Aurelius A. Zilko, Dorota Kurowicka, Rob M.P. Goverde Modeling railway disruption lengths with Copula Bayesian Networks

Transportation Research Part C 68 (2016) 350–368

The method considers the factors inﬂuencing the length of a disruption and models the dependence between them to produce a prediction. As an illustration, a model for track circuit (TC) disruptions in the Dutch railway network is presented in this paper. Factors inﬂuencing the TC disruption length are considered and a disruption length model is constructed. We show that the resulting model’s prediction power is sound and discuss its real-life use and challenges to be tackled in practice.

1. Introduction

Shorter incidents, usually referred to as **disturbances**, may require only timetable adjustment while longer incidents, usually referred to as **disruptions**, may additionally require rolling stock and crew adjustment. In this paper, **we do not distinguish the difference** between disturbance and disruption so the term disruption is used when referring to these unexpected incidents, regardless of the length.

In practice, a series of updated disruption length predictions are made as more information about the disruption is gathered. When a disruption occurs, ﬁrstly a rough prediction based on history is made. This is called the ‘‘P1” prediction and is taken to be the average of a given disruption length in the past. In the meantime, the mechanics are informed about the disruption and are tasked to repair the problem. From now on, the OCCR (**диспетч. центр**) is in close communication with the mechanics. After arriving at the site, the mechanics have 15 min to diagnose the problem after which they are required to make a prediction, based on their own judgment, regarding the repair time. This mechanics’ prediction is called the ‘‘P2” prediction. The mechanics are allowed to update the prediction later on and this updated prediction is called the ‘‘P2a” prediction. Finally, when a ﬁnal prediction can be made, the mechanics are required to update the OCCR with the so-called ‘‘P3” prediction.

Thus in current practice, the uncertainty in disruption length is handled by means of a series of predictions based on the mechanics’ expertise and judgment. One other way to tackle the uncertain disruption length problem is by representing the disruption length with a probability distribution. Having such distribution allows us to generate random samples of disruption length. This approach is relatively new in railway operation but has been used in several earlier studies in highway trafﬁc engineering.

In this paper, disruption length is the center of attention. The uncertainty of disruption length is going to be modeled with a probability distribution from historical data. Moreover, several inﬂuencing factors of disruption length are considered. The goal is to construct a dependence model between the disruption length and these inﬂuencing factors. When a disruption occurs, the model is conditionalized on the realization of the inﬂuencing factors resulting in a conditionalized disruption length distribution. This conditionalized distribution represents the disruption length specialized to a speciﬁc situation. Then, a disruption length prediction is made from this distribution.

В этой статье продолжительность сбоя находится в центре внимания. Неопределенность продол-жительности будет смоделирована с помощью распределения вероятностей на основе истории-ческих данных. Кроме того, рассматриваются несколько факторов, влияющих на длину сбоя. Цель состоит в том, чтобы построить модель зависимости между продолжительностью сбоя и этими влияющими факторами. Когда происходит сбой, модель учитывает реализацию влияющих факторов, что приводит к условному распределению продолжительности. Это условное распределение представляет собой длину сбоя, адаптированную к конкретной ситуации. Затем по этому распределению делается прогноз продолжительности сбоя.

To do this, a proposal made by Zilko et al. (2014) is followed where the dependence model is constructed using the Copula Bayesian Network (**связанная сеть Байеса**). The Bayesian Network (BN) technique has been used in transportation research ﬁeld for several different studies.

BN modeling consists of two parts: the **graphical structure** that represents (conditional) independence in the model and **the conditional probabilities** between the variables to specify the rest of the relationships. Моделирование BN состоит из двух частей: графической структуры, которая представляет (условную) независимость в модели, и условных вероятностей между переменными для определения остальных взаимосвязей.

There are different types of BNs depending on how the conditional probabilities are modeled. When all the variables are discrete, as in the case of the three studies mentioned above, the conditional probabilities are modeled using conditional probability tables (CPT). In the case of continuous variables in Gregoriades and Mouskos (2013), the variables are discretized to obtain a fully discrete model. When all the variables are continuous and Normally distributed, the Normal Bayesian Network (which uses the multivariate Normal distribution) can be used.

In this paper, the conditional probabilities between the variables are represented with copula (more on copula in Section 4). Copula is a very useful model for the dependence between continuous variables as this allows the separation of marginal distributions and the dependence. We are interested in the use of copula because our variable of interest, disruption length, is continuous. В этой статье условные вероятности между переменными представлены связкой (подробнее о связке в разделе 4). Связка (копула) — очень полезная модель зависимости между непрерывными переменными, поскольку она позволяет разделить предельные (?) распределения и зависимость. Мы заинтересованы в использовании связки, потому что интересующая нас переменная - продолжительность сбоя - непрерывна.

Moreover, because the output of the model is a probability distribution function, this also gives the OCCR full control regarding which value they want to take as a prediction. Do they want to be more conservative by choosing a value in the upper quantile of the distribution? Or do they want to be more optimistic by choosing a value in the lower quantile of the distribution? Более того, поскольку выход модели представляет собой функцию распределения вероятностей, это также дает Диспетчерам (OCCR) полный контроль над тем, какое значение они хотят принять в качестве прогноза. Они хотят быть более консервативными, выбирая значение в верхнем квантиле распределения? Или они хотят быть более оптимистичными, выбирая значение в нижнем квантиле распределения?

A Copula Bayesian Network model needs to be constructed for each disruption type. To help the model construction, a user-friendly and computationally-efﬁcient software called UNINET which implements the algorithm of the Copula Bayesian Network will be used. This software was developed at Delft University of Technology and is available at www.lighttwist.net/ wp/uninet.

First, to construct the model, 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 mechanics need to get to the disrupted site. The repair time is the length of time they need to repair the problem. Во-первых, для построения модели необходимо разделить продолжительность сбоя на два взаимоисключающих определения времени: время задержки (ожидания) и время ремонта . Время задержки — это период времени, в течение которого механики должны добраться до аварийного участка. Время ремонта — это время, необходимое для устранения проблемы.

2. The track circuit

3. Data analysis

3.1. Factors influencing the latency time **Описываются распределения для разных факторов**

3.1.1. Time Обсуждается распределение времени задержки при различных ситуациях.

3.1.2. Location

Местоположение сбоя. Две характеристики: расстояние до ближайшей станции обслуживания (механиков) и плотность потока. Эта переменная (плотность потока) по-прежнему не оказывает существенного влияния на время ожидания в сочетании с другими переменными. В результате эта переменная исключена из модели. Вводится третья переменная, которая представляет доступность к месту аварии: (3) расстояние до ближайшего железнодорожного переезда. Чтобы попасть на отказавший участок, механики должны припарковать свои машины (обычно, возле ближайшего переезда) и идти на участок. Кроме того, также учитывается другое свойство местоположения, которое является бинарной дискретной переменной, - (4) тип контракта. Вскоре будет объяснено, почему эта переменная рассматривается как свойство местоположения.

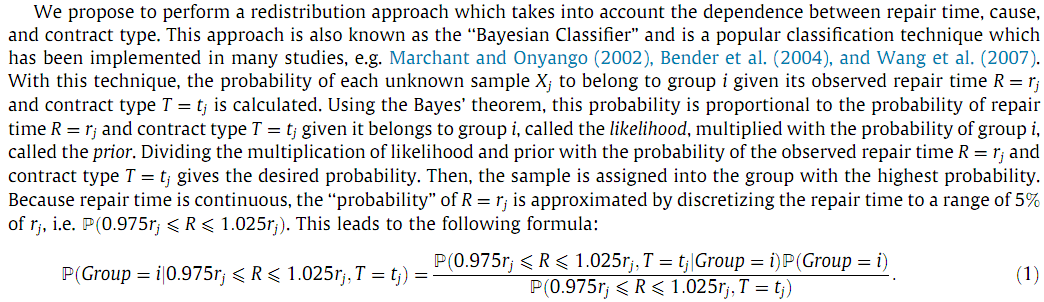
Контракторы – различные фирмы, обслуживающие инфраструктуру.

3.1.3. Weather and overlapping disruptions

3.2. Factors inﬂuencing the repair time **Описываются распределения для разных факторов**

3.2.1. Contract type. 3.2.2. The cause.

Стр. 358



Мы предлагаем использовать метод перестройки распределения (перераспределения), учитывающий зависимость между продолжительностью ремонта, причиной и типом договора. Этот метод также известен как «Байесовский классификатор» и является популярным методом распределения (классификации), который применялся во многих исследованиях, например Маршан и Оньянго (2002), Бендер и др. (2004) и Wang et al. (2007).

В этом методе рассчитывается вероятность каждой неизвестной выборки (отсчета) Xi (принадлежит группе i) при данном наблюдаемом времени ремонта *R* и типе контракта *Т*. Согласно теореме Байеса, эта вероятность пропорциональна вероятностям величин *R* и *T*, называется **правдоподобие,** умножается на вероятность группы *i* , называемую априорной.

Искомая вероятность получается путем деления произведения (правдоподобие х априорная вероятность) на вероятности величин *R* и *T*. Затем выборку относят к группе с наибольшей вероятностью.

Время ремонта – непрерывная величина, поэтому «вероятность» R аппроксимируется путем дискретизации времени ремонта на интервалы шириной 5% от ri . Это приводит к следующей формуле: (1)

With this technique, an unknown sample with short repair time is more likely to be redistributed into a group with shorter repair time, and vice versa.

При использовании этого метода неизвестная выборка - малое время ремонта - с большей вероятностью будет перераспределена в группу с более коротким временем, и наоборот.

Table 1 presents some information about the repair time of each group before and after redistribution of unknowns. Groups with high proportion (Group 2 and Group 5) receive the most assigned unknowns while groups with very low proportion (Group 1 and Others) receive no assigned unknowns. Fig. 8(b) shows the repair time distribution of Group 2, the group that receives the most redistributed unknowns, before (dashed) and after (solid) redistribution. The redistribution

appears to slightly shift the repair time distribution to the left.

В таблице 1 представлены некоторые сведения о времени ремонта по каждой группе до и после перераспределения неизвестных. Группы с высокой долей (Группа 2 и Группа 5) получают наибольшее количество назначенных неизвестных, в то время как группы с очень низкой долей (Группа 1 и Другие) не получают назначенных неизвестных. На рис. 8(b) показано распределение времени восстановления группы 2, группы, которая получает наибольшее количество перераспределенных неизвестных, до (штриховая линия) и после (сплошная линия) перераспределения. Перераспределение немного сдвигает распределение времени ремонта влево.

**4. The Copula Bayesian Network and the disruption length model**

4.1. The Bayesian Network and its structure

Зилко и др. (2014) предлагает использовать Copula Bayesian Network в качестве модели зависимо-сти между переменными, обсуждаемыми в разделе 3, чтобы сделать вывод о продолжительности сбоя с учетом реализаций влияющих переменных. Как следует из названия, байесовская сеть (BN) используется для графического представления зависимости между переменными. BN представ-ляет собой ориентированный ациклический граф, состоящий из узлов и дуг, представляющих соответственно переменные и поток влияний между переменными.

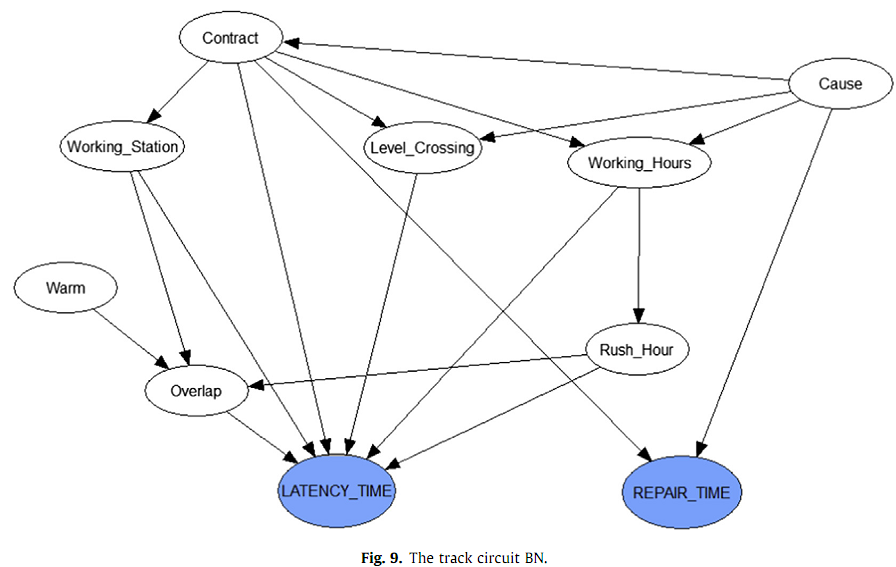


Fig. 9 presents the BN structure of the TC disruption length model in this paper. The ten nodes in the structure correspond to the ten variables in the model and the arcs represent the ﬂow of inﬂuence between the variables. The absence of an arc between two nodes indicates (conditional) independence between the variables the two nodes represent. For instance, the variable cause and latency time are conditionally independent given the variables contract, distance to the nearest level crossing (from hereon, this variable’s name is ortened as ’level crossing’), and working hours. This means that once information about the contract, level crossing, and working hours is known, the variable cause has no effect on the latency time. This is reasonable because while cause does not have direct inﬂuence on latency time, a certain cause occurs more likely in certain areas at certain time and this inﬂuences the latency time. Therefore, once information about location (represented by contract and level crossing) and time (represented by working hours) is available, the information about cause becomes irrelevant to the latency time length.

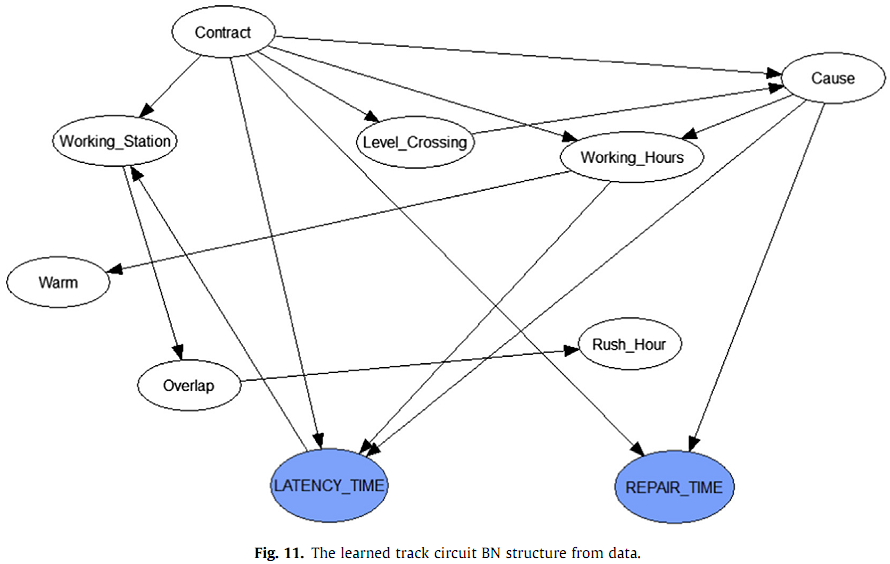
На рис. 9 представлена структура BN модели длины отказа рельсовой цепи (в этой статье). Десять узлов в структуре соответствуют десяти переменным в модели, а дуги представляют поток влияний между переменными. Отсутствие дуги между двумя узлами указывает на (условную) независимость между переменными, которые представляют два узла. Например, переменная «причина» и задержка условно независимы при данной переменной «контракт», расстояния до ближайшего переезда (отсюда и название этой переменной пишется как «железнодорожный переезд») и рабочих часов. Это означает, что когда информация о контракте, железнодорожном переезде и рабочем времени известна, переменная «причина» не влияет на время задержки. Это разумно, потому что, хотя «причина» не оказывает прямого влияния на задержку, определенная «причина» с большей вероятностью возникает в определенных областях в определенное время, и это влияет на время задержки. Следовательно, как только информация о местоположении (представленном «контрактом» и «железнодорожным переездом») и времени (представленном «рабочими часами») становится доступной, информация о «причине» становится нерелевантной для продолжительности задержки.

The structure in Fig. 9 is obtained based on the deﬁnition of the variables and the analysis of the data. Another way to determine the structure is by learning from data. A number of algorithms have been developed for this purpose when the variables at hand are all discrete. One of the algorithms is the hill-climbing greedy search in the space of all possible BN structure. It is a score-based algorithm which assigns a score to each possible BN structure based on the data and the structure that maximizes the score is chosen. The score of a structure G is deﬁned as the probability of the structure given the data D, i.e. P(G/D) Margaritis (2003). Because our model consists of six discrete and four continuous variables, we need to discretize the continuous variables to be able to use the algorithm. Because of the limited number of samples, the continuous variables are discretized into four discrete states. Performing the hill climbing search results in a BN structure as presented in Fig. 11. The structure is obtained by executing the algorithm using the package **bnlearn** in R. Scutari (2010).

Структура на рис. 9 получена на основе определения переменных и анализа данных. Другой способ определить структуру — это обучение на основе данных. Для этой цели был разработан ряд алгоритмов, когда все имеющиеся переменные являются дискретными. Одним из алгоритмов является жадный поиск (с восхождением на вершину) в пространстве всех возможных структур BN. Это алгоритм, основанный на оценке, который присваивает оценку каждой возможной структуре BN на основе данных, и выбирается структура, которая максимизирует оценку. Оценка структуры G определяется как вероятность структуры с учетом данных D, то есть P(G/D) Margaritis (2003). Поскольку наша модель состоит из шести дискретных и четырех непрерывных перемен-ных, нам необходимо дискретизировать непрерывные переменные, чтобы иметь возможность использовать алгоритм. Из-за ограниченного числа выборок непрерывные переменные дискретизируются в четыре дискретных состояния. Выполнение поиска с восхождением на холм приводит к структуре BN, как показано на рис. 11. Структура получается путем выполнения алгоритма с использованием пакета **bnlearn** в R. Scutari (2010).

Resemblance between the structure in Figs. 9 and 11 can be observed even with a few ﬂipped and missing arcs. Margaritis (2003) studies the performance of the algorithm where it is not always able to fully recover the directionality of the arcs. Moreover, the result in Fig. 11 is obtained by discretizing the continuous variables. The missing arcs and reversed direction may be artefacts of the discretization.

Сходство структуры на рис. 9 и 11 можно наблюдать даже с несколькими перевернутыми и отсутствующими дугами. Маргаритис (2003) изучает характеристику алгоритма, когда он не всегда может полностью восстановить направленность дуг. Кроме того, результат на рис. 11 получается путем дискретизации непрерывных переменных. Отсутствие дуг и обратное направление могут быть результатами дискретизации.



Two models with both structures are constructed in the next subsection to see which structure is more suitable for our data.

В следующем подразделе построены две модели с обеими структурами, чтобы увидеть, какая структура больше подходит для наших данных.

4.2. The Copula Bayesian Network and copula **4.2. Связанная сеть Байеса и связь**

The TC Disruption Length BN is quantiﬁed with the Copula Bayesian Network. The method is introduced in Kurowicka and Cooke (2005) and extended in Hanea et al. (2006) and Hanea et al. (2010) to construct a dependence model between continuous variables. This method implements the use of copula as the backbone model of dependence. A copula is the n-dimensional joint distribution in the unit hypercube of n uniform random variables. The theorem of Sklar (1959) serves as the basis of the copula application. He states that any cumulative distribution function (X1; . . . ;Xn) denoted as F1;...;n can be rewritten in terms of the corresponding copula C as ……….

where F ðX Þ denotes the marginal distribution of the i-th variable.

Протяженность отказа ТС количественно определяется с помощью связанной байесовской сети. Этот метод представлен Kurowicka and Cooke (2005) и расширен Hanea et al. (2006) и Hanea et al. (2010) для построения модели зависимости между непрерывными переменными. Этот метод реализует использование связки в качестве базовой модели зависимости. Связка — это n-мерное совместное распределение в единичном гиперкубе n равномерных случайных величин. Теорема Склара (1959) служит основой применения копулы. Она утверждает, что любую накопительную функцию распределения (X1; . . . ;Xn), обозначаемую как F1;...;n , можно переписать в терминах соответствующей связи C как ,

где Fi(Xi) обозначает предельное распределение i-й переменной.

There are many different copula families. One that is of interest in this paper is the multivariate Normal, or Gaussian, copula C. многомерная нормальная связь

There are many different copula families. One that is of interest in this paper is the multivariate Normal, or Gaussian, copula C.

Существует множество различных семейств связок. Одна из них, представляющая интерес для этой статьи, — это многомерная нормальная, или гауссовская, связка C. Она определяется как:

,

where Ф-1 denotes the inverse cumulative distribution of a univariate standard normal distribution and Фсумм denotes the cumulative joint distribution of a multivariate normal distribution with zero mean and correlation matrix R. With the Copula Bayesian Network, the parameter R corresponds to the arcs in the BN structure. We are interested in this copula because it allows conditionalization to be computed rapidly, a very useful feature in the real-time decision making environment of the OCCR.

где Ф-1 обозначает обратное суммарное распределение одномерного стандартного нормального распределения, а Фсумм обозначает суммарное совместное распределение многомерного нормального распределения с нулевым средним и корреляционной матрицей R. В байесовской сети параметр R соответствует дугам в структуре BN. Мы заинтересованы в этой связке, потому что она дает условия для быстрого вычисления, что является очень полезной функцией в среде принятия решений диспетчерским центром в реальном времени.

Copula has been proven to be an attractive model for the dependence between continuous variables. In our case however, some of the inﬂuencing variables are discrete. This results in a mixed discrete–continious model that

Было доказано, что копула является привлекательной моделью зависимости между непрерывными переменными. Однако в нашем случае некоторые из влияющих переменных являются дискретными. Это приводит к смешанной дискретно-непрерывной модели, которую необходимо построить.

Unfortunately, the use of copula in discrete model is known to be troublesome. For this reason, Zilko and Kurowicka (in press) investigate the matter further for the latency time model. A more complex copula model construction (with the so-called copula-vine approach) is also considered to perfectly represent the discrete part of the model by adding more parameters.

К сожалению, известно, что использование копулы в дискретной модели затруднительно. По этой причине Зилко и Куровицка (в печати) дополнительно исследуют этот вопрос для модели времени задержки. Также считается, что более сложная конструкция модели связки (с так называемым методом связки-лозы) идеально представляет дискретную часть модели за счет добавления большего количества параметров.