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