**A Hybrid Bayesian Network Model for Predicting Delays in Train Operations**

Javad Lessan, Liping Fu, Chao Wen

Computers & Industrial Engineering,

Volume 127 Pages 1214-1222 January 2019

DOI: https://doi.org/10.1016/j.cie.2018.03.017

**Abstract**

We present a Bayesian network-(BN) based train delay prediction model to tackle the complexity and dependency nature of train operations. Three different BN schemes, namely, heuristic hill-climbing, primitive linear and hybrid structure, are investigated using real-world train operation data from a high-speed railway line. We first use historical data to rationalize the dependency graph of the developed structures. Each BN structure is then trained with the gold standard *k*-fold cross validation approach to avoid over-fitting and evaluate its performance against the others. Overall, the validation results indicate that a BN-based model can be an efficient tool for capturing superposition and interaction effects of train delays. However, a well-designed hybrid BN structure, developed based on domain knowledge and judgments of expertise and local authorities, can outperform the other models. We present a performance comparison of the predictions obtained from the hybrid BN structure against the real-world benchmark data. The results show that the proposed model on overage can achieve over 80% accuracy in predictions within a 60-min horizon, yielding low prediction errors regarding mean absolute error (MAE), mean error (ME) and root mean square error (RMSE) measures.

Мы представляем основанную на байесовской сети (BN) модель прогнозирования задержки поездов, чтобы справиться со сложностью и зависимым характером движения поездов. Три различные схемы BN, а именно: эвристический метод восхождения, примитивная линейная и гибридная структуры - исследуются с использованием реальных данных о поездной работе на высокоскоростной железнодорожной линии. Сначала мы используем исторические данные для рационализации графа зависимостей разработанных структур. Каждая структура BN затем обучается с использованием золотого стандарта перекрестной проверки k-fold (чтобы избежать переобучения) и оценить ее производительность по сравнению с другими. В целом, результаты проверки показывают, что модель на основе BN может быть эффективным инструментом для учета эффектов суперпозиции и взаимодействия задержек поездов. Вместе с тем, хорошо спроектированная гибридная структура BN, разработанная на основе знаний предметной области и суждений экспертов и местных властей, может превзойти другие модели. Мы даем сравнение характеристик прогнозов, полученных из гибридной структуры BN, с реальными референтными данными. Результаты показывают, что предложенная модель может обеспечить точность прогнозов более 80% (в среднем) в пределах 60-минутного горизонта, дает малые ошибки прогнозирования в отношении показателей средней абсолютной ошибки (MAE), средней ошибки (ME) и среднеквадратичной ошибки (RMSE).

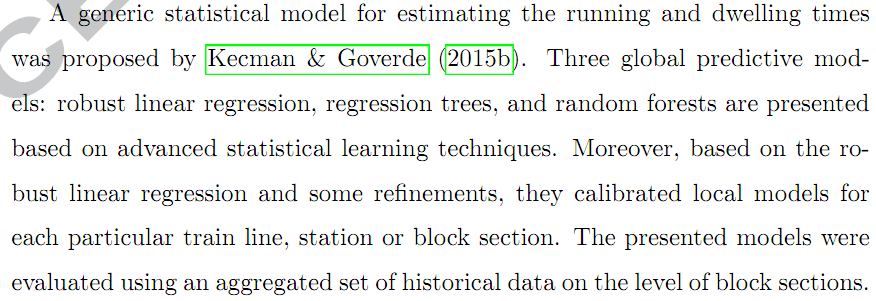
In large and dense network areas, however, the domain knowledge and expertise of local dispatchers must be supported by an advanced computational tool that can account for the interdependencies of train operations and interrelated delay factors. Creation of such an advanced tool has been hindered by two fundamental limitations. Firstly, methodologically, there has been a lack of models capable of simultaneously examining multiple components of delay incidents intertwined with stochastic operations and interaction effects. Secondly, technologically, there has been a need for collection and incorporation of massive train operation data. Recently, the integration of graph and probability theories led to the introduction of Bayesian networks (BNs) that enabled practitioners to overtake these limitations.

In this paper, we present three different BN designing architectures, namely, a heuristic, a naive, and a hybrid method, to represent the relationship and superposition of interdependent variables identified in the delay chain of trains. Using information obtained from historical data, we rationalize the contingency graph of the proposed BN structures. Next, we apply the gold standard k-fold cross-validation method to train and evaluate the proposed BNs. The hybrid BN structure, having a higher performance compared to the other models, is then tested against real-world benchmark data under different performance measures.

Авторы рационализируют вероятностный граф предложенной BN структуры с использованием информации, взятой из исторических данных. Затем используется метод взаимной проверки k-fold (k-fold cross-validation method) обучения и контроля предложенных BN. Гибридная BN структура имеет лучшие характеристики по сравнению с другими моделями. Она затем тестируется путем сравнения с реальными данными при различных характеристических показателях.

The main idea behind the hybrid structure introduced here is to distinguish between the delay due to the most recent performed operation and the delay propagated from previous operations. The proposed ideas were made possible through examining the similarities and differences between the naive and heuristic structures supported by domain knowledge and expertise of local authorities.

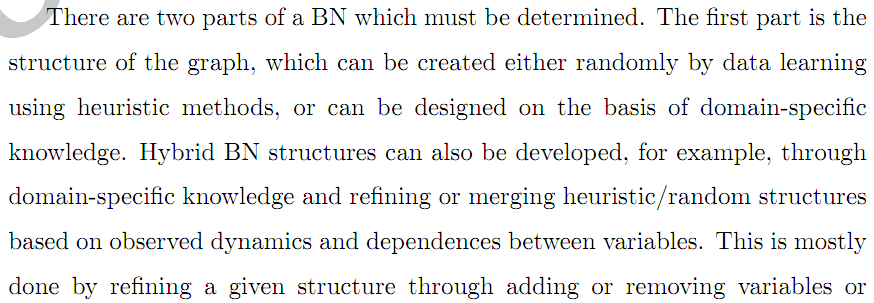
Это – первая гибридная прогностическая BN-модель. Основная идея гибридной структуры, представленной здесь, состоит в том, чтобы различать задержку, связанную с самой последней выполненной операцией, и задержку, полученную от предыдущих операций. Идея реализуется путем рассмотрения схожести и различия между примитивной (наивной) и эвристической структурами, которые поддерживаются знанием предметной области и опытом местных властей.

****

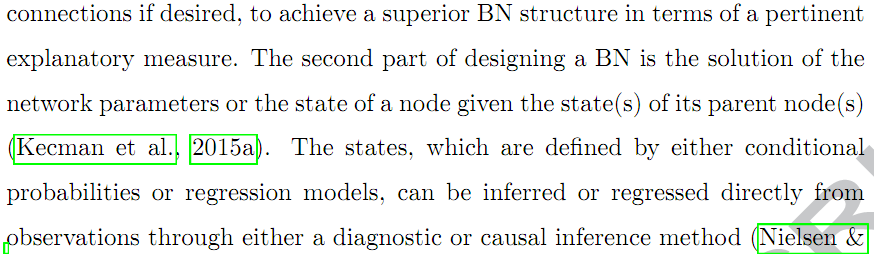
Глобальные прогнозные модели: устойчивая линейная регрессия, регрессионные деревья и вероятностные леса.

Точно так же, и статистические модели недостаточно адаптивны, чтобы объединить знание предметной области, которые имеются у (локальных) диспетчеров, и характеристики сети.

**Bayesian networks**

****

Гибридные BN структуры могут быть созданы, например, на основе знания области и обработки (улучшения) или слияния эвристических/случайных структур, основанных на динамике и зависимостях между переменными.Чаще всего – улучшение структуры посредством добавления или исключения переменных или связей (при желании) – с целью получения наилучшей структуры по показателю (pertinent explanatory measure) «соответствующая (релевантная) независимая (фиксированная) мера»

****

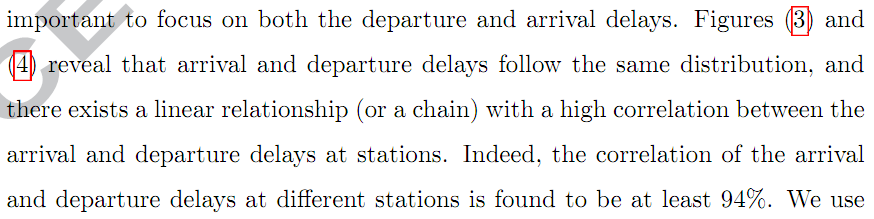
Вторая часть – определение параметров сети или состояния узла при данном состоянии родительского узла (узлов). Эти состояния, которые определяются условными вероятностями или регрессионными моделями, могут быть получены или вычислены (как коэффициенты регрессии) прямо из наблюдений, или получены путем диагностики, или причинно-следственным методом.

События отправления, прибытия, проследования станции. Каждое событие – узел BN, дуга – поездная работа. Модель показывает последовательность событий, действий, их логическую взаимосвязь.

Принципиальная идея: предсказать время выполнения каждого действия (хода, стоянки) в зависимости от состояний родительских узлов, которые отражают актуальное состояние (или задержку) непосредственных входных событий. Родительские узлы играют роль независимых переменных для каждого из их узлов-потомков.

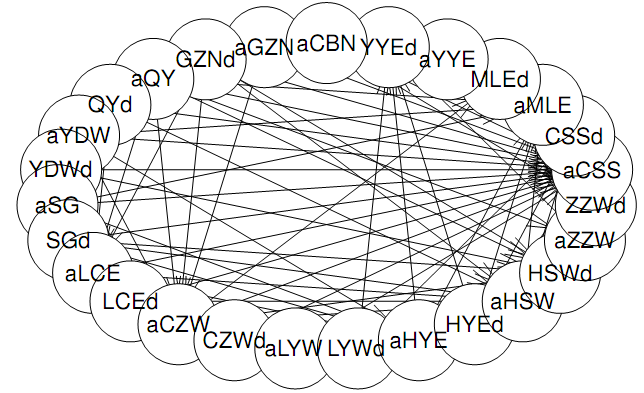
**Description of the data set**

The data used in this study come from train operation records on Wuhan-Guangzhou (WH-GZ) high-speed rail (HSR) line in China. The WH-GZ HSR connects Wuhan in Hubei province to Guangzhou in Guangdong province with a 1096 km double-track HSR line including 18 stations, as shown in Fig. 2. A total of 15 stations and 14 sections,



4.2. Обучение и подтверждение достоверности

Сетевая структура НС строится с использованием метода восхождения (**подъема в гору).** Структура изучается эвристически на основе эмпирических данных. DAG состоит из узлов, дуги добавляются одна за другой. Показатель (score) сети отражает добавочную аналитическую способность (возможность) сети, когда дуга добавляется или исключается (между парой узлов). На каждом шаге выбирается действие (добавление, исключение, разворот), дающее наибольший показатель.

На рис. 5 показана BN-структура, полученная этим методом. 

Вторая архитектура (PL) – простейшая линейная структура – в которой события соединены в порядке, соответствующем графику. Каждое событие подключено к его непосредственно предшествующему событию. То есть, события происходят в фиксированной последовательности. В вероятностном смысле, прибытийное распределение есть функция распределения при отправлении - наиболее близкой предыдущей операции.

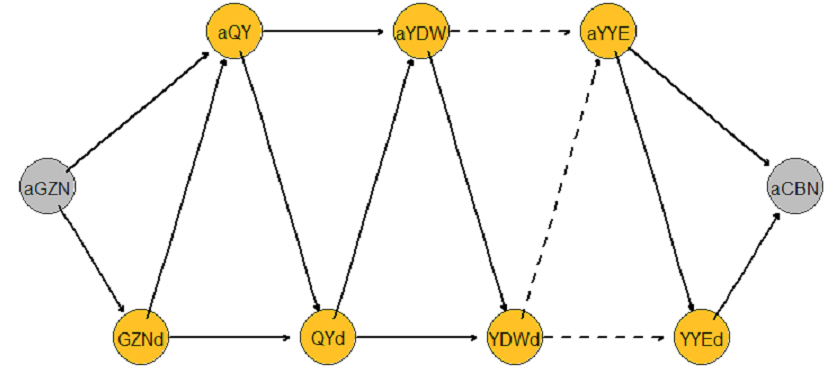
Третья предложенная модель – гибридная BN (HB) – получена из первых двух моделей на основе знания предмета, экспертных соображений о последовательности станций и связи между последовательными поездными операциями.

Из линейной модели взято то, что событие связано с непосредственно предшествующим событием. Однако видно, что в НС-структуре большинство событий связаны со вторым предыдущим событием – по отношению к анализируемому. Из полученной ранее информации следует, что задержка поезда распространяется на последующую (после ближайшей) станцию. Этот процесс может вызвать добавку к задержке, полученной на данной станции от непосредственно предшествующей станции.

На рис. 5 видно, что некоторые узлы имеют более, чем два, родителя. Авторы не смогли найти никакой логики, обобщающей это свойство.

Основная идея: определить разницу между задержками, распространяющимися от исходных операций, и более ранними возникающими задержками.

На рис. 6 – графическое представление гибридной BN-структуры для жд участка GZN-CBN. Узлы с буквой а – прибытия, с буквой d – отправления. Прибытие на ст. GZN влияет на отправление с этой станции и непосредственно на прибытие на последующую ст. QY (две дуги). За исключением первого и последнего узлов, каждый локальный узел обуславливается (формируется) двумя ближайшими предшествующими узлами. Предпоследний узел отражает вновь выполняемую операцию (результат – событие k), последний рассматривается как распространитель задержки от предшествующей операции. Результирующее распределение есть функция двух компонент: распределения времени выполнения всех предшествующих операций и прогнозного времени выполнения недавней (наиболее близкой) операции. Более точно, предшествующая часть



**Performance evaluation**

To investigate how the predictions from the proposed model match with the real-world benchmark data, we performed the following series of analyses. Firstly, a comparison of the real and predicted arrival and departure delay distributions for each station is provided in Fig. 8, Fig. 9. Moreover, in Fig. 10 the scatter plots of the observed vs. the predicted values for arrival and departure delays are presented. From these results, we can see that the predictions match well with observations,