ОБЩИЕ ВОПРОСЫ МЕТОДОЛОГИИ И МЕТОДИКИ СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

Т.Д. Алексеев
(*Москва*)

АНАЛИЗ ПОСЛЕДОВАТЕЛЬНОСТЕЙ В СОЦИОЛОГИИ: ВОЗМОЖНОСТИ, ОГРАНИЧЕНИЯ И ПОТЕНЦИАЛ ПРИМЕНЕНИЯ

В статье представлен краткий обзор применения метода анализа последовательностей в социологии. Обсуждается как контекст появления метода в социальных науках, так и примеры его основных приложений. Анализ последовательностей рассматривается в контексте классификации более широкого набора методов анализа временных данных и оценивается в плане своих сравнительных достоинств и недостатков применительно к разным исследовательским вопросам, типам данных и теоретическим допущениям.

Ключевые слова: анализ последовательностей, временные данные, визуализация временных данных

Постановка исследовательской задачи

Цель, ставящаяся в рамках этой статьи, – обозначение истоков анализа последовательностей и обзор областей его применения в социологии. Говоря об анализе последовательностей, мы имеем в виду комплекс методов и техник, в рамках которых ряды упорядоченных во времени состояний объекта рассматриваются

Тимофей Дмитриевич Алексеев – аспирант Национального исследовательского университета «Высшая школа экономики», преподаватель Государственного академического университета гуманитарных наук. E-mail: veretur@gmail.com.

холистически. Такие объекты сопоставляются и группируются специальными алгоритмами, которые выступают в качестве альтернативы статистическим методам анализа или в качестве дополнения к ним.

Когда социологические объекты меняют свои дискретные свойства во времени, исследователь сталкивается с характерными проблемами. От чего зависит состояние в данный момент времени? Имеет ли значение порядок событий, и если да, то в каком масштабе? Какие теоретически интерпретируемые свойства последовательности как целого можно измерить? Как именно их лучше измерять? Имеет ли значение время между событиями? Как лучше группировать последовательности? В чем, собственно, заключается сходство одной последовательности с другой?

В анализе данных подобного рода используются хорошо известные общие линейные модели. К примеру, в демографии, исследуя совокупность людей, для которых известны матримониальные статусы в некоторых временных периодах, мы можем сделать попытку предсказать вероятность наступления искомого статуса в интересующем нас периоде. Так же, например, мы можем оценить вероятность продвижения по карьерной лестнице. Для того чтобы сделать подобное предсказание, нужно принять ряд допущений о природе данных и, что важнее, о форме их связи. Например, предсказывая, возьмет ли Зевс в жены Фетиду в наблюдаемом периоде (зависимая переменная), мы рассматриваем только три дискретных предшествующих периода (независимые переменные) в жизни Зевса: в каждом из них он женится на Метиде, Фемиде и Гере соответственно. Нас также могут заинтересовать эффекты взаимодействия предыдущих браков, которые могут войти в модель в качестве отдельных независимых переменных. Так, возможно, проанализировав брачные сценарии других олимпийцев, мы выяснили, что каждый отдельный брак в прошлом положительно влияет на вероятность брака в наблюдаемом периоде, так как определенный матримониальный опыт понижает

тревожность перед поздним браком. Однако браки в первых двух и в первых трех периодах не аддитивны: молодые боги быстро пресыщаются отношениями подобного рода, и в четвертый раз в случаях частых союзов в начале пути в брак вступают редко. Такая модель хорошо сочетается с тем фактом, что Гера действительно останется последней женой Зевса.

Приведенный пример очень примитивен: он не учитывает времени пребывания в браке, фактора многоженства, значения кратковременных романов, типов партнеров. Более того, хорошее объяснение того, почему Зевс не женился на Фетиде, связано с событиями, которые вообще не входят в модель: Громовержец убоялся пророчества, гласящего, что их сын превзойдет отца по силе. Но какой бы сложной модель ни была, она всегда одна. Соответственно, регрессионные модели ориентированы на поиск некоей всеобщей типичности, и она тесно связана с порядком событий. Если бы мы рассматривали не только жен, но и всех любовниц Зевса, то предсказание могло радикально измениться в связи с перестановкой порядка их появления. Но можно ли бы было судить о том, что его романтическая биография в своей завершенности претерпела в связи с этим существенные изменения? Вряд ли: Зевс бы в любом случае остался тем же любвеобильным богом, вне зависимости от конкретного размещения романов и пауз между ними в рассматриваемых периодах. Что важно в его судьбе как таковой, а не в ее исходной точке? Регрессионные модели не очень хорошо подходят для поиска ответов на такие вопросы. В социологии анализ последовательностей исторически позиционировался в качестве ресурса рассмотрения рядов событий в их целостности, к классификации последовательностей со вниманием к неожиданным вариациям, которые модели могут отнести к «необъясненной дисперсии». В ряде своих приложений такой анализ действительно удачно решает задачи подобного рода.

Дисциплинарные источники и области использования анализа последовательностей

С 1980-х гг. в социологических исследованиях начал применяться метод анализа последовательностей, исходно разработанный и применявшийся преимущественно в биологии и лингвистике [1]. Появившись как оригинальный подход, альтернативный уже давно на тот момент применявшимся в социальных науках разнообразным техникам анализа временных данных, за первые десять лет использования и теоретического обсуждения он снискал как благосклонность ряда поклонников и апологетов [2] среди социологов, так и критику со стороны скептиков [3; 4]. История появления этого метода в социологии проясняет сравнительные особенности методических альтернатив для аналогичных типов данных и задач, а также основания для отнесения отдельных техник к анализу последовательностей.

Анализ последовательностей – междисциплинарный метод, направленный на классификацию сценариев изменения состояний или отдельных характеристик объектов во времени. Будучи модным в социальных науках сегодня, до 1980-х гг. он в основном развивался в рамках биоинформатики [5, p. V–VI] и в некоторой степени − в лингвистике [5, p. 163–188; 6; 7]. Знакомство пионера анализа последовательностей в социологии – Э. Эббота – с приложениями метода в других науках состоялось через Дж. Краскала, который готовил издание сборника о методах обработки речи и анализа макромолекулярных последовательностей («Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison») [8, p. 8–9]. Анализ последовательностей исходно разрабатывался как специфичный метод именно для биоинформатики [9]. В связи с этим импорт метода в социальные науки сопровождался спорами о теоретической обоснованности используемых процедур, которые были исходно связаны с содержательными представлениями о природе биологических данных.

Единицей анализа в биологии являются последовательности генов или их структурных элементов в нуклеиновых кислотах [5, p. 45–52]. Стандартные задачи, решаемые в молекулярной биологии, эволюционной биологии и биоинформатике с его помощью – построение филогенетических деревьев, аннотирование геномов, сбор баз генетического разнообразия [10; 11]. Такой анализ направлен прежде всего на выявление выделяющихся из общего контекста последовательностей качественных признаков (поиск первых мутаций) и на каталогизацию типических порядков событий и перерывов между ними (задачи биологической таксономии). Содержательно эти задачи позволяют уточнять таксоны и выдвигать обоснованные предположения об эволюционной близости видов [12, с. 430–433]. Например, первые найденные представители семейства дамановых в связи с фенотипическим сходством с морскими свинками были отнесены к отряду грызунов. Позднее структурные особенности генома этих животных помогли отнести их к одному классу со слонами и ламантинами: их общий предок, как это отражалось в сходных частях генома, мутировал в разных направлениях [13].

Основные предметные области, где используется анализ последовательностей в социологии, – жизненный путь, карьеры, взаимодействие лицом к лицу, структурные исторические изменения, брачный статус индивида [14, p. 4]. Выбор единицы анализа в них сильнее варьирует вместе с особенностями исследовательского дизайна, чем в биологии. Элементы последовательностей – события – в таких концептуально сложных объектах могут обладать свойствами, которые не позволяют, например, симметрично их заменять. А эта операция, как станет ясно в дальнейшем, – основа стандартных алгоритмов построения расстояния между последовательностями.

Конечно, в истории статистики известно множество любопытных анекдотов[1] о контексте создания того или иного метода

[1] Наиболее знаменитый за пределами естественных наук, вероятно, повествует о том, как Уильям Госсет опубликовал под псевдонимом Стьюдент статью, чтобы

анализа, что располагает подозревать проблемную и теоретическую укорененность в любом математическом формализме. Важно, однако, что прикладная статистика все же ориентирована на универсализацию разрабатываемых методов для разных научных задач, а разработка анализа последовательностей была более дисциплинарно герметичной. Задача использования метода в биологии состоит в том, чтобы создать критерий для сопоставления разных биополимеров (в первую очередь – нуклеиновых кислот) [5, p. 45–46]. С развитием методов считывания данных с генетического материала и увеличением вычислительных мощностей, автоматические техники секвенирования и сборки генома позволяют определять, т. е. кодировать в текстовые последовательности, все более протяженные последовательности оснований нуклеотидов в ДНК или РНК [17], которые в свою очередь могут быть сопоставлены с аналогичными последовательностями у разных биологических видов.

Процедуры «раннего» анализа последовательностей

Поскольку в социологии анализ последовательностей пока не используется на таких больших объемах данных, которые требуют особых техник автоматизированного отбора *внутри* отдельно взятой последовательности, мы рассмотрим детали этапа анализа, а не сбора. Подготовленная единица анализа – упорядоченный набор неких дискретных состояний. Рассмотрим базовые стадии,

избежать судебного преследования со стороны пивоварни Эдварда Гиннесса, для которой разработал тест и распределение в целях анализа процессов вызревания пива [15, p. 234–235]. Чуть менее известная, но более элегантная история рассказывает о том, что точный тест Фишера появился благодаря пари Фишера со своей коллегой – Мьюриэл Бристоль, – состоявшем в том, сможет ли она определить порядок наливания молока и чая в чашку в краткой серии слепых тестов [16].

характерные для биоинформатики и, соответственно, для раннего этапа использования метода в социологии.

1. Применение алгоритма выравнивания последовательностей («sequence alignment») с целью сопоставления единиц. Он позволяет отыскать количественную меру различий между всеми парами последовательностей и создать основу для их классификации. Определив вес операции замены знака, ввода или удаления знака, можно применить алгоритм, который найдет наименьшие веса приведения каждой последовательности к некоей другой[1] [14, p. 109–111]. В силу совокупности таких особенностей биологических данных, как большая длина последовательностей[2], большие объемы неклассифицированной информации[3], значительная межвидовая вариация длин последовательностей и косвенные дорогостоящие способы реконструирования целых последовательностей, выравнивание в биологии до сих пор остается развивающимся пространством методического и инженерного новаторства [20].

В социологии, где последовательности не столь неохватны и несопоставимы по длине, один из методов выравнивания можно считать общепринятым – это оптимальный подбор («optimal matching») с помощью алгоритма Нидлмана–Вунша. Этот алгоритм динамического программирования предназначен для выравнивания

[1] Мера удаленности называется дистанцией редактирования, а включающая именно три этих операции – при равных весах каждой из них – расстоянием Левенштейна [5, p. 37–39] [18]. Она удовлетворяет неравенству треугольника, но не транзитивна.

[2] Только генов, несущих наследственную информацию, геном человека насчитывает более 19-ти тысяч [19]. Пары нуклеотидов в цепочке ДНК человека измеряются миллиардами.

[3] Представление о доле «мусорного» ДНК в молекуле кардинально изменилось за последние 50 лет. В эпоху первых шагов биоинформатического анализа последовательностей научное сообщество разделяло мнение, что подавляющее большинство последовательностей генома не кодирует наследственную информацию.

всех последовательностей по всей длине по уже описанному выше принципу. Например, мы хотим определить расстояние Левенштейна (дистанцию редактирования) между последовательностями знаков «ПАТЕФОН» и «ТЕЛЕФОН», для чего нужно заменить подпоследовательность «ПАТ» в «ПАТЕФОН» на «ТЕЛ». Если бы операция замены знака была «дорогой» в сравнении с операциями ввода и удаления, более оптимальным способом мы бы сочли ввод двух знаков для получения «ПАТЕЛЕФОН» и удаление из него двух знаков – подпоследовательности «ПА». Сопоставив массу лексем подобным образом, мы могли бы упорядочить их в схожие и отличные группы, чтобы выдвинуть некую содержательную гипотезу о природе их морфологии. Наиболее «креативная» часть в социологическом анализе последовательностей – выбор весов для операций (например: [8, p. 479]) и ввод новых операций[1]. Динамика занимаемых должностей, очередность инициативы в разговоре, история матримониального статуса, образовательного статуса – лишь список примеров единиц социологического анализа, и каждая требует гибкости в выборе параметров алгоритма. Так, индекс может вообще не основываться на количестве вводов, удалений и замен, необходимых для приведения одной последовательности к другой. Например, есть мера сходства, в основе которой лежит доля общих подпоследовательностей [22]. В случае с Зевсом это бы значило, что он очень «похож» на некую свою вымышленную версию, взявшую в жены не только Метиду (A), Фемиду (B) и Геру (C), но и десятки остальных своих возлюбленных, так как все семь подпоследовательностей (A, B, C, AB, AC, BC, ABC) браков Зевса есть у его более ответственного двойника. При помощи дистанции редактирования получился бы иной результат: десятки операций удаления тех жен, которые у реального Зевса

[1] Например, описан алгоритм вычисления дистанции редактирования – расстояния Дамерау–Левенштейна – при вводе операции перестановки двух соседних элементов.

были лишь возлюбленными, обеспечили бы представление об этих двух версиях одного бога как об очень разных.

2. Кластеризация. Мы уже обращали внимание, что стандартной содержательной целью биологического исследования с применением анализа последовательностей является, в частности, конструирование филогенетических деревьев, которые состоят из обширных кластеров, значительно удаленных друг от друга в сконструированном при помощи выравнивания последовательностей и попарного замера расстояния между ними в метрическом пространстве дистанций редактирования. Дистанция редактирования положительно коррелирует с разницой длин сравниваемых последовательностей, поэтому, чаще всего, дистанции нормируются посредством деления на число символов в длиннейшей последовательности из пары [8, p. 482]. В качестве альтернативы можно обеспечить одинаковые длины последовательностей на этапе дизайна [23]. В дальнейшем последовательности пошагово, по одной агломерируются по принципу наименьшего расстояния между кластерами (в том числе состоящими из одной последовательности). Существует множество алгоритмов определения ближайшего кластера [24, С. 251–254, 260–267], детали которых – предмет отдельного обсуждения.

Место анализа последовательностей в ряду методов анализа временных данных

Пока анализ последовательностей существовал в социологии в описанном выше виде, он едва ли конкурировал с другими представителями широкого класса методов рассмотрения динамики событий во времени. Действительно, долгое время – приблизительно на протяжении своего существования в социологии в XX в. – он оставался не очень хорошо разработанной техникой. Открытая критика его ограниченности в сравнении с хорошо развитыми линейными моделями анализа истории событий была вынесена на обсуждение [3; 4] только в публикациях 2000 г. (специальный

выпуск «Sociological Methods & Research»). Систематические учебные пособия [14; 25] по использованию метода в социологии и пакеты для программ анализа данных [26] стали появляться не раньше десяти лет назад, а в большинстве своем − после 2010 г.

Далее мы кратко рассмотрим основных «конкурентов» анализа последовательностей в области методов обработки временных данных и возможности новых техник, которые иногда тоже обобщаются в термин «анализ последовательностей» [27]. Здесь мы сделаем попытку создать набросок классификации методов по двум основаниям: по типу используемых процедур статистического вывода и по подходу к рассмотрению времени. Прежде следует сделать две существенные оговорки. Во-первых, классификация не будет исчерпывающей. Нас здесь интересуют выделяющиеся на общем фоне методов анализа временны́х данных черты анализа последовательностей, его достоинства и недостатки, возможности и ограничения на сегодняшний день, а для этих целей достаточно некоторого количества иллюстраций. Во-вторых, большее значение, чем сами методы, имеют, пожалуй, именно признаки, по основаниям которых можно провести классификацию, потому что в них отражается сложность теоретико-методологической проблемы изучения времени как независимой и зависимой переменной в социологии.

Приведенная выше типология – набросок, подчеркивающий кажущиеся нам значительными различия в подходах к интерпретации временно́й динамики. Методы, отнесенные к разным ячейкам, не обязательно применяются для решения одних и тех же задач и идентичных типов данных. *Табл.* 1 дает общую картину различных способов концептуализации временны*х* последовательностей.

Различие между строками *табл.* 1 состоит в том, какими признаками обладает событие. Как структурный элемент или единица анализа, оно не только принимает некоторое значение, но и вступает

Таблица 1

МЕТОДОЛОГИЧЕСКАЯ КЛАССИФИКАЦИЯ ТИПОВ АНАЛИЗА ВРЕМЕННЫХ ДАННЫХ

Определения события	Способ получения результата		
	Классическая параметрическая статистика	Непараметрические статистические подходы	Алгоритмический
Время	Анализ истории событий	Последовательный анализ временного окна*	Анализ последовательностей (поздний)
Порядок	Цепи Маркова с дискретным временем	Тест серий	Анализ последовательностей (ранний)

Примечание. *«Time-window sequential analysis».

в отношения порядка[1] с другими событиями. Безусловно, порядок в последовательности всегда предполагает временной порядок, либо как прямо описанная переменная (ряд событий во времени), либо как имплицитно привязанный к временной шкале признак, имеющий отношение как минимум к порядку чтения или анализа. Даже обсуждавшиеся в начале статьи последовательности нуклеотидов ДНК не лишены временного измерения, так как ДНК-полимераза «читает» их только в одном направлении, следовательно, можно определенно сказать, какой элемент появился в более ранней фазе репликации. Между тем, существует большая разница между теми методами, аппарат которых включает физическое время в виде непрерывной переменной или набора дискретных временных интервалов известной протяженности (сколько времени прошло между событиями), и теми, где этот признак задан на ординальной шкале (в каком порядке произошли события).

Различие между столбцами состоит в основном в типе результата. Так, вероятностные методы (первые два столбца) предполагают применение статистического критерия, т. е. проверку гипотез о статистической связи или различиях, или статистическое оценивание. При этом параметрические критерии вводят допущения о параметрах распределений, выборкой из которых стали события в последовательности. Непараметрические тесты строят свои предположения «от данных». Нужно указать, что разделение на параметрические и непараметрические методы несколько условно, потому что отдельные требования к данным могут быть выражены как в отсылках к параметрам теоретического распределения, так и в чисто процедурной форме[2]. В любом случае, вероятностные методы разделяют подход к

[1] Возможно, применительно к событиям корректнее говорить не о порядке, а о предшествовании, так как строгий порядок транзитивен, нерефлексивен и асимметричен, что невозможно при повторении одного элемента в разных сегментах последовательности [22].

[2] Например, анализ простых таблиц сопряженности классифицируется как непараметрический метод, в то время как логлинейные модели, в сущности

последовательности как к паттерну, складывающемуся из результатов стохастических процессов, лежащих в основе каждого события.

Анализ последовательностей, который мы относим к «алгоритмическому» типу вывода, отличается тем, что в него не встроен явный или принципиально необходимый вероятностный аппарат. В этом случае мы имеем дело с конструированием принципов (алгоритмов) сопоставления целостных последовательностей. Один алгоритм измеряет отношения между элементами множества последовательностей (например, вычисляет дистанцию редактирования), а другой алгоритм определяет принципы агрегирования (скажем, методом полной связи). Соответственно, анализ последовательностей тяготеет к «детерминистскому» подходу к объектам, в отличие от стохастического понимания в статистических моделях и тестах.

Если обращаться к конкретным примерам из *табл*. 1, цепи Маркова с дискретным временем чаще всего используются при определении единого шага времени между событиями, но это условие не является необходимым. Так, например, этот метод исторически использовался в текстологии для исследования тенденций в расстановке гласных и согласных [30, с. 188].

Анализ истории событий представляет собой пример использования общей линейной модели. В анализе истории событий используется регрессия по Коксу, также известная как анализ наработок на отказ, в целях предсказания вероятности наступления некоего состояния в данный момент времени. Предикторами выступают предыдущие периоды времени, принимающие такие значения, как определенные события (подробнее см.: [28; 29, p. 33–41])[1].

обобщающие такой анализ на многомерные распределения, часто относят к параметрическим методам, так как они требовательны к ожидаемым частотам.

[1] Яркий пример – модель пропорциональных рисков, используемая в биомедицинской статистике для предсказания вероятности смерти пациента. Предикторами выступают периоды времени, когда применялись или не применялись некие терапевтические приёмы.

Для непараметрических методов анализа временных данных характерная эксплораторная направленность, например последовательная модификация гипотез в ходе продолжающегося сбора серийных данных. Таким образом, в группу попадают многие методы статистического последовательного анализа, один из самых простых из них – критерий Вальда–Волфовица (тест серий), проверяющий гипотезу о случайности отбора наблюдений в последовательности, например, корректно ли использовать для этой последовательности марковскую модель. Более сложным методом, учитывающим продолжительности событий и возможность их полного или частичного совпадения во времени, представляется последовательный анализ временно́го окна. Он используется в психологии для анализа обусловленного поведения [31, p. 135–137].

Анализ последовательностей, который мы отнесли к алгоритмическим методам, как известно из предыдущего раздела статьи, явным образом не учитывает количественных временных интервалов. Но этот подход начал дополняться с 2000-х гг. с развитием новых техник обобщения и визуализации в рамках анализа последовательностей. О них мы подробнее расскажем в следующем разделе.

Направления развития анализа последовательностей в XXI в.

Критика раннего анализа последовательностей основывалась в основном на «атеоретичности» алгоритмов, лежащих в его основе, т.е. отсутствию очевидной интерпретации их теоретического смысла [22, p. 4–8]. Как один из равноправных методов со своими достоинствами и недостатками, он мог бы сегодня оставаться таким, каким был знаком исследователям в 1990-х гг., потому что атеоретичность не может служить безусловным препятствием для эксплораторного использования техники. Любопытно рассмотреть

рост числа опубликованных работ, посвященных его дальнейшей методической разработке, после выхода упоминавшегося выше специального выпуска «Sociological Methods & Research» в 2000-м г. в контексте уже существовавших аргументов за использование анализа последовательностей.

Э. Эббот, горячий энтузиаст применения анализа последовательностей в социологии, известен также как автор работ по методологии, сравнивающих каузальные и детерминистские перспективы рассмотрения временны́х данных [32; 33, p. 53–63]. Аргументация за использование анализа последовательностей в качестве альтернативы вероятностным методам работы с временными данными основывалась на критике общей линейной модели с точки зрения исторической очевидности. Моделирование социального процесса предполагает неявное допущение , что процесс, изучаемый на массе *разнообразных* историй, подчиняется единому шаблону, который эксплицируется в модели. Другими словами, вариация возможных сценариев представляется статистической ошибкой, которая говорит не о качественных различиях в пространстве динамики развертывания изучаемого процесса, а о приемлемо низком качестве содержательной модели. Также моделирование предполагает универсальную значимость всех предикторов – событий в данных периодах – для конечного исхода. «Субъектами» действия в статистических методах являются переменные, в то время как для многих социальных процессов – карьер, взаимодействия лицом к лицу – это не вполне очевидно. Исторический подход к таким феноменам предполагает, что действующим лицами будут сами объекты [33, p. 54–59]. Также статистические методы обладают «короткой памятью»: в данном периоде не учитываются события всех предыдущих периодов [32, p. 144–146].

Наша гипотеза состоит в том, что дальнейшее развитие метода и переформулирование принципов его применения не в последнюю очередь было связано именно с расхождениями достоинств техники с дискурсивной стратегией, в рамках которой

оформлялись призывы к его использованию. Действительно, детерминистский и исторический взгляд на единицы анализа не очень хорошо сочетается с марковскими допущениями. Но он так же не сочетается с измерением различия последовательностей расстоянием Левенштейна, которое совершенно не чувствительно к порядку событий и их качественным характеристикам, как и с их автоматической классификацией («без учителя»), исключающей предварительную теоретическую работу с типами последовательностей. Интерпретация кластера, полученного в результате автоматической классификации, разделяет эпистемологические недостатки с общей линейной моделью.

Таким образом, запрос на реформирование раннего анализа последовательностей был связан с тем, что он не избегал проблем тех методов, в виде альтернативы которым позиционировался [3]. После начала 2000-х гг. анализ последовательностей отличался, во-первых, алгоритмическими нововведениями, которые расширили возможности теоретического осмысления результатов его применения, во-вторых, явной спецификацией переменной времени и, в-третьих, общей переориентацией цели использования метода. Последний пункт является наиболее масштабным изменением. Полемические стратегии защиты и продвижения анализа последовательностей через критику общей линейной модели сыграли важную роль в рецепции метода в социологии, но метод требовал и позитивного определения своего предназначения, несводимого к конкурентным преимуществам в сравнении с другими техниками. Так анализ последовательностей постепенно стал развиваться как иллюстративный метод, способный дополнять предсказательные модели качественной информацией, в первую очередь внятной визуализацией. Ниже мы перечислим некоторые важные нововведения, происшедшие в XXI в. в анализе последовательностей, но опустим как их подробное обсуждение, так и описание только

начинающих развиваться техник[1], поскольку сложность таких конкретных техник требует их анализа в отдельных публикациях.

Новые алгоритмы для вычисления меры сходства последовательностей искали альтернативы алгоритмам выравнивания, которые указывают на расстояние между наблюдениями в зависимости от того, как сложно преобразовать одно в другое. Например, в качестве меры использовались доли совпадающих подпоследовательностей в сравниваемых парах («NMS[2] approach») [22] или евклидово расстояние между последовательностями, каждый из элементов которых перекодирован в дихотомические признаки (геометрическое сравнение) [34]. Хотя последняя мера сильно коррелирует с получаемой алгоритмом Нидлмана–Вунша [14, p. 149–150] и так же не учитывает временно́й порядок, первая в чем-то отвечает запросу на теоретическую обоснованность (похожими здесь оказываются те последовательности, в которых фигурирует больше сходных сценариев в смысле одинаковых начала, середины и конца).

Также в рамках позднего анализа последовательностей разрабатываются одномерные статистики для отдельно взятых последовательностей: простые метрики, как число переходов и число подпоследовательностей, и более сложные показатели[3], такие как внутренняя энтропия последовательности [26, p. 22–23], турбулентность [35, p. 231–234] и индекс сложности («complexity index») [26, p. 23]. Внутренняя энтропия последовательности

[1] Например, сетевой анализ последовательностей или статистический вывод в анализе последовательностей. [14, p. 257–262].

[2] Number of Matching Sequences, т.е. число совпадающих последовательностей.

[3] Все они с небольшой разницей в подходах к построению формул отражают разнообразие и непредсказуемость последовательности. Например, энтропия максимальна при большом числе разных событий и низкой дисперсии времени между ними, а турбулентность максимальна при высокой дисперсии времени между большим числом событий.

отражает разнообразие событий в одном наблюдении. π_i – доля событий со статусом *i* в последовательности, каждый из элементов которой принимает значение от 1 до *a*. Соответственно, энтропия максимальна при равной длительности пребывания объекта наблюдения в каждом из *a* статусов, и минимальна при отсутствии смены статуса.

$$h = -\sum_{i=1}^{a} \pi_i \log \pi_i$$

Турбулентность, как и энтропия, отражает разнообразие событий. Турбулентность велика при большом числе отдельных подпоследовательностей (φ). Также она велика при низкой дисперсии периодов, непрерывно проведенных в одном статусе $\left(s_t^2\right)$.

$$T = \log_2\left(\varphi \frac{s_{t,\max}^2 + 1}{s_t^2 + 1}\right)$$

Индекс сложности растет как в связи с энтропией (*h*), скорректированной на максимальную возможную ($\log a$) при *a* возможных статусов, так и в связи с числом смен статуса (l_d) относительно длины последовательности (l_d).

$$C = \sqrt{\frac{l_d}{l} \frac{h}{\log a}}$$

Ввод этих показателей в оборот расширил познавательные возможности метода с точки зрения использования для решения задач квантифицированной оценки стандартизации или дестандартизации социальных процессов. Также этот уровень анализа согласуется с одним из существенных вопросов о наличии качественных паттернов последовательностей и измерении этого качества [25, p. 140–143] – первые приложения анализа последовательностей с оптимальным подбором и иерархической агломерацией не позволяли перевести обсуждение в такую плоскость.

Ввод континуальной переменной времени в анализ стал толчком для развития техник визуализации, широким применением которых и отличается современный анализ последовательностей. Базовая техника визуализации – это совокупная диаграмма последовательностей («sequence index plot», *рис*. 1), где по вертикальной оси расположены объекты, отсортированные по времени наступления одного из значимых событий, а горизонтальная ось отражает время [36].

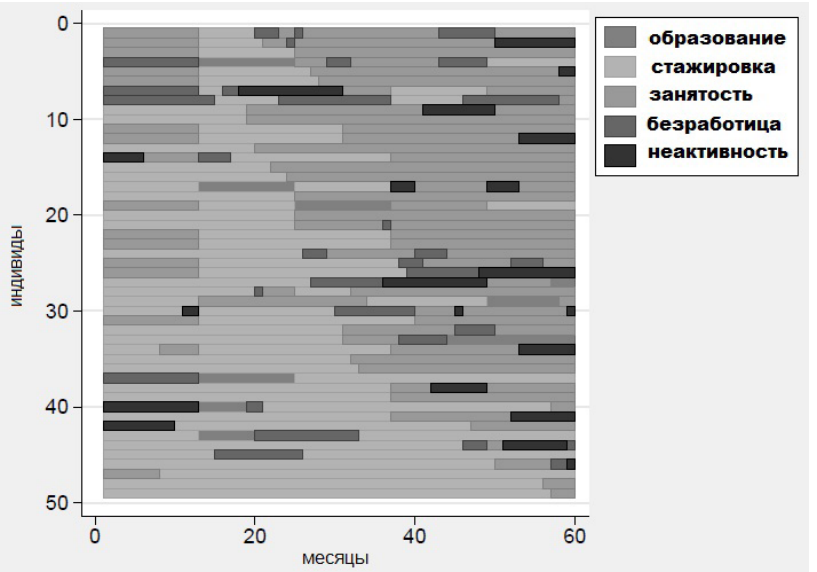

Рис. 1. Совокупная диаграмма последовательностей статусов занятости (вывод STATA)

Источник: [37].

Нахождение в том или ином статусе в данном периоде времени отражается соответствующим цветом отрезка последовательности. Недостаток этой техники заключается в невозможности усмотреть

индивидуальные значения[1] даже при скромном объеме выборки ($N = 200$) [1, p. 651–652] и в зависимости интуитивной понятности такого представления результатов от однородности данных. Для такого типа диаграмм существует модификация с использованием многомерного шкалирования [38], где строки упорядочиваются по значениям первого фактора (т.е. фактора, объясняющего большую часть дисперсии меры сходства последовательностей), полученного в результате многомерного шкалирования матрицы расстояний между наблюдениями. Такие графики в среднем более упорядочены, чем простые совокупные диаграммы последовательностей, так как фактор зависит от многих признаков, и имеет шансы быть более равномерно распределенным в наблюдаемой совокупности, чем время наступления одного из событий.

Наравне с совокупной диаграммой последовательностей используется частотная диаграмма последовательностей («sequence frequency plot»), где вертикальная ось указывает долю последовательностей в выборке. В такой диаграмме отражаются последовательности, которые встречаются чаще всего. Этот график очень чувствителен к мелко квантуемому времени, а также представляется хорошим методом визуализации только при большом объеме групп идентичных последовательностей [1, p. 653]. Глубоким усовершенствованием частотных диаграмм стали представительные диаграммы последовательностей, куда входят те последовательности, которые представительны с точки зрения мер частоты, плотности «соседей», центральности (наименее отличающиеся от других последовательностей), правдоподобия (вероятности каждого из событий последовательности произойти в данный момент) [39]. Достоинство такой диаграммы в том, что можно вычислить долю последовательностей, которую на заданной дистанции покрывают эти «типичные» образцы.

[1] К тому же часть наблюдений может не содержать искомое событие и быть упорядочена произвольно.

Также для последовательностей существует класс диаграмм, которые обобщают значимые показатели. Это простая диаграмма среднего времени для каждого статуса (время по вертикальной оси, типы событий по горизонтальной оси), а также диаграмма модального статуса на данный момент времени (*рис*. 2) [1, p. 657]. Более сложными представляются диаграмма распределения статуса (*рис*. 2), демонстрирующая все статусы и их долю в данный момент времени, и кривые выживаемости, указывающие на вероятность (по вертикали) покинуть группу с неким статусом в данном периоде (по горизонтали) [1, p. 656].

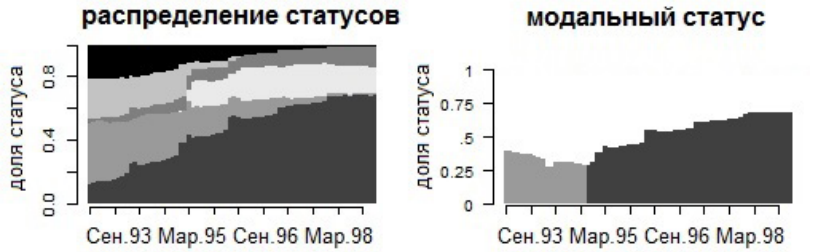

Рис. 2. Обобщающие диаграммы в анализе последовательностей (вывод *R*)

Источник: [40].

Заключение

Унаследованный из приемов решения описательных задач классификации в эволюционной биологии, в своем современном положении в социальных науках анализ последовательностей стал как успешным вспомогательным, так и полноценным самостоятельным методом. Он легко сочетается внутри одного исследовательского дизайна с другими методами, например, предполагающими статистический вывод, но при этом достаточно разветвлен, чтобы оказаться подходящим для решения некоторых

оригинальных задач, ориентированных на детерминистский подход к временным данным. Так, внутри анализа последовательностей появились как самостоятельные индексы для количественной оценки особенностей последовательности как целого, разработан ряд специфических для социологии алгоритмов классификации и вычисления оснований для классификации последовательностей. Обращаясь к содержательным аспектам применения метода, мы можем увидеть, что он применяется в основном в анализе карьерных данных, статусов занятости и брачности, однако не ограничивается таковыми, находя применение в анализе разговоров, пространственной динамике, изменений в документах, исследованиях публикационной активности издательств, и более экзотических областях эмпирической социологии.

ЛИТЕРАТУРА

1. *Fasang A.E. Liao T.F.* Visualizing Sequences in the Social Sciences: Relative Frequency Sequence Plots // Sociological Methods & Research. 2014. Vol. 43(4). P. 643–676.

2. *Abbott A., Tsay A.* Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect // Sociological Methods and Research. 2000. Vol. 29(1). P. 3–33.

3. *Wu L.L.* Some Comments on «Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect» // Sociological Methods and Research. 2000. Vol. 29(1). P. 41–64.

4. *Levine J.H.* But What Have You Done for Us Lately? Commentary on Abbott and Tsay // Sociological Methods and Research. 2000. Vol. 29(1). P. 34–40.

5. Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison / Ed. D. Sankoff, J. B. Kruskal. Reading, Mass.: Addison-Weasley, 1983. (The David Hume Series).

6. *Прокофьев П.А.* Классификация фрагментов текстов с описанием зависимостей правилами на интерпретируемом экспертами языке // Вестник ВГУ, Серия: Системный анализ и информационные технологии. 2012. № 1. С. 174–178.

7. *Ройтберг М.А.* Биоалгоритмика. [Электронный ресурс] // Компьютерра. 2001. №36 (413). URL: http://old.computerra.ru/2001/413/197950/ (дата обращения: 12.05.2017).

8. *Abbott A.* Time Matters: On Theory and Method. The University of Chicago Press, 2001.

9. *Дурбин Р. и др.* Анализ биологических последовательностей: вероятностные модели белков и нуклеиновых кислот / Р. Дурбин, Ш. Эдди, А. Крог, Г. Митчинсон; пер. с англ. А. Миронова. М.: Ин-т компьютерных исследований, 2006.

10. *Gonnet G.H.* Surprising Results on Phylogenetic Tree Building Methods Based on Molecular Sequences [online] // BMC Bioinformatics. 2012. Vol. 13.

11. *Наумов Д.Г.* Филогенетический анализ семейства белков-гомологов [Электронный ресурс] // Zbio. 2006. URL: http://zbio.net/bio/001/003.html#a6 (дата обращения: 12.05.2017).

12. *Гасфилд Д.* Строки, деревья и последовательности в алгоритмах: информатика и вычислительная биология [Электронный ресурс] / Пер. с англ. И. В. Романовского. СПб: Невский Диалект, 2003. URL: http://padaread.com/?book=10289 (дата обращения: 12.05.2017).

13. *Bininda-Emonds O.R.P., et. al.* The Delayed Rise of Present-Day Mammals // Nature. 2007. Vol. 446 (7135). P. 507–512.

14. *Cornwell B.* Social Sequence Analysis: Methods and Applications. NY: Cambridge University Press, 2015.

15. *Dodge Y.* The Concise Encyclopedia of Statistics. NY: Springer-Verlag, 2010.

16. *Fisher R.A.* Mathematics of a Lady Tasting Tea // The World of Mathematics. Vol. 3. P. 1514–1521.

17. *Staden R.* A Strategy of DNA Sequencing Employing Computer Programs // Nucleic Acids Research. 1979. Vol. 6 (7). P. 2601–2610.

18. *Левенштейн В.И.* Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады АН СССР. 1965. Т. 163. № 4. С. 845–848.

19. *Ezkurdia I.,* et. al. Multiple Evidence Strands Suggest that There May Be as Few as 19,000 Human Protein-Coding Genes // Human Molecular Genetics. 2014. Vol. 23 (22). P. 5866–5878.

20. *Мельников Б.Ф., Панин А.Г.* Параллельная реализация мультиэвристического подхода в задаче сравнения генетических последовательностей // Вектор науки ТГУ. 2012. № 4 (22). С. 83–86.

21. *Abbott A., Forrest J.* Optimal Matching Methods for Historical Sequences // Journal of Interdisciplinary History. 1986. Vol. 16(3). P. 479–494.

22. *Elzinga C. H.* Sequence Similarity: A Nonaligning Technique // Sociological Methods and Research. 2003. Vol. 32(1). P. 3–29.

23. *Stovel K., Savage M., Bearman P.* Ascription into Achievement: Models of Career Systems at Lloyds Bank, 1890–1970 // American Journal of Sociology. 1996. Vol. 102 (2). P. 358–339.

24. *Айвазян С.А., Бухштабер В.Р., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: Классификация и снижение размерности. М.: Финансы и статистика, 1989.

25. Advances in Sequence Analysis: Theory, Method, Applications / Ed. P. Blanchard, F. Bühlmann, J.–A. Gauthier. NY: Springer, 2014. (Life Course Research and Social Policies).

26. *Gabadinho A., et. al.* Analyzing and Visualizing State Sequences in R with TraMineR // Journal of Statistical Software. 2011. Vol. 40(4).

27. *Abbott A.* Sequence Analysis: New Methods for Old Ideas // Annual Review of Sociology. 1995. Vol. 21. P. 93–113.

28. *Кокс Д.Р., Оукс Д.* Анализ данных типа времени жизни / Пер. с англ. О.В. Селезнева, под ред. Ю.К. Беляева. М.: Финансы и статистика, 1988.

29. *Allison P.D.* Event History Analysis: Regression for Longitudinal Event Data. Iowa City: Sage Publications, 1984. (Quantitative Applications in the Social Sciences).

30. *Майстров Л.Е.* Развитие понятия вероятности. М.: Наука, 1980.

31. *Baker M., Quera V.* Sequential Analysis and Observational Methods for Social Sciences. NY: Cambridge University Press, 2011.

32. *Abbott A.* Conception of Time and Events in Social Science Methods: Causal and Narrative Approaches // Historical Methods: A Journal of Quantitative and Interdisciplinary History. 1990. Vol. 23 (4). P. 140–150.

33. What Is a Case: Exploring the Foundations of Social Inquiry / Ed. C.C. Ragin, H.S. Becker. NY: Cambridge University Press, 1992.

34. *Robette N., Bry X.* Harpoon or Bait? A Comparison of Various Metrics in Fishing for Sequence Patterns // Bulletin of Sociological Methodology. 2012. Vol 116. P. 5–24.

35. *Elzinga C.H., Liefbroer A.C.* De-Standardization of Family-Life Trajectories of Young Adults: A Cross-National Comparison Using Sequence Analysis // European Journal of Population. 2007. Vol. 23. P. 225–250.

36. *Scherer S.* Early Career Patterns: A Comparison between Great Britain and West Germany // European Sociological Review. 2001. Vol 17. P. 114–119.

37. *Kohler U., Brzinsky-Fay C.* Stata Tip 25: Sequence Index Plots [online] // The Stata Journal. 2005. Vol. 4(5). URL: http://www.stata-journal.com/sjpdf.html?articlenum=gr0022 (date of access: 12.05.2017).

38. *Piccarreta R., Lior O.* Exploring Sequences: A Graphical Tool Based on Multi-dimensional Scaling // Journal of The Royal Statistical Society: Series A. 2010. Vol. 173. P. 165–184.

39. *Gabadinho A., et. al.* Extracting and Rendering Representative Sequences // Knowledge Discovery, Knowledge Engineering and Knowledge Management. 2011. Vol. 128. P. 94–106.

40. Computing and Visualizing Descriptive Statistics [online] // TraMineR: Sequence Analysis in R. Geneva: IDESCO, University of Geneva, Switzerland. URL: http://traminer.unige.ch/preview-describing.shtml (date of access: 12.05.2017).

Alekseyev Timofey
National Research University Higher School of Economics (NRU HSE), Moscow
State Academic University for the Humanities, Moscow, veretur@gmail.com

Sequence analysis in sociology: resources, limitations and possible applications

The article provides a brief review of social sequence analysis as an algorithmic deterministic approach to the classification of event series. The method is discussed in the context of its reception in social sciences in early 1980s with the help of a pioneering research enthusiast A. Abbott. The specificity of sociological applications of sequence analysis under certain data assumptions inherited from bioinformatics, e.g. universal interchangeability of events, arbitrary censoring, rank time variable, is considered. The article classifies a broad set of methods of time–ordered data analysis to provide a base for epistemological confrontment and pinpoint the advantages and shortcomings of sequence analysis compared to nonparametric statistics and general linear models of ordered events. The bases of classification are the dichotomies of time/order event definition and algorithm / statistical inference method of result acquiring. The comparison covers different methods' applications in cases of varied research goals, data types and theoretical assumptions. The article provides a sketch of sequence analysis development over time, considering its aggressive movements towards positioning on the bases of philosophy of history and narrative criticism of general linear models. The roots of its recent orientation towards visualization techniques are discussed as revealed in the scope of early 2000's controversy over the capability of the use of sequence analysis to solve the theoretical problems stemming from the limitations of general linear models.
Key words: sequence analysis, optimal matching, time–ordered data, time–ordered data visualization

References

1. Fasang A. E. Liao T. F. "Visualizing Sequences in the Social Sciences: Relative Frequency Sequence Plots", *Sociological Methods & Research*, 2014, 43(4), 643–676.
2. Abbott A., Tsay A. "Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect", *Sociological Methods & Research*, 2000, 29(1), 3–33.

3. Wu L. L. “Some Comments on ‘Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect”, *Sociological Methods & Research,* 2000, 29 (1), 41–64.
4. Levine J. H. “But What Have You Done for Us Lately? Commentary on Abbott and Tsay”, *Sociological Methods &Research,* 2000, 29(1), 34–40.
5. Sankoff D., Kruskal J. B. (eds.) *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Reading, Mass: Addison-Weasley, 1983.
6. Prokofyev P. A. “Text fragments classification by representation of dependencies using rules interpreted by experts” (in Russian), *Vestnik VGU, Seriya: Sistemnyj analiz i informacionnye tehnologii (Proceedings of Voronezh State University. Series: Systems analysis and information technologies),* 2012, 1, 174–178.
7. Rojtberg M. A. “Bioalgorithmics” (in Russian), *Kompyuterra,* 2001, 36 (413). URL: http://old.computerra.ru/2001/413/197950/ (date of access: 12.05.2017)
8. Abbott A. *Time Matters: On Theory and Method*. The University of Chicago Press, 2001.
9. Durbin R. et al. *Analysis of biological sequences: probabilistic models of proteins and nucleic acids* (in Russian). M.: In-t komp’yuternyh issledovanij, 2006.
10. Gonnet G. H. “Surprising Results on Phylogenetic Tree Building Methods Based on Molecular Sequences”, *BMC Bioinformatics*, 2012, 13.
11. Naumov D. G. “Phylogenetic analysis of the family of homologous proteins” (in Russian), *Zbio*. 2006. URL: http://zbio.net/bio/001/003.html#a6 (date of access: 12.05.2017)
12. Gusfield D. *Algorithms on strings, trees and sequences: computer science and computational biology* (trans., in Russian). SPb: Nevskij Dialekt, 2003. URL: http://padaread.com/?book=10289 (date of access: 12.05.2017)
13. Bininda-Emonds O. R. P. et. al. “The Delayed Rise of Present-Day Mammals”, *Nature*, 2007, 446 (7135), 507–512
14. Cornwell B. *Social Sequence Analysis: Methods and Applications*. NY: Cambridge University Press, 2015.
15. Dodge Y. *The Concise Encyclopedia of Statistics*. NY: Springer-Verlag, 2010.
16. Fisher R. A. “Mathematics of a Lady Tasting Tea”, *The World of Mathematics,* 1956, 3, 1514–1521.
17. Staden R. A “Strategy of DNA Sequencing Employing Computer Programs”, *Nucleic Acids Research,* 1979, 6 (7), 2601–2610.
18. Levenshtejn V. I. “Binary codes with correction of fallouts, inserts and substitutions of symbols”, *Doklady AN SSSR (Reports of the Academy of Sciences of the USSR),* 1965, 163 (4), 845–848.
19. Ezkurdia I., et. al. “Multiple Evidence Strands Suggest that There May Be as Few as 19,000 Human Protein-Coding Genes”, *Human Molecular Genetics*, 2014, 23 (22), 5866–5878.

20. Melnikov B. F., Panin A. G. “The parallel implementation of the multiheuristical approach in the nucleotide sequence comparison problem”, *Vektor nauki TGU (Vector of science of Togliatti State University),* 2012, 4 (22), 83–86.
21. Abbott A., Forrest J. “Optimal Matching Methods for Historical Sequences”, *Journal of Interdisciplinary History*, 1986, 16 (3), 479–494.
22. Elzinga C. H. “Sequence Similarity: A Nonaligning Technique”, *Sociological Methods & Research,* 2003, 32 (1), 3–29.
23. Stovel K., Savage M., Bearman P. “Ascription into Achievement: Models of Career Systems at Lloyds Bank, 1890–1970”, *American Journal of Sociology,* 1996, 102 (2), P. 358–339.
24. Ajvazyan S. A., Buhshtaber V. R., Enyukov I. S., Meshalkin L. D. *Applied statistics: classification and dimension reduction* (in Russian). M.: Finansy i statistika, 1989.
25. Blanchard P., Bühlmann F. (eds.) *Advances in Sequence Analysis: Theory, Method, Applications*. Gauthier. NY: Springer, 2014.
26. Gabadinho A., et. al. “Analyzing and Visualizing State Sequences in R with TraMineR”, *Journal of Statistical Software*, 2011, 40 (4).
27. Abbott A. “Sequence Analysis: New Methods for Old Ideas”, *Annual Review of Sociology*, 1995, 21, 93–113.
28. Cox D.R., Oakes D. *Analysis of Survival Data* (transl., in Russian). M.: Finansy i statistika, 1988.
29. Allison P. D. *Event History Analysis: Regression for Longitudinal Event Data*. Iowa City: Sage Publications, 1984.
30. Majstrov L. E. *The development of the concept of probability* (in Russian). M.: Nauka, 1980.
31. Baker M., Quera V. *Sequential Analysis and Observational Methods for Social Sciences*. NY: Cambridge University Press, 2011.
32. Abbott A. “Conception of Time and Events in Social Science Methods: Causal and Narrative Approaches”, *Historical Methods: A Journal of Quantitative and Interdisciplinary History,* 1990, 23 (4),140–150.
33. Ragin C. C., Becker H. S. *What Is a Case: Exploring the Foundations of Social Inquiry*. NY: Cambridge University Press, 1992.
34. Robette N., Bry X. “Harpoon or Bait? A Comparison of Various Metrics in Fishing for Sequence Patterns”, *Bulletin of Sociological Methodology*, 2012, 116, 5–24.
35. Elzinga C. H., Liefbroer A. C. “De-Standardization of Family-Life Trajectories of Young Adults: A Cross-National Comparison Using Sequence Analysis”, *European Journal of Population*, 2007, 23, 225–250.
36. Scherer S. “Early Career Patterns: A Comparison between Great Britain and West Germany”, *European Sociological Review*, 2001, 17, 114–119.

37. Kohler U., Brzinsky-Fay C. “Stata Tip 25: Sequence Index Plots”, *The Stata Journal*, 2005, 4 (5). URL: http://www.stata-journal.com/sjpdf.html?articlenum=gr0022 (date of access: 12.05.2017)
38. Piccarreta R., Lior O. “Exploring Sequences: A Graphical Tool Based on Multi-dimensional Scaling”, *Journal of The Royal Statistical Society: Series A,* 2010, 173, 165–184.
39. Gabadinho A., et. al. “Extracting and Rendering Representative Sequences”, *Knowledge Discovery, Knowledge Engineering and Knowledge Management,* 2011, 128, 94–106
40. “Computing and Visualizing Descriptive Statistics”, *TraMineR: Sequence Analysis in R. Geneva: IDESCO, University of Geneva, Switzerland*. URL: http://traminer.unige.ch/preview-describing.shtml (date of access: 12.05.2017)