Анализ общесетевых закономерностей задержек железнодорожных перевозок с использованием Байесовского сетевого обучения

Mehmet Baran Ulak c, Anil Yazici a, Yun Zhang b

Ключевые слова:

Transit delays - ?

Crowdsourced data -?

Delay pattern identification - ?

Bayesian network learning - ?

Network dependency metrics -?

**Предисловие:**

Задержки в железнодорожном сообщении, как правило, рассматриваются с точки зрения своевременности выполнения рейсов или проблем на отдельных остановках. При таком подходе к масштабированию остановок игнорируется тот факт, что задержки также вызываются и закрепляются факторами, охватывающими всю сеть (например, узкими местами, возникающими из-за совместного использования путей несколькими транзитными линиями). Целью данной статьи является разработка сетевой модели и показателей, которые могут количественно определять зависимость задержек между остановками транзитной сети и выявлять локальные источники проблем в масштабах всей сети. Для этой цели используется байесовское сетевое обучение (на стыке машинного обучения и network science). На основе вычисленных байесовских сетей (BNs) были разработаны сетевые показатели (индуцирующие и восприимчивые) для количественной оценки влияния задержек на остановках в масштабах всей сети. Для реализации предложенной структуры были учтены задержки на Long Island Rail Road (LIRR)были собраны с помощью краудсорсингового приложения OnTime для сбора информации о транзите в режиме реального времени. Разработанная модель BN была протестирована путем перекрестной проверки, дала многообещающие результаты по точности, успешно выявила проблемные остановки на основе отчетов LIRR и предоставила дальнейшее изучение влияния сети. Модель BN и разработанные показатели были дополнительно протестированы с использованием естественного эксперимента, то есть исследования "до" и "после", посвященного недавно завершенному проекту расширения железнодорожных путей в LIRR. Полученные результаты свидетельствуют о том, что BN learning может успешно выявлять сетевые зависимости и указывать железнодорожные соединения/коридоры, которые являются наилучшими кандидатами для последующих инвестиций в улучшение. В целом, разработанные показатели позволяют количественно оценить зависимость задержек между остановками и могут быть использованы разработчиками политики и практиками для принятия решений об инвестициях и улучшениях.

**1. Введение**

Общественный транспорт является важным компонентом здоровой транспортной системы. Для поддержания эффективной системы общественного транспорта требуется высокий уровень надежности, который определяется как уверенность в том, что предоставляемые услуги будут обслуживаться в соответствии с заданным расписанием. На надежность влияют задержки в перевозках, которые могут возникать по различным причинам: нерегулярные перемещения пассажиров или большое количество пассажиров, совершающих посадку/высадку; недостаточная пропускная способность железнодорожных путей; сцепки поездов, узкие места (например, разветвления) в железнодорожной сети; условия окружающей среды (например, погодные условия); отказы и поломки оборудования; аварии. Эти причины могут быть классифицированы как первичные или вторичные, в зависимости от характера события, вызвавшего задержку.

Первичные задержки могут быть вызваны такими событиями, как отказ оборудования или аварии. Вторичные задержки (также известные как knoch-on delays) являются следствием первичных задержек в других поездах. Кроме того, исходя из существующей терминологии, задержки в системах железнодорожного транспорта можно разделить на повторяющиеся и разовые. Несистематические события, такие как отказы оборудования, поломки или сбои в работе, могут привести к разовым. С другой стороны, периодические задержки вызваны ежедневными операциями (например, высадкой и посадкой пассажиров), систематической зависимостью между различными маршрутами (например, приоритезацией или задержкой поездов, приближающихся к одной и той же остановке/пути), а также эксплуатационными ограничениями, такими как пропускная способность путей, обусловленная инфраструктурой. Более того, задержки могут увеличиваться из-за взаимосвязанности на железных дорогах и использования общих маршрутов другими поездами и в конечном итоге могут вызвать эффект домино для всей системы. Такой эффект домино указывает на важность сетевого подхода, т.е. на зависимость между остановками на железной дороге с точки зрения увеличения или уменьшения существующих задержек.

Для общесетевого анализа требуются данные о задержках по всей железнодорожной сети, чтобы выявить закономерности и зависимости от задержек. К счастью, более широкое использование транспортных агентств для отслеживания поездов с помощью GPS и краудсорсинговых приложений, которые собирают информацию о прибытии поездов, могут предоставить необходимые данные. Разработки в области машинного обучения и сетевых технологий параллельно с растущей доступностью больших наборов данных позволяют исследователям решать задачу анализа сетевых задержек.

Соответственно, в данном исследовании изучаются данные о задержках в пути для достижения следующих целей:

1. разработка сетевой модели, которая может идентифицировать взаимосвязь между шаблонами задержек на транзитных остановках, т.е. какие остановки вызывают, распространяют и/или сводят на нет задержки по всей железнодорожной сети.
2. количественная оценка зависимостей между остановками трафика с помощью сетевых показателей, чтобы помочь политикам и практикам определить приоритеты инвестиций и/или проектов в железнодорожной сети.

Таким образом, в данном исследовании задержки рассматриваются с точки зрения всей сети, поскольку были проанализированы остановки во всей транзитной сети, а не факторы на уровне станций (например, движение пассажиров, количество пассажиров и т.д.), влияющие на задержки. Для этой цели было использовано байесовское сетевое обучение. На основе результатов разработанных байесовских сетей (BNs) были сформулированы сетевые показатели для количественной оценки влияния задержек, возникающих на отдельных остановках, на всю сеть. Для реализации предложенной концепции задержки (в минутах), возникшие на остановках на Long Island Rail Road (LIRR) были собраны с помощью краудсорсингового приложения OnTime, предоставляющего информацию о транзите в режиме реального времени.

Это исследование вносит вклад в общее исследование задержек в железнодорожной сети следующим образом:

1. Впервые был использован байесовский подход к изучению сети для выявления закономерностей задержек в сложной транзитной сети;
2. Были разработаны показатели, основанные на байесовской сети, для получения ощутимого показателя, который может быть использован для принятие решений и оценки эффективности работы.

В следующих разделах, во-первых, представлен обзор литературы о задержках в транзитных сетях и байесовском обучении сети, за которым следует описание области исследования и данных. Во-вторых, обсуждается байесовский сетевой подход к обучению, объясняется обработка данных и формулируются сетевые показатели. Затем разработанная модель BN тестируется на точность, подтвержденную с помощью естественного эксперимента в рамках двухпутного проекта before & after в филиале LIRR в Ронконкоме. Наконец, результаты обобщаются в разделе "Заключение" вместе с предостережениями и направлениями будущих исследований.

**2. Обзор литературы**

Ранее в нескольких исследованиях изучались прогнозирование задержек, прогнозирование распространения задержек, идентификация шаблона задержек в железнодорожных сетях и зависимость компонентов транзита. Широко известно, что выявление зависимости задержек между различными элементами (например, поездами, станциями) железнодорожной сети имеет решающее значение для поддержания надежности и эффективности железнодорожного транспорта. Однако обнаружение таких зависимостей от задержки является сложной задачей и требует обширных знаний об инфраструктуре. Тем не менее, Конте и Шобель (Conte and Schobel, 2007) определили зависимости задержки без такой исчерпывающей информации о системе, используя так называемый трехграфический (Tri-Graph) подход из семейства графических методов (Wille and Bühlmann, 2004). Этот подход, по сути, определяет полную условную модель для событий (например, отправления или прибытия поездов) посредством моделирования подгрупп событий за один раз, чтобы найти условную зависимость между этими событиями. Однако такая группировка требует тщательного изучения расписания движения поездов, и, следовательно, ее трудно реализовать в сложных транзитных сетях. Поэтому авторы применили модель к железнодорожной сети дальнего следования, состоящей из 8 станций с 928 событиями, и неясно, будет ли этот подход может быть полезен для сложной сети и больших объемов данных. Тем не менее, авторы показали, что модели на основе графов хорошо подходят для определения зависимостей от задержек в железнодорожной системе. Флайер и др. (2009) разработали набор алгоритмов для определения зависимости между поездами, курсирующими по швейцарской железнодорожной сети, от расписания движения. Их исследование показало, что для точного определения задержек в крупной железнодорожной сети необходимо выявить основную взаимосвязь между поездными рейсами. Однако, их подход был разработан для обнаружения зависимостей на одной остановке, а не для определения зависимостей задержки по всей сети.

Распространение задержки в крупных сетях также является проблемой для поддержания эффективной работы (Бюкер и Сейболд, 2012; Серенсен и др., 2017; Валландер и Макитало, 2012). Сосредоточившись на этом вопросе, Бюкер и Сейболд (2012) разработали стохастический подход, который вычисляет потенциальное распространение задержек в крупных сетях. Утверждается, что этот подход может помочь оценить надежность железнодорожных расписаний. Руководствуясь той же мотивацией, Уолландер и Макитало (2012) применили “анализ последовательности”, основанный на ассоциации анализ (т.е. одновременное возникновение событий) для выявления задержек поездов, которые происходят последовательно, что указывает на распространяющуюся закономерность. Авторы стремились выявить взаимосвязь между задержками поездов дальнего следования на финских железных дорогах и выявить причины этих задержек. Однако исследование было сосредоточено на задержках поездов, прибывающих на одну станцию и отправляющихся с нее, и не рассматривало всю железнодорожную сеть в целом. Аналогичный анализ связи был также проведен Ябуки и соавторами (2015), чтобы оценить эффективность контрмер, которые принимающихся для уменьшения задержек и распространения задержек.

Несколько подходов, таких как методы опорных векторов (Markovic et al., 2015), нейронные сети (NN) (Milinkovic et al., 2013; Yaghini et al., 2013), глубокое и поверхностное обучение (Oneto et al., 2017), подходы, основанные на данных (Wang and Work, 2015)., 2019), марковские модели (Гаурав и Шривастава, 2018), эконометрические модели (Агбели и Либнао, 2018; Горман, 2009), гибридные модели (Онето и др., 2020) и байесовские сети (Корман и Кечман, 2018; Хуан и др., 2018). др., 2020; Лессан и Фу, 2019) были упомянуты в литературе. Эти подходы помогают прогнозировать задержки на железнодорожном транспорте и сравнивать альтернативные варианты в контексте задержек. Сообщалось, что нейросетевой подход превосходит дерево решений и мультиномиальную логистическую регрессию при прогнозировании задержек (Yaghini et al., 2013). Было обнаружено, что по сравнению с нейронными сетями машинное моделирование с использованием опорных векторов лучше отражает взаимосвязь между задержками и особенностями железнодорожной системы (Markovic et al., 2015).

Байесовское сетевое обучение было впервые рассмотрено в контексте анализа задержки Conte (2007), однако лишь недавно оно было использовано для прогнозирования задержки Lessan and Fu (2019), Corman and Kecman (2018) и Huang et al. (2020). Все недавние исследования использовали BN learning для построения взаимосвязей между узлами и разработки моделей прогнозирования задержек. Corman and Kecman (2018) предложили стохастическую модель прогнозирования задержек, в которой прогнозы задержки изменяются с использованием потока данных в реальном времени. Lessan and Fu (2019), наоборот протестировали эффективность трех различных подходов к обучению BN, а именно: восхождение на холм, примитивный линейный подход и гибридный подход BN для прогнозирования задержки. Исследование показало, что гибридный подход BN работает лучше, чем два других, особенно с точки зрения вариативности результатов; однако их результаты показывают, что между этими подходами нет большой разницы с точки зрения средней ошибки прогнозирования. В недавнем исследовании Huang et al. (2020) была разработана гибридная модель BN для прогнозирования задержек, количества поездов, на которые распространяется действие, и общей задержки путем построения взаимосвязей между этими тремя переменными. Тем не менее, ни в одном из этих исследований не изучались закономерности задержек и зависимости задержек между компонентами (т.е. остановками) сложных транзитных сетей с использованием байесовского сетевого обучения.

Выявление закономерностей и источников задержек (например, узких мест в сети) очень важно для повышения надежности транзитных систем (Cule et al., 2011). Однако выявить истинную причину задержек, как правило, сложно из-за комплексности сети. Тем не менее, существуют исследования, направленные на решение этой проблемы путем выявления последовательности движения поездов, вызывающей задержки (Cule et al., 2011); разработки алгоритмов для анализа распространения задержек на отдельных путях (Sørensen et al., 2017); или оптимизации расписания движения поездов (Andersson et al., 2015). Чтобы понять проблемы, связанные с систематическими задержками на бельгийских железных дорогах, и выявить потенциальные причины, лежащие в основе задержек, Куле и др. использовали методы анализа частых ситуаций. Исследователи не использовали продолжительность задержек, а рассматривали поезда как задерживающиеся или нет, чтобы выявить последовательность поездов, которые вызывают задержки. Хотя этот подход полезен для систематического определения последовательности задержек поездов, масштабы зависимостей в масштабах всей сети полностью не раскрыты. В большинстве исследований, приведенных в литературе, основное внимание уделяется задержкам поездов дальнего следования, которые курсируют по относительно менее сложным железнодорожным сетям. Однако в очень немногих исследованиях предпринята попытка изучить закономерности задержек в пригородных сетях, которые, как правило, страдают от большого объема перевозок, низкой пропускной способности путей и сложной инфраструктуры. Данное исследование направлено на устранение этого пробела путем разработки байесовской сетевой модели и показателей, которые могут объяснить частые задержки в транзитных системах, количественно оценить зависимость задержек между остановками в транзитной сети и выявить локальные источники проблем в масштабах всей сети.



Рисунок 1- Полная сеть Long Island Rail Road (MTA, 2019b) и самый высокий пассажиропоток.

**3. Тематическая область исследования и данные**

3.1 Область исследования

Объектом исследования является система Long Island Rail Road (LIRR), которая соединяет Нью-Йорк с Лонг-Айлендом, штат Нью-Йорк, несколькими линиями (рис. 1). Система LIRR имеет в общей сложности 11 ответвлений (то есть железнодорожных линий), охватывающих 620 миль путей вдоль Лонг-Айленда со 124 остановками. В Нью-Йорке есть две конечные станции, а именно Пенсильванский вокзал и Атлантический вокзал. В 2017 году LIRR перевезла 89,2 миллиона пассажиров на своих 735 ежедневных поездах. Эти статистические данные делают LIRR самой загруженной пригородной железной дорогой в Северной Америке с примерной выручкой 710 миллионов долларов (Metropolitan Transportation Authority, 2017). В 2017 году LIRR сообщила о 91,4% своевременных рейсах (т.е. поездках с задержкой на остановках менее 5 минут и 59 секунд).

В таблице 1 показано ежедневное количество поездов на каждой остановке LIRR за период данных в 32 дня с 23.12.2017 по 23.01.2018. Цифры в таблице 1 рассчитаны на основе запланированных поездок в течение исследуемого периода (MTA, 2019a). На Ямайском вокзале проходит самое большое количество поездов - в среднем 543 поезда в день, за которым следует Пенсильванский вокзал со средним количеством поездов 470 в день.



3.2 Краудсорсинговые данные о задержке – оперативные журналы пользователей

OnTime - это приложение для получения информации о транзите в режиме реального времени, которое имеет функцию краудсорсинга, с помощью которой пользователи могут зарегистрироваться для отслеживания статуса поезда; таким образом, пользователи позволяют приложению использовать их GPS-координаты во время поездки. Координаты пользователей собираются анонимно и используются для расчета расстояния до следующей остановки и скорости поезда, и, соответственно, расчетного времени прибытия (ETA) для каждой станции. Задержки поездов на каждой остановке рассчитываются путем сравнения этих данных с запланированным временем прибытия, т.е. опубликованным расписанием столичного транспортного управления (MTA, 2019a; Transitfeeds, 2019). Об этих задержках сообщается всем пассажирам, которые пользуются приложением. Обратите внимание, что пользователи приложения не принимают активного участия (путем ввода информации) в расчете предполагаемого времени прибытия и задержек. Возможно (хотя и очень редко), что поезд может прибыть на остановку раньше запланированного времени прибытия. В таких случаях задержка для анализа не регистрируется. В таких случаях раннего прибытия поезд не отправляется до запланированного времени. Следовательно, с точки зрения распространения задержки и моделирования, раннее прибытие и отсутствие задержки практически эквивалентны, и раннее прибытие не влияет на результаты моделирования.

Данные краудсорсингового журнала включают 372 502 измерений с отметками времени пользователей OnTime в системе LIRR за период с 23.12.2017 по 23.01.2018 (32 дня). Среди этих журналов есть записи с очень большими задержками, которые были признаны либо ошибочными, либо возникшими в результате чрезвычайной ситуации. Чтобы избежать предвзятых и неточных оценок из-за таких записей, задержки, превышающие два часы были выброшены. Обратите внимание, что эти записи соответствуют только 0,2% данных (756 из 372 502 журналов); следовательно, исключение этих записей с выбросами не приведет к искажению данных (особенно в рамках вероятностных моделей, таких как байесовские сети). Окончательный набор данных содержит 371 746 оценок задержки в сети LIRR. Инжир. На рис. 2 показана выборка этих задержек на всех трассах LIRR, иллюстрирующая, что величина задержки увеличивается по направлению к Нью-Йорку (запад).



Рисунок 2 - Пример задержек, возникших в системе LIRR.

**4. Методология (теория?)**

4.1 Байесовское сетевое обучение и логический вывод

Взаимосвязь между остановками с точки зрения возникающих задержек была исследована с использованием байесовского сетевого подхода к обучению (BN). Байесовская сеть - это статистическая модель, в которой условные зависимости случайных величин представлены направленным ациклическим графом (то есть графом одностороннего взаимодействия) (DAG) (Buntine, 1996; Charniak, 1991). Таким образом, DAG - это тип графа, ребра которого (т.е. пути) являются направленными, а не циклическими (т.е. имеют одностороннюю связь). Байесовские сети полезны для представления статистически значимых взаимосвязи между событиями, которые, возможно, имеют причинно-следственную зависимость друг от друга как родительский и дочерний узел. Таким образом, байесовские сети полезны для прогнозирования вероятности возникновения набора событий, учитывая, что наблюдались другие наборы событий. Популярным примером в этом контексте является то, что симптомы, проявляющиеся у пациента, могут быть использованы для оценки вероятности заболеваний, которые могут быть у пациента (Buntine, 1996). Подход BN недавно был "адаптирован" для прогнозирования железнодорожных перевозок и задержек (Conte, 2007; Corman and Kecman, 2018; Huang et al., 2020; Lessan and Fu, 2019; Zilko et al., 2016). Эти исследования показали, что BN learning очень эффективен при анализе и прогнозировании задержек на узлах железнодорожной сети. Учитывая многообещающие результаты, BN был принят и в текущем исследовании. Однако обратите внимание, что точность прогнозирования модели также зависит от данных, даже несмотря на то, что возможности прогнозирования с помощью обучения BN были показаны в литературе. В случае этого исследования пространственно-временной охват модели краудсорсинговые данные являются важным фактором точности модели, т.е. чем больше количество записей о задержках пользователей для каждой остановки в разное время суток, тем выше точность прогноза. В целом, onTime записи обеспечивают достаточный пространственно-временной охват для хорошей точности прогнозирования, что было продемонстрировано в последующих разделах с помощью перекрестной проверки.

Байесовская сетевая структура может быть представлена DAG, которая имеет n узлов (1, 2, ∞, n), где каждый узел связан с случайная величина Xj∊(X1,X2,⋯,Xn). С учетом этого DAG условные зависимости n случайных величин X могут быть представлены в виде байесовской сети (Buntine, 1996; Charniak, 1991).:



ПЕРЕПЕЧАТАТЬ

где ΠXj - родительские узлы узла j. На рис.3 показана простая байесовская сеть, имеющая семь узлов и семь ребер. На этом графе узлы A и B являются родительскими узлами узла D, а узлы F и G - дочерними узлами узла D. Эта сеть указывает, что узел D зависит от узлов A и B. Аналогично, узел F зависит от узла D, тогда как он условно независим от A и B. Пожалуйста, обратитесь к Buntine (1996) and Scutari (2010) для получения более подробной информации о теории и применении байесовских сетей.

  
Рисунок 3 -Пример Байесовской сети.

В этом исследовании байесовское сетевое обучение было реализовано с использованием R-project и пакета “bnlearn”. Оценка максимального правдоподобия (MLE- Maximum likelihood estimation) использовалась для оценки взаимосвязи между узлами (т.е. родительско–дочерним) и алгоритмом обучения гибридной структуры. (Lessan and Fu, 2019), а именно метод максимального и минимального восхождения на холм (mmhc- max-min hill–climbing), был реализован для эвристического изучения графовых структур. Конечные графы представляют собой байесовские сети, состоящие из статистически значимых ребер, которые идентифицируются на основе значений p тестов условной независимости с пороговым значением 0,05 (т.е. 95% доверительный интервал) (Scutari, 2010; Scutari and Ness, 2019).

4.2 Реализация байесовского сетевого подхода

4.2.1 Обработка данных

Собранные с помощью краудсорсинга данные состоят из журналов с предполагаемыми задержками. Эти журналы были дополнительно обработаны для определения наиболее точных оценок задержек. То есть, как правило, более чем один пассажир пригородного сообщения использует приложение OnTime для одновременной проверки задержек; таким образом, существует несколько записей, которые предоставляют оценки задержек для одного и того же набора “поездка-пункт назначения-время”. На рисунке 4 показаны семь записей для гипотетического примера *поезда 1*. Наиболее точная задержка для этого *поезда 1,* прибывающего на гипотетическую *остановку B*, указана в последней записи (запись №7), поскольку она ближе всего к *остановке B*. В наборе данных, насчитывающем в общей сложности 371 746 записей, в анализ была включена 45 621 такая запись (например, запись №7 на рисунке 4). Другими словами, 326 125 пользовательских записей из 371 746 дают менее точные оценки задержки, чем последние записи, например, записи с 1 по 6 по сравнению с записью №7 на рис. 4. Эти более ранние записи не были учтены, и для оценки задержки поезда были использованы последние (следовательно, наиболее точные) записи.

Временной интервал 10:00 - 11:00



Задержки на Остановке В между 10:00 и 11:00 = (Поезд 1- Запись 7) + (Поезд 2 - Запись 6) + (Поезд 3 - Запись 4)

Рисунок 4 - Пример распределения записей, записанных приложением за один и тот же интервал времени остановки.

Чтобы разработать байесовскую сеть остановок, задержки остановок были использованы для разработки матрицы данных о задержках в сети (таблица 2). Для этой цели каждая запись была классифицирована в соответствии со временем суток (например, 17:00-18:00), направлением поездки (Восток-Запад) и пунктом прибытия. Затем задержки суммировались на основе этих временных интервалов остановки и определялась общая задержка при остановке за данный временной интервал. Этот процесс показан на рис. 4, где три поезда прибывают на воображаемую остановку В между 10:00 и 11:00. Общая задержка на остановке B для этого временного интервала была рассчитана путем суммирования трех оценок задержки, полученных из записи №7, записи №6 и записи №4 для воображаемого *Поезда 1, поезда 2 и поезда 3* соответственно. В результате была получена матрица задержки с 768 временными строками путем деления общего периода данных (32 дня) на часовые интервалы (32 дня × 24 часа в сутки = всего 768 интервалов). Наконец, матрица задержки пар "остановка-временной интервал" (краткий обзор приведен в таблице 2) была введена в учебный процесс для определения взаимодействий с задержкой. Строки матрицы задержек рассматривались как отдельные наблюдения, которые указывают на состояние задержки в сети в данный промежуток времени.



Таблица 2 - Снимок матрицы данных об общей задержке остановок для поездок в западном направлении в утренний час пик.

Анализ проводился для двух периодов пиковой нагрузки на основе часов пик LIRR: Утренний час пик (поезда прибывают на терминалы Нью-Йорка с 18:00 до 10:00), Вечерний час пик (поезда отправляются с терминалов Нью-Йорка с 16:00 до 20:00). Для анализа утренних и вечерних часов пик были выделены строки с временными интервалами, которые соответствуют утренним и вечерним часам пик в будний день. Наконец, суммарные задержки были разделены на количество поездов (в течение рассматриваемого периода, например, вечерний час пик), проходящих через каждую остановку, в порядке чтобы рассчитать задержку для каждого поезда. Обратите внимание, что из-за зависимости от пользователей краудсорсинговые данные не включают записи о задержках для всех периодов и остановок. Для определения недостающих значений временной задержки (например, 2117 временных интервалов из 4410 для вечернего часа пика в восточном направлении) мы внедрили инструмент “structural.em” в пакет “bnlearn” для оценки недостающих значений (Scutari, 2010; Scutari and Ness, 2019). Чтобы предсказать недостающие значения, необходимо заранее знать структуру байесовской сети. Однако структура сети также должна быть оценивается с использованием недостающих данных, что создает проблему с курицей или яйцом.

Подход к изучению байесовской сетевой структуры с использованием неполных данных был впервые разработан Friedman (1997) с использованием расширения алгоритма максимизации математического ожидания для поиска большого числа возможных сетевых структур и определения наиболее подходящей из них. Friedman (1997) также привел пример, в котором предложенный подход позволяет успешно обучать сетевые структуры при наличии нескольких переменных, в которых в общей сложности не хватает 30% значений. Пожалуйста, обратитесь к исследованию Friedman (1997) для получения теоретической основы и результатов. Инструмент “structural.em” в пакете “bnlearn” (Scutari and Ness, 2019) основан на подходе , предложенном Friedman (1997). Оценка пропущенных значений и байесовских сетей повторялась в течение 100 раз (т.е. процесс начальной загрузки) для каждой группы анализа. В итоговые графы были включены только часто выявляемые (более 80% при повторном анализе) зависимости между узлами.

4.2.2. Топология сети

Структура байесовской сети (BN) влияет на достоверность и интерпретацию результатов; таким образом, она является одним из важнейших компонентов обучения байесовской сети. Существует два основных подхода к определению структуры BN:

1) построение структуры до подачи данных и определения взаимодействия между узлами;

2) обучение сети непосредственно на основе данных без каких-либо предварительных ограничений.

По сути, первый подход возможен при наличии экспертных знаний о структуре сети; принимая во внимание, что второй подход является более распространенным (и менее предвзятым), поскольку, как правило, трудно получить предварительные знания о сетевых взаимодействиях.

В этом исследовании был применен второй подход для определения структуры сети без какого-либо предварительного проектирования. Однако из-за физических соединений в железнодорожной сети зависимость задержки между остановками определяется физической сетью и запланированными поездками. Например, поезда от Пенсильванского вокзала (рис. 5 – группа 22) до Атлантического терминала (рис. 5 – Группа 21) не ходят, и они соединяются через станцию Ямайка (рис. 5 – группа 18). Таким образом, зависимости задержек между Пенсильванским вокзалом и Атлантическим терминалом могут возникать только через станцию Ямайка. Чтобы отразить эти зависимости, были введены ограничения “черного списка”, т.е. принуждающие BN учиться игнорировать определенные грани (т.е. взаимодействия) (Scutari, 2010). Использование черного списка помогает избежать взаимодействия с узлами (остановками транзитной системы), которое физически невозможно. Соответственно, топология физической железнодорожной сети включена в анализ BN с точки зрения смежности на основе маршрутов. Можно утверждать, что принятый подход можно было бы охарактеризовать как “гибридный подход”, который преимущественно обучает сеть на основе данных.

Смежность маршрутов была разработана с использованием схемы движения поездов LIRR. Для этой цели остановки были сгруппированы на основе “сегментов”, разделенных разветвлениями железнодорожной сети. Обратите внимание, что всего существует 13 разветвлений, приводящих к появлению 12 стоп-групп (одна группа между двумя разветвлениями) и 11 конечных ответвлений, приводящих примерно к 10 стоп-группам; в сумме получается 22 стоп-группы (рис. 5). Две стоп-группы считались смежными, если поезд мог добраться с одной из этих остановок на другую. Для примера, есть запланированные транзитные поездки от остановок группы 4 до остановок группы 22 (MTA, 2019 a), в то время как между остановками группы 4 и группы 20 таких поездок нет. Матрица смежности, основанная на номерах групп, показана на рисунке. 6. Структура смежности гарантирует, что задержки двух несмежных остановок могут быть связаны только с остановкой, которая находится рядом с обеими этими остановками.



Рисунок 5 - Назначенные групп остановок.



Рисунок 6 - Матрица смежности остановок, созданная на основе присвоенных номеров групп. Зеленый прямоугольник указывает на наличие смежности между двумя группами.

4.3 Показатели производительности сети

Байесовские сети (BNs) извлекают зависимости между остановками с точки зрения возникающих задержек; однако байесовские сети не всегда просты для понимания из-за большого количества узлов и связей в железнодорожной сети. В контексте взаимодействия с задержкой и распространения задержки можно разделить роли, которые играют остановки, на две категории в зависимости от их воздействия на сеть:

1) Остановка может быть источником задержки, распространяющейся по сети;

2) на остановку может влиять задержки, распространяющиеся внутри сети.

Это также можно аллегорически описать как “отношения жертвы и хищника”. Такой “источник задержки” по сравнению с схемой “под влиянием задержки” также может рассматриваться как концепция первичной задержки в сравнении со вторичной задержкой в литературе по эксплуатации железных дорог (Huang et al., 2020; Lee et al., 2016). Мы предположили, что эти два аспекта воздействия на всю сеть различны, т.е. первый из них является причиной задержек, тогда как второй является наиболее чувствительным. Таким образом, были разработаны два показателя, а именно “инициатор” и “Восприимчивость” для количественной оценки зависимости задержек в железнодорожной сети. Другими словами, показатели количественно оценивают влияние остановок на задержки в сети в целом, а также их уязвимость перед задержками, вызванными другими остановками в железнодорожной сети. С этой точки зрения, показатель "инициатор" определяет и количественно оценивает влияние остановок на общие задержки, возникающие в сети. Показатель "Восприимчивость" определяет и количественно оценивает уязвимость остановок перед воздействием задержек, вызванных другими остановками.

Для оценки задержки "инициатор", выбранной остановке присваивается ее средняя задержка (наблюдаемая в данных), а всем остальным остановкам присваивается нулевая задержка. Затем, используя байесовскую сеть, задержки на всех остальных остановках оцениваются одна за другой с учетом заданных начальных условий (т.е. средняя задержка на остановке индуктора и нулевые задержки на всех остальных остановках). Затем расчетные задержки умножаются на ежедневное количество поездок на соответствующих остановках (на основе опубликованного расписания поездок LIRR), и показатель "инициатор", вызывающий задержки, рассчитывается путем усреднения всех значений (Уравнение (2)).,

Для оценки задержки "восприимчивость" всем остановкам, кроме выбранной, присваиваются соответствующие средние задержки (наблюдаемые в данных), одна за другой. Затем задержка на "восприимчивой" остановке оценивается с использованием разработанной байесовской сети с учетом этих начальных условий. Наконец, рассчитанная средняя задержка умножается на ежедневное количество поездок на "восприимчивой" остановке (уравнение (3)). Итоговые баллы представляют собой минуты задержки на каждой остановке, т.е. среднюю задержку, вызванную (причиной) или находящуюся под влиянием (уязвимой) каждой остановки в Байесовской сети.



2 - "инициатор"

3 - "восприимчивость", где Xi - задержка остановки i, Vi - ежедневное количество поездок на/с остановки i, n - общее количество остановок, x - среднее значение задержки переменной x. E(X|Y = y) - условное математическое ожидание X, если Y равно y, и E(X|Y = y) =



- функция общей плотности вероятности, найденная с использованием разработанной байесовской сети.

Для оценки задержек, используемых в метриках, была использована функция прогнозирования, предоставляемая в пакете “bnlearn”, а также метод bayes-lw, который использует всю байесовскую сеть (не только родительские узлы) для вычисления ожидаемых значений путем усреднения результатов моделирования с взвешиванием вероятности (Scutari, 2010; Scutari and Ness, 2019). Поскольку это стохастический процесс, анализ был повторен несколько раз (100 повторений), и были найдены средние условные ожидания задержек. В качестве примера в таблице 3 показано, как задержка на Станции "Ямайка" влияет на задержки на других остановках и как задержки на других остановках влияют на станцию "Ямайка". Например, строки “Задержка от ”Ямайка“ показывают задержки, вызванные станцией ”Ямайка" на других остановках, когда среднее значение задержки на станции "Ямайка" (34,89 мин) вводится в BN. Строки "Задержка на Ямайке" иллюстрируют подверженность Ямайки задержкам из-за задержек на других станциях LIRR.

Обратите внимание, что сформулированные показатели напрямую не связаны с моделированием BN. Сформулированные показатели используют взаимодействия с задержками между станциями и оценки задержек. Эти необходимые исходные данные могут быть получены с помощью моделей/подходов, отличных от BNs. Тем не менее, моделирование BN является наиболее подходящим подходом для получения взаимодействий с задержками в железнодорожной сети; следовательно, ожидается, что сформулированные показатели будут хорошим дополнением к принятому подходу, т.е. к обучению BN.