Е.В. Горбунова, В.В. Ульянов
(*Москва*)

# ДИСКРЕТНЫЕ МОДЕЛИ АНАЛИЗА НАСТУПЛЕНИЯ СОБЫТИЙ: РАЗРАБОТКА ПОДХОДОВ К СОВМЕЩЕНИЮ ДАННЫХ, ИМЕЮЩИХ РАЗНУЮ ПЕРИОДИЧНОСТЬ

Проблема совмещения данных с разной периодичностью встречается в различных дисциплинарных областях: астрономии, экономике, медицине, социологии. Настоящая работа посвящена данной проблеме на примере изучения факторов выбытия студентов из американских вузов. В этом исследовании возникла задача совмещения триместровых и семестровых данных, описывающих историю обучения студентов. Предложено три метода решения данной проблемы: агрегирование до года, интерполяция до интервала в полтора месяца, сведение семестровой системы к триместровой с использованием распределений вероятностей наступления событий. Эти подходы носят общий характер и допускают применения в задачах совмещения периодичностей другого типа.

*Ключевые слова*: анализ наступления событий, анализ выживаемости, совмещение разных периодичностей, выбытие студентов, отчисление студентов

## *Введение*

Проблема совмещения данных с разной периодичностью нередко возникает в различных дисциплинарных областях: в

**Елена Васильевна Горбунова** – аналитик Центра внутреннего мониторинга, аналитик Института образования Национального исследовательского университета «Высшая школа экономики». E-mail: evgorbunova@hse.ru.

**Владимир Васильевич Ульянов** – доктор физико-математических наук, профессор факультета социальных наук, факультета компьютерных наук Национального исследовательского университета «Высшая школа экономики». E-mail: vulyanov@hse.ru.

экономических исследованиях, медицине, социологии. С одной стороны, это происходит по причине нерегулярности сбора данных о каком-либо процессе. Например, в исследованиях клинической медицины состояние здоровья пациента может фиксироваться в нерегулярные временные интервалы, и для разных пациентов моменты наблюдения также не совпадают. Естественные катаклизмы, такие как землетрясения, наводнения имеют нерегулярную частоту наблюдения. При проведении лонгитюдных социологических исследований замер данных также может происходить с неодинаковой частотой. Например, в лонгитюдном исследовании «Early Childhood Longitudinal Survey-Kindergarten Cohort», посвященном изучению развития детей, сбор данных происходил в семи волнах исследования с разными интервалами: осенью и весной в детском саду, осенью и весной в первом классе, весной в третьем классе, весной в пятом классе, весной в восьмом классе [12]. Другой пример – лонгитюдное исследование жителей Китая, когда сбор информации о доходах индивидов проходил с неодинаковой частотой (1955, 1960, 1965, 1975, 1978, 1984, 1987, 1991, 1992, 1993, 1994 гг.) [15].

В других случаях измерение проводится с регулярной периодичностью, однако по некоторым периодам данные могут быть пропущены, либо имеют неодинаковую периодичность по разным переменным. В частности, в эконометрических исследованиях, посвященных прогнозированию уровня ВВП, часть данных регистрируются ежемесячно, часть – раз в квартал, некоторые – каждый день. Анализ таких данных требует особого подхода, позволяющего совместить разные периодичности.

Так, в эконометрических исследованиях, использующих анализ временных рядов, существуют различные подходы к совмещению данных с высокой и низкой периодичностью. Чаще всего исследователи прибегают к процедуре агрегирования данных с более высокой периодичностью. Метод агрегирования зависит от специфики данных, однако чаще используется приведение к

среднему значению. Чтобы избежать проблемы потери данных, используют интерполяцию – приведение данных с низкой частотой к высокой частоте. Существуют и более сложные методы для совмещения разных периодичностей, например, модели в пространстве состояний (State-space models) [4], модели временных рядов со смешанными частотами (Mixed data sampling) [6], разночастотная векторная авторегрессия (Mixed frequency vector autoregressive models) [7] и др. Для более подробного ознакомления читатель может обратиться к указанным работам, а также обзорным статьям [5; 14].

При анализе наступления событий (который также обозначается как анализ длительностей, анализ надежности, событийный анализ, анализ выживаемости, далее обозначается как АНС) часто возникает проблема интервального цензурирования, связанная с тем, что время наступления события фиксируется не точно, а лишь в определенные интервалы. Причем длина этих интервалов может быть неодинаковой, интервалы могут пересекаться. Таким образом, переменные, входящие в анализ, могут иметь разную периодичность, что представляет сложность для анализа, поскольку предпосылкой применения АНС выступает единая размерность данных на «входе». В зависимости от специфики данных, существуют различные стратегии для решения этой проблемы.

Проблема интервального цензурирования может решаться с помощью процедуры вменения пропущенных данных (или импутации), когда зависимой переменной в каждый временной интервал приписывается какое-то значение. Чаще всего используется процедура приведения к среднему значению, что позволяет использовать модели для непрерывных данных. Однако использование этого метода имеет свои недостатки, такие как смещение оценок параметров в случае больших по длительности интервалов, недооценка стандартных ошибок [10; 11]. Когда данные по периодам сгруппированы (известны данные по всем наблюдениям в одни и те же моменты времени), проблема интервального цензурирования

может решаться в рамках работы с дискретными данными. В статье [8] сравниваются различные методы для построения модели выживаемости по сгруппированным данным с интервальным цензурированием, а именно, когда регистрация состояния проводилась с нерегулярной периодичностью на протяжении 21 года: раз в год, раз в два года, раз в шесть лет. Сравниваются результаты, полученные с применением импутации данных по пропущенным периодам и без импутации, с применением дискретных моделей.

Другой случай, когда интервальное цензурирование представляет сложности для анализа, связан с тем, что в модель включаются несколько зависящих от времени переменных, и каждая имеет свою периодичность. Например, при анализе надежности работы автомобиля, в качестве зависимого параметра используются два признака с разной периодичностью: возраст и пробег автомобиля. В некоторых исследованиях в модели используются сразу оба признака, это обеспечивается за счет совмещения шкал с разной периодичностью [9].

Еще одним примером может послужить ситуация, когда объекты, входящие в предмет изучения, по одному и тому же признаку могут иметь разную периодичность. Например, в исследовании факторов завершения американскими студентами обучения [3] использовались данные по нескольким вузам, имеющим разную периодичность рубежного контроля. А именно: в части вузов данные регистрировались в триместровой системе (учебный год состоит из трех триместров, каждый приблизительной продолжительностью три месяца), а в остальных вузах – в семестровой системе (учебный год состоит из двух семестров, каждый приблизительной продолжительностью четыре с половиной месяца). Для использования метода анализа наступления события автор объединял эти данные с помощью процедуры агрегирования до года.

Наше исследование посвящено разработке подходов к совмещению данных с разной периодичностью на примере анализа факторов выбытия студентов из вузов. В данном исследовании

потребовалось совместить данные об истории обучения студентов по восьми американским вузам, имеющим триместровую и семестровую периодичность рубежного контроля. Однако использование процедуры агрегирования было нежелательно, поскольку общий период наблюдения оказался невелик и составлял всего два года и один учебный период (осенний триместр/семестр третьего года обучения). Таким образом, возникла необходимость разработки дополнительных подходов к совмещению триместровой и семестровой периодичностей. Кроме того, помимо процедуры агрегирования, здесь рассматривается метод интерполяции, применяемый в эконометрических исследованиях [5; 14]. А также авторами предлагается собственный алгоритм, который ранее не применялся для задачи совмещения разных периодичностей, но позволяет им свести семестровую систему к триместровой с использованием распределений вероятностей наступления событий. Далее подробно описаны особенности реализации третьего подхода, обсуждаются его плюсы и минусы в сравнении с процедурой агрегирования и интерполяции учебных периодов. Несмотря на то что работа фокусируется на обсуждении особенностей совмещения триместровой и семестровой периодичностей, ее результаты носят общий характер и могут применяться для задачи совмещения периодичностей другого типа.

## *Особенности реализации дискретного подхода к анализу наступления событий*

Задача совмещения данных с разной периодичностью возникает в исследованиях, в которых время наступления события фиксируется не точно. В частности, в рассматриваемом исследовании отсутствуют данные о точном времени выбытия студента из вуза. История обучения студентов фиксировалась в семестрах (в пяти

вузах) или триместрах (в трех вузах)[1]. Таким образом, поскольку время выбытия определяется дискретными величинами, для анализа выбраны дискретные модели АНС. Анализ наступления событий наилучшим образом подходит для лонгитюдных данных, содержащих информацию как о факте наступления события, так и о времени его наступления. Мы не будем здесь подробно останавливаться на описании моделей АНС. Заинтересованный читатель может обратиться к отечественным работам [17; 18], либо к иностранным книгам [1; 2; 13].

Будет рассмотрена модель, основанная на логистической регрессии (иначе, модель с логит-связкой):

$$\text{logit } h(t_j) = [a_1 D_1 + a_2 D_2 + ... + a_j D_j] + [\beta_1 X_{1j} + \beta_2 X_{2j} + ... + \beta_p X_{pj}],$$

где $i$ – период наблюдения, $h(t_j)$ – риск наступления события (условная вероятность наступления события в период $j$, при условии, что оно не наступило ранее, и при учете значений постоянных и динамических ковариат $(X_{1j}, ..., X_{pj})$, $D_1, ..., D_j$ – временные периоды, $a_1, ..., a_j$ – значения логарифма риска наступления события в период $i$ при условии, когда все ковариаты принимают значение «нуль», $X_{1j}, ..., X_{pj}$ – совокупность постоянных и динамических ковариат, $\beta_1, ..., \beta_p$ – коэффициенты при соответствующих ковариатах.

## Подходы к объединению данных, имеющих разную периодичность, в АНС

Здесь рассматриваются три метода решения проблемы совмещения данных: агрегирование до года, интерполяция до интервала в полтора месяца, сведение семестровой системы к триместровой с использованием распределений вероятностей наступления событий. Далее будут описаны каждый из подходов.

[1] Временем выбытия определяется последний учебный период, перед которым студент прекратил обучение. Студент считается выбывшим из вуза, если он прервал свое обучение в данном вузе на срок более года (непрерывно).

*Агрегирование периодов*

Данный подход является наиболее простым решением данной методологической задачи. Он заключается в укрупнении временных интервалов, в которых фиксируется наступление события, а также динамические ковариаты. Исследователю необходимо самому определить, каким методом будет проводиться агрегирование данных, а также выбрать размерность новой периодичности. В настоящем исследовании такой подход заключается в укрупнении временных периодов до года (общая периодичность для семестровой и триместровой систем обучения).

*Интерполяция данных по семестрам и триместрам до полуторамесячных периодов*

В противоположность предыдущему подходу, данные за семестры и триместры дезагрегируются до периода, являющегося наибольшим общим делителем. В данном случае, это полуторамесячные интервалы. Таким образом, один период триместра преобразуется в два полуторамесячных периода, а один период семестра – в три периода.

Методы интерполяции могут быть различными:

– исследователь самостоятельно решает, какому из новых периодов приписать наступление события;

– применение алгоритмов, используемых для восстановления пропущенных данных;

– использование алгоритма, аналогичного описанному в третьем подходе ниже.

*Приведение данных более низкой частоты (семестры) к более высоким частотам (триместры), используя распределения вероятностей наступления событий*

Такой подход позволяет для каждого учебного года привести периодичность «два семестра» к периодичности «три триместра»[1]. А именно, каждое событие в семестре считается наступившим в один из двух соответствующих триместров (для осеннего семестра – в осенний или зимний, для весеннего – в зимний или весенний[2]) с вероятностями, которые определяются на основе наблюдаемых данных о доле выбывших по вузам с триместровой и семестровой системами обучения.

Важным условием применения данного подхода является схожесть распределений выбывших студентов в обеих системах периодичности. Графически сравнить распределения долей выбывших студентов между группами вузов с триместровой и семестровой системами обучения позволяют *рис*. 1 и 2. Они иллюстрируют, что в обеих системах периодичности паттерны выбытия схожи: на весенний период приходится пик выбытия студентов; в первый год обучения интенсивность выбытия выше, чем в последующие года. Это дает основание определять вероятности отнесения событий из семестров к одному из двух соответствующих триместров на основе данных о выбытии по вузам с триместровой системой обучения.

Далее описано применение названного алгоритма для данных, которые не содержат динамических ковариат, т.е. значения во

[1] Данный подход позволяет также привести триместровые данные к семестровым, однако он видится менее предпочтительным, поскольку приводит к потере детальной информации по триместрам.

[2] Также во всех вузах есть летний период, когда студент также может изучать курсы, однако большинство из них предпочитают не делать этого. К тому же, продолжительность летнего периода в вузах с семестровой и триместровой системами обучения приблизительно одинакова, поэтому он не рассматривается при совмещении данных.

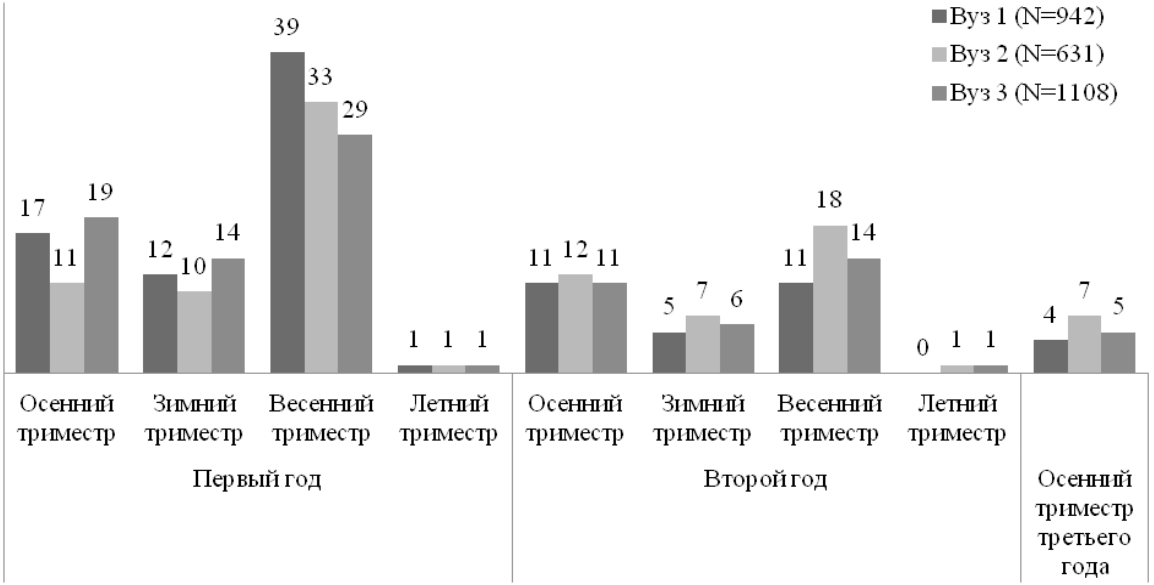


*Рис. 1.* Распределение выбывших студентов в вузах с триместровой системой обучения (доля выбывших в конкретный учебный период по отношению ко всем выбывшим из данного вуза)

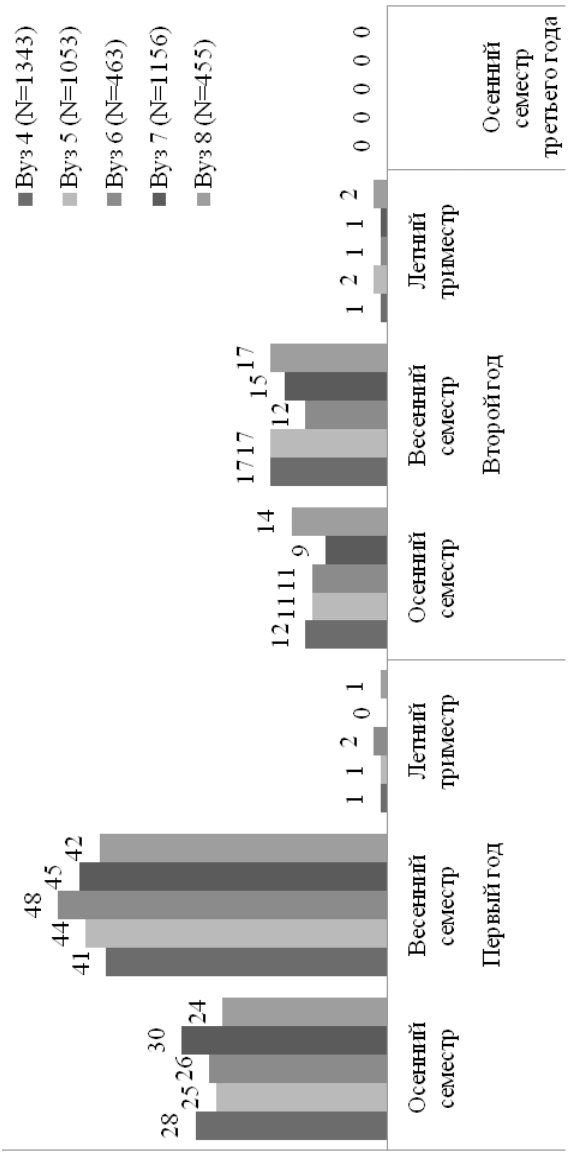


*Рис. 2.* Распределение доли выбывших студентов в вузах с семестровой системой обучения (доля выбывших в конкретный учебный период по отношению ко всем выбывшим из вуза)

времени меняет лишь зависимая переменная[1]. Рассматривается представление данных об истории события в «длинном» формате, т.е. каждая отдельная строка представляет информацию по отдельному индивиду в каждый наблюдаемый период.

Трансформация данных по ковариатам, не изменяющимся во времени, сводится к заполнению пропусков в системе новой периодичности теми же значениями, что и в исходной периодичности. Той же процедуре будут подвергаться значения по зависимой переменной для индивидов, для которых событие не наступило на протяжении всего периода наблюдения.

Для наступивших событий будет определяться, к какому из двух периодов новой размерности их отнести, с помощью описываемого ниже алгоритма. Более подробно данная процедура будет описана на данных о выбытии студентов за первый год обучения.

Введем обозначения: $a_i$ – доля выбывших в $i$-м семестре по всем вузам с семестровой системой обучения, $i = 1, 2$ (осенний и весенний семестры), $b_j$ – доля выбывших в $j$-м триместре по всем вузам с триместровой системой обучения, $j = 1, 2, 3$ (осенний, зимний, весенний триместры), $p_{ij}$ – конструируемая вероятность, с которой событие из семестра $i$ относится к триместру $j$.

Сделаем естественное предположение о невозможности отнесения событий из первого семестра к третьему триместру, а также из второго семестра к первому триместру:

$$p_{12} = p_{21} = 0.$$

Из обозначений вытекают равенства:

$$a_1 + a_2 = b_1 + b_2 + b_3 = p_{11} + p_{12} = p_{22} + p_{23} = 1.$$

Рассмотрим построение $p_{11}$ и $p_{12}$.

Ко второму триместру ($b_2$) необходимо отнести события как из первого ($a_1$), так и второго семестра ($a_2$). Сначала найдем $b_{21}$ и $b_{22}$, удовлетворяющие равенствам:

[1] Применение данного алгоритма с динамическими ковариатами будет обсуждаться дальше.

$$\begin{cases} b_{21} + b_{22} = b_2, & (1) \\ \dfrac{b_1 + b_{21}}{b_{22} + b_3} = \dfrac{a_1}{a_2}. & (2) \end{cases}$$

Содержательно выполнение условия (2) означает, что та доля от $a_1$, которая попадает во второй триместр, в соответствие с конструируемой вероятностью $p_{12}$, при сложении с $b_1$ дает число, относящееся к аналогичной сумме, построенной по $a_2$, как $a_1$ к $a_2$. Естественность такого условия обусловлена схожестью распределений долей выбывших студентов между вузами с семестровой и триместровой системами обучения (что было описано ранее).

Решая систему уравнений (1) и (2) относительно $b_{21}$ и $b_{22}$, получаем:

$$b_{22} = b_1 + b_2 - a_1,$$
$$b_{21} = b_2 + b_3 - a_2.$$

Формулы для расчета итоговых вероятностей определяются следующим образом:

$$p_{11} = \frac{b_1}{b_1 + b_{21}} = \frac{b_1}{a_1},$$

$$p_{12} = 1 - \frac{b_1}{a_1},$$

$$p_{23} = \frac{b_3}{b_3 + b_{22}} = \frac{b_3}{a_2},$$

$$p_{22} = 1 - \frac{b_3}{a_2}.$$

Согласно наблюдаемым данным по вузам, $a_1 = 0{,}386$; $a_2 = 0{,}614$; $b_1 = 0{,}265$; $b_2 = 0{,}197$; $b_3 = 0{,}538$.

Таким образом, для $p_{ij}$ рассчитываются следующие значения:

$$p_{11} = \frac{0{,}256}{0{,}386} = 0{,}687;\ p_{12} = 0{,}313;$$

$$p_{23} = \frac{0{,}538}{0{,}614} = 0{,}876;\ p_{22} = 0{,}124.$$

Отметим, что полученные вероятности используются для каждого из пяти вузов с семестровой системой обучения. После того как вероятности получены, начинается процесс трансформации данных последовательно для каждого периода *i*. Каждое событие, наступившее в период $i = 1$, относится к триместру $j = 1$ с вероятностью $p_{11} = 0{,}687$, либо относится к триместру $j = 2$ с вероятностью $p_{12} = 0{,}313$. Событие, наступившее в период $i = 2$, относится к триместру $j = 2$ с вероятностью $p_{22} = 0{,}124$, либо к триместру $j = 3$ с вероятностью $p_{23} = 0{,}876$.

Практическая реализация этого шага может быть осуществлена в любом статистическом пакете. В массиве данных необходимо создать новую переменную, принимающую случайные значения (например, переменную «randomvar» с помощью функции «gen randomvar = runiform()» в пакете STATA), и задать логическое условие, описывающее отнесение события из семестров к соответствующему триместру, учитывая полученные вероятности $p_{ij}$. Например, для первого семестра логическое условие будет задано следующим образом: если $randomvar \leq 0{,}687$, то событие относится к первому триместру, а если $randomvar > 0{,}687$, то событие относится ко второму триместру.

## *Сравнение подходов к объединению данных с разными периодичностями*

Данный раздел посвящен обсуждению критериев, которые могут применяться при выборе конкретного метода объединения данных. А именно, объем исходных данных, насколько новая периодичность естественна для описания изучаемого процесса,

соответствие новой периодичности изучаемой задаче (описательная или изучение причинно-следственных связей), включение в модель динамических ковариат.

*Объем данных (число периодов наблюдения в исходных данных)*

Когда число периодов наблюдения велико, адекватным методом становится агрегирование данных. Подобная процедура, например, применялась в [3], что было оправдано, поскольку период наблюдения был равен восьми годам. Однако в нашем примере данные ограничены двумя с половиной годами наблюдения, и агрегирование означает уменьшение количества рассматриваемых периодов до двух или трех, что приводит к потере слишком большого объема информации.

*Насколько новая периодичность естественна для модели (модели описания истории наступления событий)*

В силу высокой гибкости американской системы высшего образования студенты могут формировать нелинейные и гибкие образовательные траектории: менять университет или программу обучения (с сохранением накопленных кредитов), делать перерыв в обучении с целью получить опыт работы или по другим обстоятельствам. Как правило, такие изменения, связанные с временным или окончательным прекращением обучения в данном вузе, осуществляются по окончании учебного года. Таким образом, годовая периодичность вполне естественна для описания процесса выбытия студентов из американских вузов, хотя и приводит к потере детальной информации. В случае сведéния семестровой периодичности к триместровой, все данные приводятся к триместровой системе учебного года, которая также является естественной для описания истории обучения. Тогда как полуторамесячная периодичность представляется наименее естественной для описания выбытия студентов.

*Соответствие новой периодичности изучаемой задаче (описать историю наступления событий или установить причинно-следственные связи)*

В случае, когда задачей исследования становится описание функции выживаемости одновременно для вузов с триместровой и семестровой системами, наиболее подходящим методом является третий подход – сведéние семестров к триместрам с использованием распределений вероятностей наступления событий. Он позволяет привести обе системы учебного года к единой шкале с триместрами, таким образом, изменения кривой выживаемости легко интерпретировать (*рис*. 3). Тогда как в случае представления данных в полуторамесячной периодичности интерпретация функции выживаемости менее очевидна – число периодов велико, и исследователю необходимо соотносить, какому семестру/триместру соответствует каждый период новой размерности (*рис*. 4). К тому же по нескольким периодами данные пропущены (2, 6, 8 периоды и т.д.), расстояния между периодами, когда зафиксированы наблюдения, неодинаковы (что отчетливо видно на *рис*. 4: расстояние между первым и третьим периодом такое же, как и между третьим и четвертым периодами). Несмотря на то что данные по вузам с семестровой и триместровой системами объединены на графике, их сложно интерпретировать вместе, поскольку часть периодов принадлежит исключительно семестровой системе (периоды 4, 12), а часть периодов – триместровой системе (периоды 3, 5, 11, 13). Таким образом, резкое падение функции выживаемости от третьего до четвертого периода трудно содержательно описать, поскольку третий период представляет зимний триместр, а четвертый – весенний семестр, они принадлежат к разным системам учебного года и представляют собой совокупность разных университетов. Тогда как в случае процедуры сведéния семестров к триместрам данные по всем вузам представляются в единой триместровой периодичности (см. *рис*. 4).

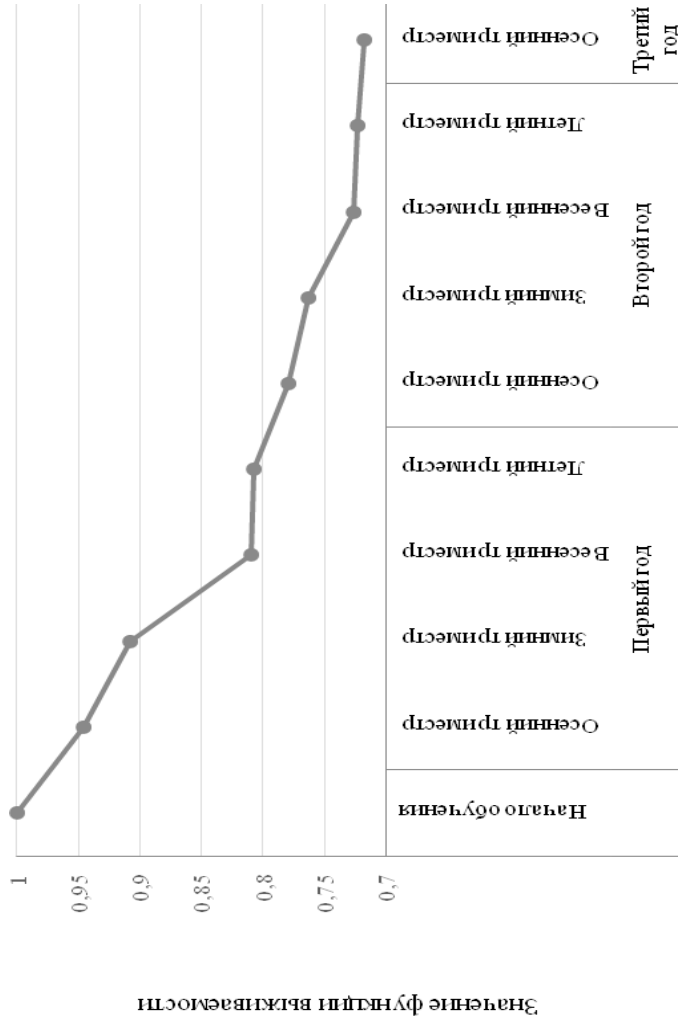


*Рис. 3.* Кривая функции выживаемости после процедуры сведения семестров к триместрам

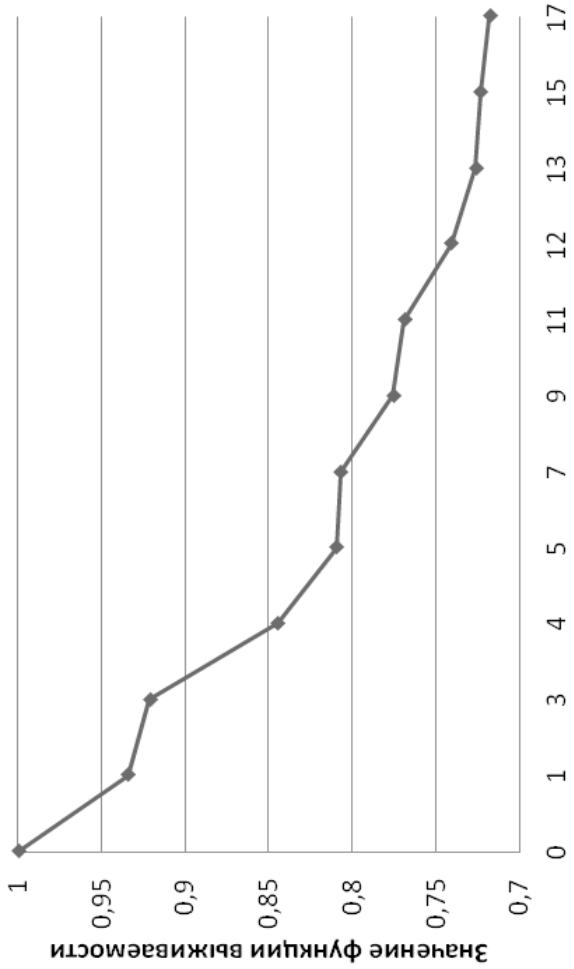


*Рис. 4.* Кривая функции выживаемости после процедуры интерполяции семестров и триместров к полуторамесячным интервалам

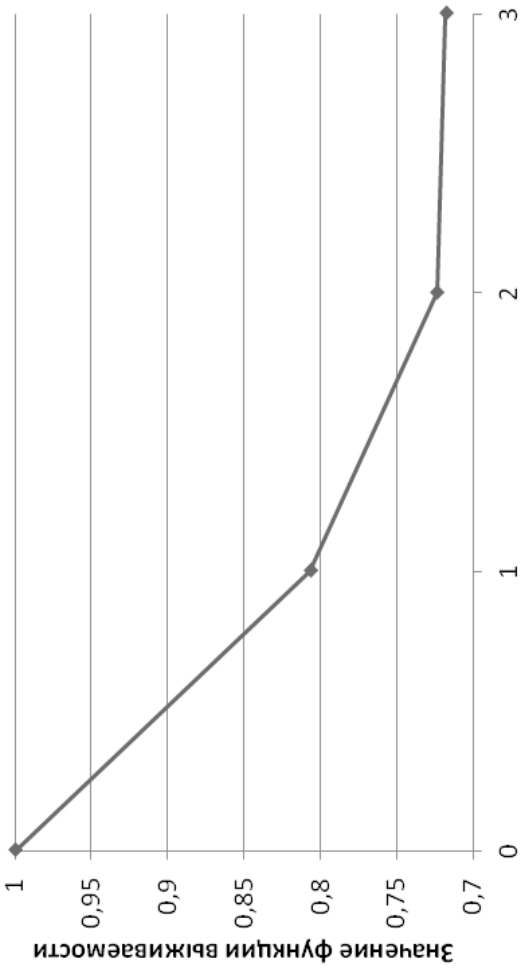


*Рис. 5.* Кривая функции выживаемости после процедуры агрегирования данных до года

Процедура агрегирования данных оказалась наименее подходящей для описания кривой выживаемости, поскольку теряется детальность описания процесса выбытия по семестрам и триместрам. Это хорошо видно на *рис*. 5.

Когда цель анализа – установление причинно-следственных связей и модель не содержит динамические ковариаты, возможно применение всех трех рассматриваемых подходов. Включение в анализ ковариат, изменяющих свои значения во времени (например, успеваемость студента по семестрам/триместрам, получение финансовой помощи для обучения, наличие работы), требует дополнительной трансформации данных при переходе от одной системы периодичности к другой. Данная проблема обсуждается ниже.

## Включение в модель динамических ковариат

При первом подходе, укрупнении данных до годового интервала, значения динамических ковариат необходимо также агрегировать. У данного метода есть ряд ограничений. С одной стороны, исследователю необходимо выбрать методику агрегирования, что становится субъективным решением и может отразиться на оценках коэффициентов. Например, рассмотрим показатель успеваемости студента, являющийся динамической ковариатой. От того, будут ли усреднен данный показатель, взято максимальное значение или значение успеваемости в начале или конце года, зависят результаты оценки данного коэффициента. С другой стороны, при агрегировании данных теряется детальность анализа динамической взаимосвязи независимой и зависимой переменной.

При втором подходе, интерполяции до полуторамесячных интервалов, значения динамических ковариат необходимо также дезагрегировать. С одной стороны, данная процедура представляется достаточно простой: заполнить значения по независимой переменной в новой размерности теми же значениями, что она принимала в исходной периодичности. И это кажется вполне ло-

гичным: если студент получал финансовую помощь в триместре $j = 1$, то он получал финансовую помощь в каждом из двух полуторамесячных интервалов, относящихся к этому триместру. Однако задача дезагрегирования осложняется необходимостью сохранить причинно-следственную связь динамической ковариаты и зависимой переменной. Дальнейший пример описывает эту проблему.

Введем новое обозначение: $k$ – номер полуторамесячного интервала, $k = 1, ..., 8$.

В *табл*. 1 приведен фрагмент исходных данных – запись истории обучения для индивида № 1, который выбыл во втором триместре обучения. Обратим внимание на вариацию признака «получение финансовой помощи»: в первый триместр индивид получал финансовую помощь, тогда как во второй триместр он ее не получал. Анализируя этот фрагмент данных, можно предположить о зависимости переменной «получение финансовой помощи» и выбытия: студент выбыл, когда перестал получать помощь.

*Таблица* 1

ИЛЛЮСТРАЦИЯ ФРАГМЕНТА ИСХОДНЫХ ДАННЫХ

| Номер индивида | Период (триместр) | Наступление события | Получение финансовой помощи |
|---|---|---|---|
| 1 | 1 | 0 | 1 |
| 1 | 2 | 1 | 0 |

Когда триместровая система переводится в полуторамесячную, теоретически возможно отнести событие к любому из двух периодов новой размерности (которые определены как $k = 3$, $k = 4$).

Далее будет описано, какие проблемы возникают при выборе каждого из этих периодов.

В одном случае событие относится к последнему периоду ($k = 4$ в *табл*. 2). Поскольку студент не получал финансовую помощь на протяжении всего второго триместра обучения, содержательно

имеет смысл присвоить обоим периодам ($k = 3$, $k = 4$) значение ковариаты, равное «нулю». Однако при таком решении возникает проблема установления связи – переменная «получение финансовой помощи» будет принимать значение «ноль» и в период $k = 3$, когда событие не наступило, и в период $k = 4$, когда событие наступило. Тогда как в исходных данных (см. *табл.* 1) отсутствие финансовой помощи имело однозначную зависимость с наступлением события. Таким образом, данное решение нарушает структуру взаимосвязи динамической ковариаты и зависимой переменной.

*Таблица* 2

ФРАГМЕНТ ДАННЫХ ПРИ ПРОЦЕДУРЕ ДЕЗАГРЕГИРОВАНИЯ
(приписывание наступления события последнему полуторамесячному периоду новой размерности)

| Номер индивида | Период (полуторамесячные интервалы) | Наступление события | Получение финансовой помощи |
|---|---|---|---|
| 1 | 1 | 0 | 1 |
| 1 | 2 | 0 | 1 |
| 1 | 3 | 0 | ? |
| 1 | 4 | 1 | ? |

В другом случае, когда событие относится к первому полуторамесячному периоду ($k = 3$), проблем с присвоением значений динамической ковариаты не возникает (в период $k = 3$ ковариата принимает то же значение, что и в соответствующем триместре, а по всем последующим полуторамесячным периодам данные цензурированы, т.е. не наблюдаются).

При реализации третьего подхода возникает схожая проблема, что и при процедуре интерполяции данных до полуторамесячных интервалов. Иначе говоря, возможны нарушения структуры взаимосвязи динамической ковариаты и зависимой переменной, когда событие относится к последнему триместру. Вместе с тем

при данном подходе исследователь не может заранее отнести все события к первому из двух триместров, поскольку это противоречит принципам используемого алгоритма – событие из семестра относится к одному из двух триместров с определенными вероятностями.

Таким образом, наиболее удобным и гибким методом, позволяющим включить в анализ динамические ковариаты, оказался подход интерполяции данных в полуторамесячные интервалы, когда исследователь заранее относит событие к первому периоду.

## Заключение

Здесь мы рассматривали проблему совмещения данных с разной периодичностью. На примере данных о выбытии студентов в вузах с разной периодичностью учебного контроля были предложены три метода решения задачи совмещения семестровых и триместровых данных: агрегирование до года, интерполяция до интервала в полтора месяца, сведéние семестровой системы к триместровой с использованием распределений вероятностей наступления событий. Последний метод является авторским и обсуждается впервые.

Особенность предлагаемых методов заключается в том, что они позволяют привести данные к единой периодичности перед оцениванием единой статистической модели по этим данным. Отметим, что в работе [16] предложен еще один подход к решению задачи совмещения триместровых и семестровых данных, основанный на стратификации по типу университета (в первую страту входили университеты с семестровой системой, во вторую – с триместровой). Данный подход позволяет включать динамические ковариаты, сохраняет детальность данных, а также не приводит к созданию фиктивных переменных.

В статье показывается, что выбор конкретного подхода зависит от особенностей данных и изучаемой задачи. В частности,

для изучения факторов выбытия студентов с применением регрессионной модели и включением динамических ковариат наиболее удобен метод интерполяции до интервала в полтора месяца, когда исследователь заранее относит событие к первому периоду. Для описания функции выживаемости наиболее подходит метод сведения семестров к триместрам с использованием распределений вероятностей наступления событий. Процедура агрегирования наименее предпочтительна для данной эмпирической задачи в силу небольшого числа периодов наблюдения.

Как было сказано ранее, проблема совмещения разных периодичностей существует в различных дисциплинарных областях. Рассмотренные в статье методы универсальны и могут быть применимы и к другим изучаемым объектам с различными периодичностями.

## ЛИТЕРАТУРА

1. *Allison P.D*. Discrete-time Methods for the Analysis of Event Histories // Sociological Methodology. 1982. No. 13. P. 61–99.

2. *Box-Steffensmeier J.M., Bradford S.J*. Event History Modeling: A guide for Social Scientists. Cambridge Univ. Press, 2004.

3. *Chiang S.-C*. Applying Event History Analysis to Investigate the Impacts of Developmental Education on Emerging Adults' Degree Completion. Ph.D. dissertation, Ohio State University, 2012.

4. *Evans M.D.D.* Where are we now? Real-time Estimates of the Macroeconomy // International Journal of Central Banking. 2005. Vol. 1(6). P. 127–175.

5. *Foroni C., Marcellino M.G.* A Survey of Econometric Methods for Mixed-Frequency Data. 2013. Norges Bank Research Working Paper 2013-06.

6. *Ghysels E., Santa-Clara P., Valkanov R*. Predicting Volatility: Getting the Most Out of Return Data Sampled at Different Frequencies // Journal of Econometrics. 2006. Vol. 131. No. 1. P. 59–95.

7. *Ghysels E.* Macroeconomics and the Reality of Mixed Frequency Data // Journal of Econometrics. 2016. Vol. 193. No. 2. P. 294–314.

8. *Giolo S.R., Colosimo E.A., Demétrio C.G.B*. Different Approaches for Modeling Grouped Survival Data: A Mango Tree Study // Journal of Agricultural, Biological, and Environmental Statistics. 2009. Vol. 14. No. 2. P. 154.

9. *Jiang R., Jardine A.K.S.* Composite Scale Modeling in the Presence of Censored Data // Reliability Engineering & System Safety. 2006. Vol. 91. No. 7. P. 756–764.

10. *Kim J.S.* Maximum Likelihood Estimation for the Proportional Hazards Model with Partly Interval-censored Data // Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2003. Vol. 65. No. 2. P. 489–502.

11. *Law C.G., Brookmeyer R*. Effects of Mid-point Imputation on the Analysis of Doubly Censored Data // Statistics in Medicine. 1992. Vol. 11. No. 12. P. 1569–1578.

12. *Millimet D.L., McDonough I.K.* Dynamic Panel Data Models With Irregular Spacing: With an Application to Early Childhood Development // Journal of Applied Econometrics. 2017. Vol. 32. No. 4. P. 725–740.

13. *Singer J.D., Willett J.B.* Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. New York: Oxford University Press, 2003.

14. *Wohlrabe K.* Forecasting with Mixed-frequency Time Series Models. Ph.D. dissertation, University Munich, 2009.

15. *Zhou X.* Economic Transformation and Income Inequality in Urban China: Evidence from Panel Data // American Journal of Sociology. 2000. Vol. 105. No. 4. P. 1135–1174.

16. *Горбунова Е.В., Ульянов В.В., Фурманов К.К.* Построение модели выбытия студентов по данным университетов с разной периодичностью рубежного контроля // Прикладная эконометрика. 2017. Т. 45. С. 116–135.

17. *Кокс Д., Оукс Д.* Анализ данных типа времени жизни. М.: Финансы и статистика, 1988.

18. *Ратникова Т.А., Фурманов К.К.* Анализ панельных данных и данных о длительности состояний. М.: НИУ ВШЭ, 2014.

***Gorbunova Elena***
National Research University Higher School of Economics (NRU HSE), Moscow, evgorbunova@hse.ru

***Ulyanov Vladimir***
National Research University Higher School of Economics (NRU HSE), Moscow, vulyanov@hse.ru

**Discrete-time methods of event history analysis: developing approaches to combining mixed-frequency data**

The problem of combining mixed-frequency data is found in various disciplinary areas: astronomy, economics, medicine, sociology. This article is devoted to this problem on the example of studying the factors of student expulsion from American universities. In this study the task was to combine trimester and semester data describing the student education trajectory. Three methods of solving this problem were proposed: aggregation up to a year, interpolation to an interval of one and a half months, reduction of the semester system to a trimester system using the probability distribution of occurrence of events. These approaches are of a general nature and allow applications to the tasks of combining other data types.
*Key words:* event history analysis, survival data analysis, combining mixed-frequency data, student expulsion

**References**

1. Allison P.D. “Discrete-time methods for the analysis of event histories”, *Sociological Methodology*, 1982, 13, 61-99.
2. Box-Steffensmeier J.M., Bradford S.J. *Event history modeling: A guide for social scientists.* Cambridge Univ. Press, 2004.
3. Chiang S.-C. *Applying Event History Analysis to Investigate the Impacts of Developmental Education on Emerging Adults’ Degree Completion*. Ph.D. dissertation, Ohio State University, 2012.
4. Evans M.D.D. “Where are we now? Real-time estimates of the macroeconomy”, *International Journal of Central Banking*, 2005, 1(6), 127–175.
5. Foroni C., Marcellino M.G. *A Survey of Econometric Methods for Mixed-Frequency Data. 2013*. Norges Bank Research Working Paper 2013-06.
6. Ghysels E., Santa-Clara P., Valkanov R. “Predicting volatility: getting the most out of return data sampled at different frequencies”, *Journal of Econometrics,* 2006, 131 (1), 59–95.

7. Ghysels E. “Macroeconomics and the reality of mixed frequency data”, *Journal of Econometrics,* 2016, 193 (2), 294–314.
8. Giolo S.R., Colosimo E.A., Demétrio C.G.B. “Different approaches for modeling grouped survival data: A mango tree study”, *Journal of agricultural, biological, and environmental statistics,* 2009, 14 (2), 154.
9. Jiang R., Jardine A. K. S. “Composite scale modeling in the presence of censored data”, *Reliability Engineering & System Safety,* 2006, 91 (7), 756–764.
10. Kim J.S. “Maximum likelihood estimation for the proportional hazards model with partly interval-censored data”, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2003, 65 (2), 489–502.
11. Law C.G., Brookmeyer R. “Effects of mid-point imputation on the analysis of doubly censored data”, *Statistics in medicine*, 1992, 11 (12), 1569–1578.
12. Millimet D.L., McDonough I.K. “Dynamic Panel Data Models With Irregular Spacing: With an Application to Early Childhood Development”, *Journal of Applied Econometrics*, 2017, 32 (4), 725–740.
13. Singer J.D., Willett J.B. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press, 2003.
14. Wohlrabe K. *Forecasting with Mixed-frequency Time Series Models*. Ph.D. dissertation, University Munich, 2009.
15. Zhou X. “Economic transformation and income inequality in urban China: evidence from panel data”, *American Journal of Sociology,* 2000, 105 (4), 1135–1174.
16. Gorbunova E.V., Ulyanov V.V., Furmanov K.K. “Using data from universities with different structure of academic year to model student attrition” (in Russian), *Prikladnaya ekonometrika (Applied Econometrics)*, 2017, 45, 116–135.
17. Cox D.R., Oakes D. *Analysis of Survival Data* (transl., in Russian). M.: Finansy i statistika, 1988.
18. Ratnikova T.A., Furmanov K.K. *Analysis of panel data and data on the duration of states* (in Russian). M.: HSE, 2014.