Матрицы корреспонденций и анализ пассажирских потоков

Д.Е. Намиот, М.Н. Некраплённая, О.Н. Покусаев, А.Е. Чекмарев

***Аннотация*— В настоящей статье речь идет о подходах к оценке использования станций метрополитена на основе матриц корреспонденции, описывающих перемещения пассажиров. Телекоммуникационные операторы в настоящее время поддерживают мобильную связь в метро. Это приводит к тому, что операторы могут отслеживать входы и выходы пассажиров из метро, определяя моменты, когда их мобильный абонент переключается на базовую станцию, размещенную в метро (входит в метро) или на базовую станцию в городе (выходит из метро). Такие анонимные данные могут быть сгруппированы по времени, чтобы исключить отслеживание индивидуальных абонентов и представлены для анализа. Итоговые данные представляют собой так называемую матрицу корреспонденций: за некоторый временной интервал для каждой пары станций известно количество перемещающихся между этими станциями пассажиров. Обычно, при анализе транспортных систем, восстановление такой матрицы (то есть, фактически, прогноз пассажиропотоков) и является основной задачей. В данном же случае, прогноз становится не нужен – все пассажиропотоки известны. Обсуждению того, что может являться целью анализа при такой структуре исходных данных и посвящена настоящая статья.**

***Ключевые слова*—транспорт, метро, матрицы корреспонденции.**

I. Введение

Развитие технологий изменило, во многих случаях, подходы к анализу транспортных данных в городах. Традиционно, исследование транспорта предполагало (так сложилось исторически) какой-то прогноз. Анализ транспортных потоков это, в конечном счете, какой-то прогноз – какое будет движение, сколько человек поедут и т.д. Но давайте посмотрим на технологический фон этих процессов. На пригородных железнодорожных станциях пассажиры сами отмечают свои проездные документы в начале и конце поездки. Естественно, что эта отметка регистрируется в некоторой базе данных. Соответственно, мы можем анализировать точно измеренные данные по всем поездкам между двумя любыми станциями. Что здесь можно прогнозировать? Если мы рассмотрим метро в Москве, то там пассажиры “отмечаются” только при входе. Конечный пункт маршрута необходимо оценивать по каким-то эвристикам. Например, после “первого” использования карты “Тройка” (проездного билета) можно попробовать определить, на какой станции эта же карта была использована во “второй” раз. С большой вероятностью эта станция и была конечной в поездке. Но, на самом деле, здесь приходят на помощь телекоммуникационные операторы. Поддержка мобильной связи в метро означает наличие там базовых станций. А это, в свою очередь, означает, что оператор может фиксировать моменты, когда его абонент (мобильное устройство) переключился на подземную станцию (то есть – вошел в метро) и когда наоборот, переключился с подземной станции на наземную (то есть – вышел из метро). Исходя из этого, оператор знает начальную и конечную точку любой поездки. Далее эти данные могут быть агрегированы по времени, так, чтобы исключалась возможность отслеживания единственной поездки [1] и результатом будет так называемая матрица корреспонденций [2]. В англоязычной литературе – OD matrix (origin – destination) [3]. Такая матрица привязана к определенному временному интервалу (периоду, за который агрегировались данные) и показывает количество пассажиров, которые перемещались от одной станции метро до другой. Опять-таки – как можно прогнозировать потоки в такой системе и, главное, зачем, если они, фактически полностью измеряются? Одно дело, когда матрица корреспонденции является конечной частью задачи анализа транспортных потоков (то есть, когда необходимо понять, как именно люди перемещаются в городе [4]), и другое дело – когда такая матрица является исходными данными.

Потоки (поездки) могут, естественно, как-то изменяться при перестроении сети, появлении новых маршрутов автобусных маршрутов из пригородов, которые подвозят новых пассажиров к конечным станциям метро, открытии пересадочных узлов с городской железной дорогой и т.д. Но информация обо всех этих событиях отсутствует в матрице корреспонденции. Матрица корреспонденции будет изменяться под воздействием этих событий. Эта матрица будет являться индикатором этих событий. В этом и есть основная идея ее анализа – оценка эффекта от событий, произошедших в городе. Или – уведомления об изменениях в транспортном поведении, для которого необходимо

Статья получена 10 февраля 2020.
Д.Е. Намиот - МГУ имени М.В. Ломоносова (e-mail: dnamiot@gmail.com)
М.Н. Некраплённая – МГУ имени М.В. Ломоносова (email: maria240398@mail.ru)
О.Н. Покусаев – Центр цифровых высокоскоростных транспортных систем РУТ (МИИТ) (email: o.pokusaev@rut.digital)
А.Е. Чекмарев - Центр цифровых высокоскоростных транспортных систем РУТ (МИИТ) (email: a.chekmarev@rut.digital)

определить причину. В целом, оба эти направления можно описать как анализ поведения (транспортного поведения) пассажиров в городе. Транспортное поведение – это одна из основных характеристик Умного города [5].

Данные (факты) полученные посредством анализа матрицы корреспонденции сами служат исходной информацией для прогнозирования потоков. Например, режим функционирования новой пересадочной линий для городской железной дороги определяется, очевидно, характеристиками пассажиропотока на соответствующей станции метро и т.д.

Заметим, что по похожей схеме телеком может строить матрицы корреспонденции, привязанные просто к географическим квадратам. То есть, можно разбить город (с пригородом) географической сеткой (достижимый размер, зависящий от плотности базовых станций, составляет порядка 500 х 500 м) и получить количество перемещений между произвольными квадратами [6]. Эти перемещения, естественно, как-то должны быть привязаны к транспорту, в частности, присутствовать матрице перемещений для метро.

Матрицы корреспонденции для метро (железной дороги) проще для обработки, поскольку здесь, по очевидным причинам, маршруты фиксированы, происходящее при самом перемещении мы не рассматриваем, и в используемых моделях есть только один тип выделенных объектов – станции.

В данной статье мы хотели бы остановиться на том, какие именно задачи могут решаться при анализе матриц корреспонденций метро. Работа является продолжением серии исследований [2, 7, 8].

II. О моделях использования станций

Как может использоваться информация о потоках пассажиров по станциям?

Во-первых, самое простое – это расчет возможных перераспределений пассажирских потоков. Такие данные могут использоваться, например, при вводе новых станций (маршрутов) в дополнение к существующим или наоборот, закрытии каких-либо станций (маршрутов). Исходя из того, что пассажиры будут выбирать наиболее короткие по времени маршруты можно рассчитать, кто (сколько) из существующих пассажиров изменит свои маршруты. Но это вряд ли можно назвать прогнозом, поскольку расчеты здесь абсолютно детерминированные.

Эти расчеты, конечно, отличаются от полностью статически расчетов типа анализа центральности станций [9], но, в целом, не представляют большой проблемы. Такая информация будет полезна, например, при проектировании новых пересадочных узлов, объединении метро и городской железной дороги. Единственно, что можно отметить, что в некоторых случаях может оказаться недостаточным просто указать новое количество пассажиров, которое появится на станции. Ведь это количество также будет как-то распределено по времени. И эта информация о распределении также может быть существенной. Например, новые пассажиры, добавленные к существующему утреннему пику, могут оказаться критичным изменением.

Следующее, что необходимо отметить – это замечание о том, что в описанном выше представлении матрица корреспонденций представляет собой замкнутую систему. Здесь нет информации о том, например, что открылся новый маршрут из пригородов, который увеличит количество пользователей системы (пассажиров). В системе нет информации о каких-то культурных событиях, которые на какое-то время увеличивают общее число пассажиров и так далее.

Метро реагирует на какие-то новые потоки пассажиров (они отражаются в матрице), но информация об этих новых потоках в системе отсутствует. Конкретный пример. Переключение пригородных автобусов с автовокзала около метро Тушинская на новый автовокзал около метро Ховрино вызовет (вызвало) некоторый пик в загрузке утром на зеленой линии метро, но это можно было только зарегистрировать постфактум. Предсказать это никаким образом было невозможно, поскольку информация об автобусных маршрутах является внешней по отношению к матрице корреспонденции.

Отсюда следует другое важное утверждение. Матрица корреспонденций, полученная описанным выше способом – это некоторый измерительный инструмент, с помощью которого можно определять наступление каких-то событий в городе или проверять результаты каких-либо действий. Изменения в поведении пассажиров (изменение маршрутов, времени поездок) – это то, теоретически можно определить и каждое такое изменение должно, естественно, быть как-то объяснено с точки зрения изменений в городе. Эффект от проведения каких-либо массовых событий также будет отображаться в транспортных перемещениях (участникам нужно приехать и уехать, что должно соответствовать отклонению в матрице корреспонденций по сравнению с “обычными” значениями).

В настоящее время указанные матрицы корреспонденции используются просто для суммирования потоков. Это имеет отношение к экономике (перевезено столько-то пассажиров). Технически это не вызывает никаких проблем и полностью понимаемо теми, кто занимается управлением городским хозяйством.

Но совершенно очевидно, что при таком подходе полностью теряется информация о пространственно-временных характеристиках потоков. У нас есть не просто какое-то количество пассажиров, перевезенных от станции А до станции В. Эти пассажиры были как-то распределены по времени, пиковые значения по входу

должны укладываться в возможности станций, работа наземного транспорта должна быть согласована с пиковыми значениями по выходу пассажиров со станции и т.д. Анализ матриц корреспонденции должен обслуживать тех, кто занимается эксплуатацией транспорта, организацией движения и городским планированием.

Именно для них, основной вопрос – это пространственно-временные характеристики потока. Как он распределен на конкретной станции по входу и выходу в течение дня, остается ли такое распределение стабильным, как определить выбросы (аномалии), как понять, что бывшие аномалии стали новой нормой и т.д. Это все имеет самое прямое отношение к тому, что в Умном городе называется мобильность как сервис (мобильность как услуга). Для представления сервиса, естественно, необходимо понимать, как этот сервис (спрос на услугу) устроен.

Следующее важное замечание. На сегодняшний день нет какого-то индикатора (индикаторов/метрик), которые бы определяли, что именно нужно вычислять по матрицам корреспонденций. Можно сказать, что такие работы только начинаются. Интересно, что большая часть существующих работ выполнена в Китае, где в силу очевидных причин (многочисленность пассажиров) вопрос анализа пространственно-временных характеристик потоков (по сути – безопасности перевозок) очень актуален. Из существующих проектов можно отметить, например, транспортное поведение от OECD [10]. Но это опять некоторые интегральные показатели. Такой большой город как Москва передвигается в разных районах по-разному.

Соответственно далее – это наши предложения (предположения), базирующиеся на анализе литературы и предшествующих разработок, относительно того, что могло бы быть использовано для описания характеристик таких матриц. Это именно обсуждение метрик и того, как они могут использоваться. Математические модели (аппарат) для их определения может быть различным.

Во-первых, это, конечно, распределение входов/выходов по времени. Причины важности понимания этих параметров очевидны. Пики в этих распределениях должны укладываться в пропускные способности станций. Работа наземного транспорта должна быть как-то синхронизована с потоками пассажиров. Последовательность нескольких станций с одинаковым утренним пиком по входу, например, может привести к невозможности (затруднению) входа на каких-то станциях. Это означает, что такие распределения должны рассматриваться совместно.

Далее – это аномалии во временных распределениях входов и выходов. Это как раз оценка эффектов от каких-либо событий в городе. Технически – это поиск аномалий во временных рядах.

Следующая возможная характеристика – это подобие распределений входов/выходов по времени. При отсутствии каких-либо изменений маршруты будут стабильны. Отсутствие стабильности – это как раз и есть сигнал о необходимости поиска объяснений этого. Например, возможное различие в распределении для рабочих дней и выходных может помочь в определении точек притяжения (куда ездят пассажиры) выходного дня.

Может представлять интерес исследование зависимости в трафике между станциями. Очевидно, что для большой системы, которой является московское метро, разовое увеличение количества пассажиров, вошедших на конечной станции на севере города, будет “не замеченным” где-то на юге. Но на ближайших станциях это будет, конечно, заметно. По сути – это некоторая структурная характеристика потоков – как связаны потоки на разных станциях. Эти зависимости разные, что проверялось и в наших работах.

В общем, изложенные выше метрики можно представить как структурные описания характеристик использования станций. Соответственно, в финальной точке, мы хотим создать систему, которая сможет описывать характер использования станций, определять аномалии, находить моменты времени, когда по каким-то причинам изменились режимы использования станций (перегонов) и т.д.

Как пример предыдущих работ можно привести следующее. Первый шаг к описанию характера использования станций, это попытаться описать разные типы (режимы) их использования. Режим использования определяется характеристиками потока на станции. Поток для станции – это входы и выходы пассажиров во времени и производные от них – отношения входов и выходов, минимумы и максимумы. Тогда классификация станций – это есть кластеризация на базе данных о потоках. Входы и выходы по времени – это временные ряды. Соответственно, речь идет о кластеризации временных рядов [11, 12].

Построенные кластеры будут разделять станции по моделям (шаблонам) использования. Объяснения того, как кластеры формировались – это некоторая классификация станций. При этом необходимо отметить, что какого-то эталонного классификатора не существует. Проверить такой классификатор можно только по каким-то косвенным признакам. Одним из главных критериев “достоверности” его устойчивость (воспроизводимость на других временных интервалах)

Если получить такой классификатор, то возможные последующие изменения в нем – это изменение режима использования станции. Которое, в свою очередь должно иметь какие-то “городские” причины. Изменение структуры кластеров (переход станций из

одной группы в другую) служит признаком того, что нужно искать причины этого. Соответственно, очень важно, чтобы такой классификатор был “объясняющим”. Причины разделения на конкретные группы должны быть понятны тем, кто отвечает за организацию движения.

Вот в предыдущих работах [2] использовалась информация об утренних и вечерних пиках на станциях. Режимы работы (группы в классификаторе) описывали так:

больше входит (уезжает со станции) утром – больше выходит вечером: “спальный” район – жители уезжают утром на работу, вечером возвращаются

больше выходит (приезжает на станцию) утром – больше входит вечером: “рабочий” район – утром приезжают на работу, вечером уезжают домой

и т.д.

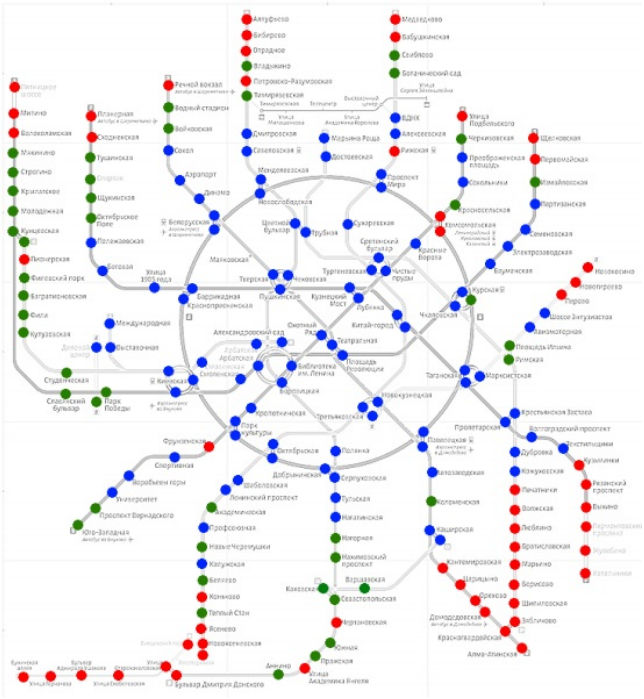

Рис. 1. Зонирование города по пикам вход/выход [2]

Другие классификаторы сравнивали изменение трафика в выходные дни по сравнению с рабочими днями. Станции, связанные с точками притяжения выходного дня (магазины, парки) теряли меньше трафика в выходные [7].

Эти результаты действительно объяснимы, но просто несколько упрощают картину. Например, наши исследования по железнодорожному трафику показали, что пики также могут быть разными. Они более растянуты по времени, если есть подвоз пассажиров, а более острые (выраженные) – если станцией пользуются люди живущие рядом. Причина очевидна – автобусы в случае подвоза не всегда приходят в одно и то же время.

Режим использования станций в выходные дни также может отличаться в зависимости от разных точек притяжения. Например, станции метро около стадионов (Динамо, ЦСКА, Аэропорт, Октябрьское Поле) имеют пики по трафику, связанные именно с проведением футбольных матчей, которых не будет во время перерывов в чемпионате.

Ну и простые соображения о жизни большого города говорят о том, что трафик не столь сильно уже привязан к режиму работы 9:00-18:00. Множество людей имеет свободный график, есть разъездные работы, переработки и т.д.

Соответственно, отсюда и возникла идея о том, что, возможно, было бы правильно описать большее количество характеристик трафика, которые могли бы быть использованы для классификации станций. Например, рассматривать пики трафика в каждый из дней недели, подсчитывать пики внутри дня на часовых или двухчасовых интервалах, учитывать изменение трафика в субботу и воскресенье по сравнению с выходными днями и т.д.

Выписав такой набор потенциальных характеристик, можно провести классический анализ признаков (feature analysis, e.g. [13, 14]), после чего использовать отобранный набор характеристик для, возможно, более точной (по сравнению с указанной выше) классификации станций.

Как можно будет проверить такой классификатор? Например, рассчитать его для зимнего месяца (например, декабрь) и летнего (июнь). По оценкам, выполненным на основе данных мобильных операторов, летняя миграция из Москвы на дачи – это около одного миллиона человек. Это, естественно, должно отражаться на режимах использования станций метро около ЖД вокзалов. Соответственно, такие станции должны, скорее всего, поменять свои группы летом по сравнению с зимой.

Но самое главное – этот классификатор опять-таки должен быть “объясняемым”. Любая группа станций, которую выделит такая кластеризация – это, в итоге, какой-то шаблон по входам и выходам. Смысл (правила) разделения на группы должен быть как-то восстановлен.

III. Заключение

В настоящей статье мы остановились на возможных задачах анализа пассажиропотоков для случаев, когда нам известна матрица корреспонденций. На первый план в этом случае, по нашему мнению, выходят задачи поведенческого анализа. Поведение пассажиров на станции (режим использования станции) определяется распределениями входов и выходов по времени в течении дня, распределением соотношений входов и выходов в течение дня, а также внутренними характеристиками распределений входов и выходов (максимами и минимумами). Необходимо понимание режимов работы станций, устойчивость этих характеристик и, соответственно, детектирование отклонений от некоторых сложившихся моделей (режимов) использования. Временные отклонения будут соответствовать каким-то событиям в городе. Матрица корреспонденции в этом случае будет выступать как индикатор и средство измерения при обнаружении и

оценке событий с точки зрения вовлеченности людей. Структурные изменения в шаблоне использования какой-то станции будут соответствовать каким-то постоянным изменениям в структуре пассажиропотока (например, открытие нового пересадочного узла, нового автобусного маршрута из пригородов и т.д.). В любом случае, матрица корреспонденций играет роль сенсора (измерительного инструмента).

Библиография

[1] Sweeney, Latanya. "k-anonymity: A model for protecting privacy." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05 (2002): 557-570.

[2] Nekraplonna, Mariia, and Dmitry Namiot. "Metro correspondence matrix analysis." International Journal of Open Information Technologies 7.7 (2019): 68-80.

[3] Zhang, Yi, et al. "Daily OD matrix estimation using cellular probe data." 89th Annual Meeting Transportation Research Board. Vol. 9. 2010.

[4] Djukic, Tamara, et al. "Efficient real time OD matrix estimation based on Principal Component Analysis." 2012 15th International IEEE Conference on Intelligent Transportation Systems. IEEE, 2012.

[5] Pucher, John. "Urban travel behavior as the outcome of public policy: the example of modal-split in Western Europe and North America." Journal of the American Planning Association 54.4 (1988): 509-520.

[6] Bulygin M. V., Namiot D. E. Об использовании данных мобильных абонентов в цифровой урбанистике //Международный научный журнал «Современные информационные технологии и ИТ-образование». – 2019. – Т. 15. – №. 3. – С. 755-766.

[7] Поматилов Ф. С., Намиот Д. Е. Об анализе пассажиропотоков Московского метрополитена //Современные информационные технологии и ИТ-образование. – 2019. – Т. 15. – №. 2.

[8] Misharin, A., D. Namiot, and O. Pokusaev. "On Processing of Correspondence Matrices in Transport Systems." 2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon). IEEE, 2019.

[9] Намиот Д. Е., Покусаев О. Н., Лазуткина В. С. О моделях пассажирского потока для городских железных дорог //International Journal of Open Information Technologies. – 2018. – Т. 6. – №. 3.

[10] OECD transport statistics https://www.oecd-ilibrary.org/transport/data/itf-transport-statistics_trsprt-data-en Retrieved: Mar, 2020

[11] Liao, T. Warren. "Clustering of time series data—a survey." Pattern recognition 38.11 (2005): 1857-1874.

[12] Rani, Sangeeta, and Geeta Sikka. "Recent techniques of clustering of time series data: a survey." International Journal of Computer Applications 52.15 (2012).

[13] Wang, Xiaozhe, Anthony Wirth, and Liang Wang. "Structure-based statistical features and multivariate time series clustering." Seventh IEEE International Conference on Data Mining (ICDM 2007). IEEE, 2007.

[14] Hautamaki, Ville, Pekka Nykanen, and Pasi Franti. "Time-series clustering by approximate prototypes." 2008 19th International Conference on Pattern Recognition. IEEE, 2008.

OD-matrix and passenger flow analysis

Dmitry Namiot, Mariia Nekraplonna, Oleg Pokusaev, Alexander Chekmarev

***Abstract*— This article deals with approaches to assessing the use of metro stations based on correspondence matrixes describing passenger movements. Telecom operators currently maintain mobile communication in the metro. This results in operators being able to track the entry and exit of passengers from the metro, determining when their mobile subscriber switches to a base station located in the metro (enters the metro) or to a base station in the city (exits the metro). Such anonymous data can be grouped by time to exclude the tracking of individual subscribers and presented for analysis. The final data are a so-called OD-matrix: for a certain time interval for each pair of stations, the number of passengers moving between these stations is known. Usually, when analyzing transport systems, restoring such a matrix (i.e. actually forecasting passenger flows) is the main task. In this case, the forecast is not needed - all passenger flows are known. This article is devoted to the discussion of what can be the purpose of analysis under such a structure of source data.**

***Keywords*— transport, metro, OD-matrix.**

REFERENCES

[1] Sweeney, Latanya. "k-anonymity: A model for protecting privacy." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05 (2002): 557-570.

[2] Nekraplonna, Mariia, and Dmitry Namiot. "Metro correspondence matrix analysis." International Journal of Open Information Technologies 7.7 (2019): 68-80.

[3] Zhang, Yi, et al. "Daily OD matrix estimation using cellular probe data." 89th Annual Meeting Transportation Research Board. Vol. 9. 2010.

[4] Djukic, Tamara, et al. "Efficient real time OD matrix estimation based on Principal Component Analysis." 2012 15th International IEEE Conference on Intelligent Transportation Systems. IEEE, 2012.

[5] Pucher, John. "Urban travel behavior as the outcome of public policy: the example of modal-split in Western Europe and North America." Journal of the American Planning Association 54.4 (1988): 509-520.

[6] Bulygin M. V., Namiot D. E. Ob ispol'zovanii dannyh mobil'nyh abonentov v cifrovoj urbanistike //Mezhdunarodnyj nauchnyj zhurnal «Sovremennye informacionnye tehnologii i IT-obrazovanie». – 2019. – T. 15. – #. 3. – S. 755-766.

[7] Pomatilov F. S., Namiot D. E. Ob analize passazhiropotokov Moskovskogo metropolitena //Sovremennye informacionnye tehnologii i IT-obrazovanie. – 2019. – T. 15. – #. 2.

[8] Misharin, A., D. Namiot, and O. Pokusaev. "On Processing of Correspondence Matrices in Transport Systems." 2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon). IEEE, 2019.

[9] Namiot D. E., Pokusaev O. N., Lazutkina V. S. O modeljah passazhirskogo potoka dlja gorodskih zheleznyh dorog //International Journal of Open Information Technologies. – 2018. – T. 6. – #. 3.

[10] OECD transport statistics https://www.oecd-ilibrary.org/transport/data/itf-transport-statistics_trsprt-data-en Retrieved: Mar, 2020

[11] Liao, T. Warren. "Clustering of time series data—a survey." Pattern recognition 38.11 (2005): 1857-1874.

[12] Rani, Sangeeta, and Geeta Sikka. "Recent techniques of clustering of time series data: a survey." International Journal of Computer Applications 52.15 (2012).

[13] Wang, Xiaozhe, Anthony Wirth, and Liang Wang. "Structure-based statistical features and multivariate time series clustering." Seventh IEEE International Conference on Data Mining (ICDM 2007). IEEE, 2007.

[14] Hautamaki, Ville, Pekka Nykanen, and Pasi Franti. "Time-series clustering by approximate prototypes." 2008 19th International Conference on Pattern Recognition. IEEE, 2008.