with the smallest p-value being 0.4902. Also, the hypothesis that the distribution of V is not different from a Chi-square distribution with 3 degrees of freedom is not rejected with the KS-test with a p-value of 0.2828.

Unsurprisingly, when the parametric distributions as discussed in subsection 4.1 are used, the conclusion is different. First of all, each hypothesis that U_i is uniformly distributed is rejected by both the KS-Test and the AD-Test with the highest p-value being $2.84\text{e}{-7}$. And then, the hypothesis that V is distributed with Chi-square distribution with 3 degrees of freedom is also rejected by the KS-Test with p-value of $3.21\text{e}{-8}$. This is because in subsection 4.1, it is mentioned that the hypothesis that the three continuous variables can be represented by these parametric marginal distributions is not accepted. Therefore, this means that each $\tilde{U}_i$ in the vector **U** is not even uniform to begin with for this case. This confirms the decision not to use the fitted parametric distributions in subsection 4.1.

Discrete Variables

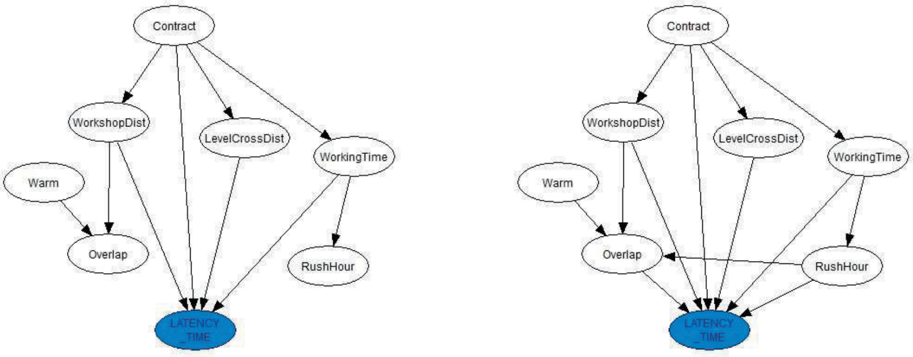
To test the Normal copula for the discrete variables, a simple comparison between the joint distribution of the discrete data and the joint distribution implied by the Normal copula is performed. Because there are five discrete variables, the joint discrete distribution has $2^5 = 32$ cells.

Calculating the squared difference between the number of observation in the data for each cell with the number of expected observation using the full Normal copula model without assuming any conditional independence, the standard Pearson's squared difference test yields a p-value of 0.0234. Even though technically the p-value is below the usual norm of 5%, visual observation of the two joint distribution reveals that the two distributions are actually very similar. The difference between each corresponding cell, in probability, is only in the order of 10^{-3} or smaller. The largest squared difference occurs in the cell corresponding to the situation where the contract type is PGO, there is no overlapping disruption, the temperature is below 25^oC, the incident occurs not during rush hour, and not during the repair team's working hours. In the data, 30.71% of all samples are of this situation, while the Normal copula models this situation with 29.87% of the TC disruptions.

Therefore, it is still believed that the Normal copula reasonably represents the empirical joint distribution.

4.3 The BN Structure

The latency time BN in this paper is a mixed BN with the presence of both discrete and continuous nodes. A number of algorithms has been developed to learn the structure of a BN from the data. The algorithm that is used is the hill-climbing greedy search in the space of all possible BN structures. This algorithm is a score-based algorithm which assigns a score to each possible structure of the BN and the structure that maximizes this score is chosen. In short, the score of a structure $\mathcal{G}$ is defined as the probability of the structure given the data $\mathcal{D}$, i.e. $\mathbb{P}(\mathcal{G}|\mathcal{D})$ (Magaritis. (2003)). This algorithm is actually developed for a purely discrete BN. Therefore, the continuous variables need to be discretized for the structure learning process. Due to the limited number of samples, each continuous variable cannot be discretized into a lot of states and so a number of four states per variable is chosen. In this paper, the algorithm is executed using the package `bnlearn` in R (Scutari (2010)).



(a) Learned BN Structure from Data.

(b) Intuitive BN Structure.

Figure 7: The Structure of the latency time Bayesian Network.

All BN figures in this paper are produced with the software UNINET. Figure 7(a) presents the result. Originally, the hill-climbing greedy search puts an arc originating from the node ”Overlap” to ”Warm”. This, however, is not intuitive because whether the temperature is warm or not (the ”Warm” node) should affect the presence of an overlapping disruption (the ”Overlap” node), not the other way around. An arc reversal implies a different (conditional) independence statement carried on by the graph. The reversal is justified upon performing the test of independence between the variable ”Warm” and ”Working-Time” which confirms the independence between the two variables, thus the structure in Figure 7(a).

The BN structure in Figure 7(a) appears to model the variables well because it looks rather close to the expected structure presented in Figure 7(b) with a few arcs missing. The structure in Figure 7(b) is obtained simply by deducting how it should be with the information of what each variable represents. This also means that the data models the variables well. The ”missing” arcs in the BN in Figure 7(a) may be an artefact of the discretization.

The Akaike information criterion (AIC) is going to be used to measure which structure models the data better. The AIC is defined as

$$AIC = 2k - 2\ln(L)$$

where k denotes the number of parameters and L is the maximized likelihood for the model. The structure with the lower AIC is the better structure to model the data. After calculating the maximum likelihood value for both structures, the AIC score for the structure in Figure 7(a) is 9220.6 while the score for the structure in Figure 7(b) is 9216.9. For this reason, the structure in Figure 7(b) is chosen.

4.4 The Model Validation

The latency time NPBN model with structure depicted in Figure 7(b) needs to be validated first. To test the validity of this model, a validation test is going to be performed. For each sample, conditionalization on all variables but the latency time with the NPBN model is performed, resulting in a conditional latency time distribution for the sample. From this

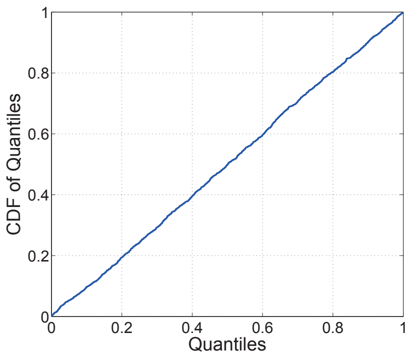


Figure 8: The distribution of the quantiles.

conditional distribution, the quantile that corresponds to the observed latency time is computed. The BN models the data well if these quantiles are distributed uniformly on $(0, 1)$.

Figure 8 shows the result where the distribution of the quantiles is uniformly distributed. This conclusion is supported by the KS-Test too with a p-value of 0.7773. Therefore, it can be said that the latency time NPBN model is a good model for the latency time.

4.5 Case Study

How the model can be used in real life application is going to be presented next. Figure 9 presents the *unconditional* latency time BN where the empirical distributions of each variable is shown in the nodes. It shows that when it is only known that there is a TC problem, the mean of the latency time is 43.3 minutes with standard deviation of 31 minutes.

Now, suppose a TC problem occur somewhere in the Dutch railway network. Let it be known that there is an overlapping disruption, the incident occurs in a place with the old OPC contract, the temperature is higher than 25^oC, the incident occurs not during rush hour, not during the repair team's working hours, it is 45 km away from the contractor's post, and the site is 250 meter from the nearest level crossing. Intuitively, this is a bad condition for the latency time, i.e. the latency time should be longer than the "average" (unconditional) situation.

Conditioning the BN, which is performed rapidly in UNINET, on this information yields the *conditioned* BN as shown in Figure 10. Figure 10 shows that the model, indeed, concludes longer latency time under this condition. The mean of the latency time is 65.3 minutes and the standard deviation is 51.9 minutes.

Conditionalization can, of course, also be performed when only part of the information is available. For example, let another TC problem occurs with the following known information: it occurs in a place with the new PGO contract, during the repair team's working

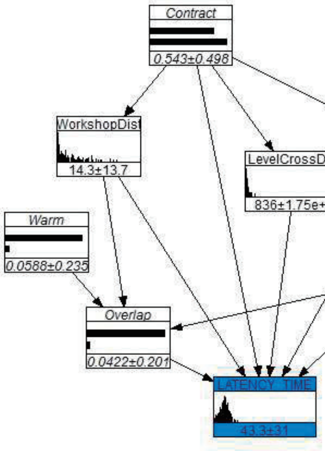


Figure 9: The unconditional latency time BN.

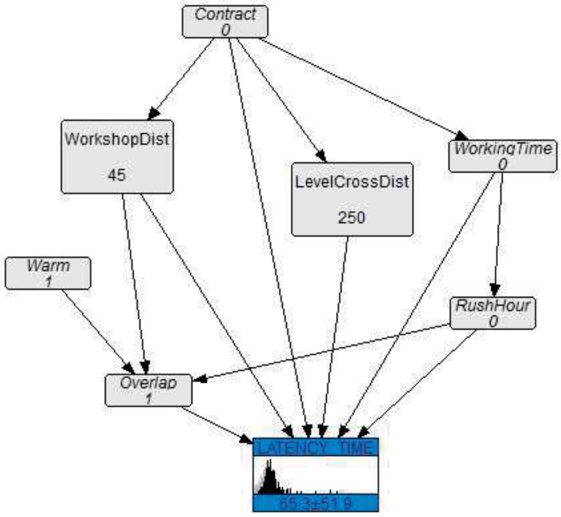


Figure 10: The conditioned latency time BN for the first case.

hours but not during the rush hour. Intuitively, this is a good condition for the latency time, i.e. the latency time should be shorter than the ”average” situation. Conditioning the BN model on this information yields the conditioned BN as shown in Figure 11.

The model, indeed, yields shorter latency time under this condition. The mean of the latency time is 35.7 minutes and the standard deviation is 23.9 minutes.

Validating the performance of the model in practice has not been done yet at this point

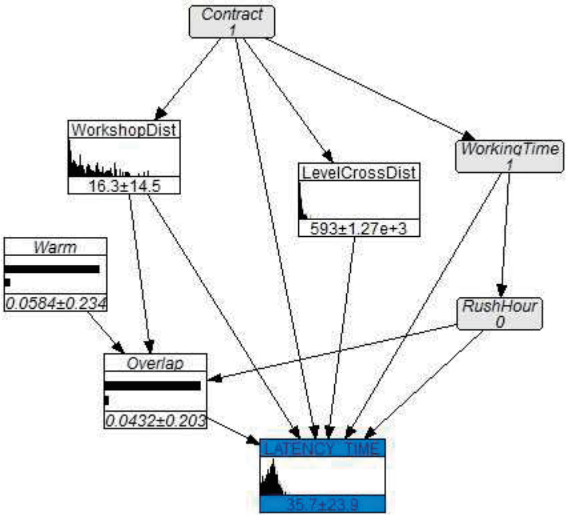


Figure 11: The conditioned latency time BN for the second case.

due to the lack of data for validation. This is certainly one part of the future work and will be discussed further in the subsequent section. Nevertheless, through observing the performance of the NPBN method with the two examples above, the strength of the method can be appreciated. The conditioning step is performed rapidly and the conclusion drawn from it on the latency time is reasonable.

5 Conclusions and Future Work

At this point, the disruption length model is certainly still incomplete. In this paper, a model of one part of the model, the latency time, has been constructed. The model construction reveals that some of the influencing variables of the latency time are discrete, hence resulting in a mixed discrete-continuous BN model. This raises some concerns because the NPBN method is originally designed to work with continuous variables. Fortunately, the Normal copula models the discrete and continuous part of the model well, hence the NPBN method with the Normal copula can still be performed. Execution of the constructed latency time models reveals that the NPBN method yields reasonable conclusions, thus an approving sign that the NPBN method is a promising probabilistic model of disruption length.

The mixed discrete-continuous nature of the model raises some interesting questions on the theoretical side. As mentioned in subsection 4.2, validation of the Normal copula for the mixed discrete-continuous part of the model is one of them. Searching the structure of a BN model with the presence of mixed discrete-continuous nodes is also of interest.

A model extension to include the repair time is certainly planned for the near future. One big problem that currently prevents the repair time model development to progress in this project is the data source's non-informativeness on the technical cause of the disruption. A technical disruption, like the one caused by a TC problem to name one, can occur because of several different causes. A different cause leads to a different action which leads to a different repair time. Unfortunately, while the repair time concerning all technical disruptions is available in the data, the data set does not reach the preferred detailed level in the disruption's technicality to be able to construct a repair time model.

Two strategies are in store to tackle this problem and both involve defining several different common causes of each technical disruption. The first one involves digging the data set for more information deeper, hopefully resulting in an improved data to construct the repair time model. Secondly, a structured expert judgment exercise can be performed to fill in this information gap in the data. A group of experts on the related technical railway components needs to be assembled. From this group of experts, the repair time length based on the different common causes is going to be elicited.

Another plan for the future work is the model validation step. At the moment, new data with better quality information is being gathered in the Netherlands. Later on, when enough data has been collected, the model is going to be validated on the new data by comparing the prediction supplied by the NPBN method and the reality observed in the field.

The new data will also be useful in determining which value of the conditional distribution of disruption length will be used as a prediction. For instance, one can choose the median of the conditional distribution as the prediction. However, this choice means that there is a 50% chance the actual length will be longer than this value according to the model. Consequently, one may opt to be on the safer side by choosing a higher quantile of the distribution as the prediction.

Acknowledgements

This work is funded by ExploRail, a partnership programme of the Dutch Technology Foundation STW and ProRail, project no. 12257: ”Smart information and decision support for railway operation control centres (SmartOCCR)”. The authors would like to acknowledge Dirk Kes and André Duinmeijer from ProRail for the supervision and access to the dataset.

References

Anderson, T.W., Darling, D.A. 2002. “Asymptotic theory of certain ”goodness-of-fit” criteria based on stochastic processes”, *Annals of Mathematical Statistics*, vol. 23, pp. 193-212.

Breymann, W., Dias, A., Embrechts, P. 2002. “Dependence Structures for Multivariate High-Frequency Data in Finance”, *Quantitative Finance*, vol. 3, pp. 1-14.

Carley, H. 2002. “Maximum and Minimum Extensions of Finite Subcopulas”, *Comm. Stats. Theory Methods*, vol. 31, pp. 2151-2166.

Clayton, D.G. 1978. “model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence”, *Biometrika*, vol. 65, pp. 141-151.

Genest, C., Nešlehová, J. 2007. “A primer on copulas for count data”, *The Astin Bulletin*, vol. 37, pp. 475-515.

Hanea, A., Kurowicka, D., Cooke, R. 2006. “Hybrid method for quantifying and analyzing Bayesian belief nets”, *Quality and reliability Engineering International*, vol. 22(6), pp. 709-729.

Hanea, A., Kurowicka, D., Cooke, R., Ababei, D. 2010. “Mining and visualizing ordinal data with non-parametric continuous BBNs”, *Computational Statistics & Data Analysis*, vol. 54(3), pp. 668-687.

Kurowicka, D., Cooke, R. 2005. “Distribution-free continuous Bayesian belief nets”, In: Wilson, A., Limnios, M., Keller-McNulty, S. & Armijo, Y. (eds.), *Modern Statistical and mathematical Methods in Reliability*, pp. 309–323, Singapore: World Scientific Publishing.

Magaritis, D. 2008. *Learning Bayesian Network Model Structure from Data*, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

Rosenblatt, M. 1952. “Remarks on a multivariate transformation”, *The Annals of Mathematical Statistics*, vol. 23, pp. 470-472.

Scutari, M. 2010. “Learning Bayesian Networks with the bnlearn R Package”, *Journal of Statistical Software*, vol. 35(3), pp. 1-22.

Sklar, A. 1959. “Fonctions de repartition ‘a n dimensions et leurs marges”, *Publ. Inst. Statist. Univ. Paris*, vol. 8, pp. 229-231.

Swaroop, R., Brownlow, J.D., Ashworth, G.R., Winter, W.R. 1980. *Bivariate Normal Conditional and Rectangular Probabilities: A Computer Program with Applications*, NASA Technical Memorandum, Edwards, California.

Zilko, A.A., Hanea, A.M., Kurowicka, D., Goverde, R.M.P. 2014. “Non-Parametric Bayesian Network to Forecast Railway Disruption Lengths”, In: *The Second International Conference on Railway Technology: Research, Development and Maintenance (Railways 2014)*, Ajaccio, France.