ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 050: [002:001.8]

А.Н. Либкинд, В.А. Маркусова, И.А. Либкинд

К вопросу определения динамики показателей периода полужизни журналов по *Journal Citation Reports**

Обсуждаются проблемы, возникающие при определении значений показателей периода полужизни журнальных статей по ежегодным выпускам Journal Citation Reports за период 1997-2018 гг. В выпусках до 2017 г. точное значение этих показателей не указывалось, когда оно превышало 10 лет. Доля журналов с такими показателями нередко достигает нескольких десятков процентов. Предлагаются приемы, позволяющие обойти это ограничение и достоверно оценивать динамику периода полужизни статей по категориям Web of Science, что важно при неправомерном значении, придаваемом библиометрическим показателям.

Ключевые слова: *старение литературы, период полужизни, Cited Half-life, Citing Half-life, Journal Citation Reports*

DOI: 10.36535/0548-0027-2020-05-4

ВВЕДЕНИЕ

Возможность более точно измерять многие закономерности, присущие журнальным публикациям, в наукометрии появилась после создания Ю. Гарфилдом в 1964 г. «Указателя библиографических ссылок в естественных науках». Однако многие из этих закономерностей начали изучаться значительно раньше. Обсуждая способы оценки старения литературы в определенных областях знания, Дж. Бернал ещё в 1958 г. предложил для этого термин «период полужизни» журнальных статей по аналогии с периодом полураспада радиоактивных веществ. Его идею использовали американские ученые Р. Бартон и Р. Кеблер [1, с. 20]. Они разработали в 1960 г. способ вычисления этого показателя, который измеряли интервалом времени, в течение которого была опубликована половина всей используемой в настоящее время литературы по какой-либо отрасли или предмету. На современном языке это означает, что они имели в виду время опубликования половины статей, цитируемых в журналах данного года. По мере совершенствования библиометрического инструментария стал приниматься во внимание и возраст библиографических ссылок в самих журналах этого года, т.е. цитирующих статей.

Со временем выяснилось, что период полужизни отражает не только старение научной литературы, но и её рост, о чём выразительно написал Д. Прайс: «В течение нескольких лет после публикации спрашиваемость статьи или её относительная цитируемость уменьшается крайне медленно (по параболе, если считать по логарифмам прошедших лет). Даже через столетие возможность цитирования уменьшается только на порядок. Большинство ссылок падает на работы последних лет потому, что этих работ большинство, и очень сомнительно, чтобы это вызывалось эффектом немедленности, связанным с быстрым старением» [2, с. 292]. Эту проблему старения подробно изучали ученые-информатики: английский – М. Лайн [3] и российский – В.М. Мотылев [4]. М. Лайн писал: «Период полужизни литературы вынужден быть тем короче, чем быстрее она растёт, если среднее число цитирований на одну статью за это время не уменьшается. Если каждая статья имеет одинаковую вероятность быть использованной или процитированной, более новая литература используется чаще просто потому, что ее больше».

Для теоретиков информатики и историков науки важно учитывать старение литературы в чистом виде, а для информационных работников и библиотекарей период полужизни является важным практическим показателем и продолжает широко использоваться. В наше время, когда библиометрические показатели служат (хотя и не всегда оправдано) оценкой научной деятельности ученых, организаций и даже стран,

* Работа выполнены при поддержке Российского фонда фундаментальных исследований (проекты РФФИ 20-07-00014 и 20-010-00179).

важно понимать, как вычисляется и что реально значит такой показатель, как период полужизни журнальных статей [5, 6].

В настоящем исследовании в качестве исходных данных нами были использованы ежегодные выпуски «*Journal Citation Reports – Science Edition» (JCR-SE)* и «*Journal Citation Reports – Social Science Edition» (JCR-SSE)* за период 1997–2018 гг. Этот аналитический инструмент располагается на платформе *Web of Science (WoS)* компании *Clarivate Analytics*. Среди ряда показателей, которые приводятся в каждом ежегодном выпуске *JCR* для каждого включенного в этот выпуск журнала, нас, прежде всего, интересовали показатели, известные как показатели «периода полужизни» – *Cited Half-life* и *Citing Half-life*, и кроме того принадлежность журнала к той или иной тематической категории *WoS (subject categories WoS или WoS categories)*, а также соответствующие этим категориям показатели «периода полужизни».

Напомним основные определения, связанные с понятием «период полужизни», воспользовавшись определениями, содержащимися в разделе «*Help*» информационной платформы *WoS.* (https://help.incites.clarivate.com/ incitesLiveJCR/ overviewGroup/ overviewJCR. html)[1]:

- цитируемый период полужизни некоторого журнала (*Cited Half-life – CdHL*) представляет собой медианный возраст (в годах) массива тех статей из этого журнала, которые в заданном году выпуска «*Journal Citation Report» (JCR)* были процитированы. Это значит, что половина процитированных статей из указанного журнала были опубликованы ранее цитируемого периода полужизни этого журнала;
- цитирующий период полужизни журнала (*Citing Half-life – CgHL*) некоторого журнала представляет собой медианный возраст (в годах) массива тех статей, которые цитировал этот журнал в данном году выпуска *JCR*. Это значит, что половина статей, которые цитировал данный журнал, были опубликованы ранее периода полужизни;
- агрегированный цитируемый период полужизни (*Aggregate Cited Half-life – AgrCdHL*), характеризующий некоторую тематическую категорию *Web of Science (WoS category)*, представляет собой медианный возраст в годах публикаций в журналах, соответствующих этой категории (значения этого показателя получаются путем агрегирования соответствующих журнальных данных);
- агрегированный цитирующий период полужизни (*Aggregate Citing Half-life – AgrCgHL*), характеризующий некоторую тематическую категорию *WoS*, представляет собой медианный возраст статей, которые в год выпуска *JCR* цитировали журналы, соответствующие данной категории (значения этого показателя получаются путем агрегирования соответствующих журнальных данных).

ПРОБЛЕМЫ ПРИ ОПРЕДЕЛЕНИИ ПЕРИОДА ПОЛУЖИЗНИ ЖУРНАЛОВ

Приступая к задаче определения динамики показателей *Cited Half-life* (*CdHL*) и *Citing Half-life* (*CgHL*), мы были вынуждены учитывать ряд проблем. Во-первых, единственным источником статистических данных для решения задачи определения динамики показателей периода полужизни журналов может послужить информация, содержащаяся в ежегодных выпусках информационной системы *Journal Citation Reports* (*JCR*). Во-вторых, выполненный нами предварительный анализ *JCR* за период 1997–2018 гг. показал, что значения показателей *CdHL* и *CgHL* для каждого журнала, включенного в соответствующий ежегодный выпуск *JCR*, обычно приводятся в числовой форме с точностью до 0,1 года. При этом в ежегодных выпусках *JCR* за 2017 г. и 2018 г., данные о значениях *CdHL*/*CgHL* для каждого из журналов приводятся полностью, т.е. в числовой форме. К сожалению, во всех остальных ежегодных выпусках *JCR* (за период 1997–2016 гг.) картина иная. А именно, если значение соответствующего показателя для некоторого журнала превышает 10 лет, то для этого журнала вместо точного числового значения указывается только текстовое выражение вида «>10.0». Попытки определения динамики значений рассматриваемых показателей при использовании только тех совокупностей журналов, для которых значения показателей приводятся в числовой форме, приведут, в лучшем случае, к существенным искажениям. Действительно, использовать «напрямую» текстовые данные вида «>10.0» при соответствующих вычислениях не удастся и, следовательно, этими данными придется просто пренебречь. И это притом, что доля журналов, у которых значения *CdHL* больше 10 лет в выпусках *JCR-SE*, значительна и находится в пределах 15,7–21,5%, и со временем она возрастает. Так, нижний предел этой доли (15,7%) соответствует выпуску 1997 г., а верхний (21,5%) – выпуску 2016 г. В случае *CgHL* в тех же выпусках *JCR-SE* доля таких журналов еще выше и находится в пределах 29–31%. Доля журналов, для которых приводятся такие неточные данные, в случае *JCR-SSE* для *CdHL* находится в пределах 11,9–31%, а для *CgHL* – в пределах 25–50,9% (1997 г. и 2016 г. соответственно). Таким образом, приходится констатировать, что подавляющее большинство (20 из 22) выпусков *JCR* содержат неполные данные о значениях *CdHL* и *CgHL* конкретных журналов.

К сожалению, ситуация с такими неполными данными для тематических категорий *WoS* еще сложнее. Действительно, для всех без исключения ежегодных выпусков *JCR* и всех категорий *WoS* за весь период 1997–2018 гг. при значениях показателей, превы-

[1] Cited Half-life some journal is median age of the articles from this journal that were cited in the JCR year. It means that half of a journal's cited articles were published more recently than the cited half-life.

Citing Half-life is median age of articles cited by the journal in the JCR year. It means that the articles cited by this journal were published more recently than the citing half-life.

Aggregate Cited half-life is the median age, in years, of items in any journal in the category that were cited during the JCR year. Shows the distribution by cited year of citations to articles published in journals in the category in the JCR year.

Aggregate Citing half-life is the median age of articles cited by journals in the category in the JCR Year based on aggregated journal data. Shows the distribution by cited year of citations from journals in the category made in the JCR year.

шающих 10 лет, в *JCR* приводятся не числовые, а текстовые значения вида «>10.0», что не позволяет использовать эти значения для дальнейших вычислений[2].

Отметим, что если игнорировать журналы, для которых в *JCR* в качестве значений *CdHL* (*CgHL*) указаны текстовые выражения «>10», то с каждым годом в расчетные значения этих показателей будут вноситься всё бо́льшие искажения. В итоге, при положительной динамике показателей результаты расчетов могут дать динамику отрицательную, что совершенно недопустимо.

Если рассмотреть эту ситуацию с методической точки зрения, то здесь необходимо указать следующее. Казалось бы, исходя из чисто статистических соображений, даже с указанными выше потерями можно было бы пренебречь, если бы не одно существенное обстоятельство. А именно, при попытке ограничиться только выборкой журналов, которые имеют числовые значения показателей, мы тем самым отбрасываем именно те журналы, которые характеризуются самыми большими значениями показателей. Таким образом, здесь мы имеем дело не с обычной случайной выборкой из генеральной совокупности и, следовательно, чисто статистический подход здесь недопустим.

Из всего изложенного следует, что прежде чем приступить к решению проблемы определения динамики показателей *CdHL* и *CgHL*, как для журналов, так и для категорий *WoS*, необходимо решить задачу оценки средних (точнее, средневзвешенных) значений этих показателей для каждого года из рассматриваемого периода**.** То есть**,** нам необходимо каким-либо образом восполнить недостающие данные в годовых распределениях журналов по этим показателям за 1997–2016 гг. Именно решению этой задачи и, в конечном счете, разработке методов для определения динамики значений показателей периода полужизни, характеризующих соответствующие совокупности журналов, посвящена настоящая статья. Прежде чем перейти к непосредственному рассмотрению этой задачи, а также соответствующих совокупностей журналов и их значений по показателям *CdHL/CgHL*, введем необходимые понятия и определения.

Определение 1. Будем называть распределением журналов по значениям показателя *CdHL* таблицу, в первой графе (колонке) которой последовательно в порядке возрастания приводятся значения *CdHL*, а во второй графе против каждого значения *CdHL* из первой колонки указано число журналов, каждый из которых имеет именно это значение *CdHL*. В дальнейшем для краткости вместо выражения «распределение журналов по значениям показателя» будем применять – «распределение».

Характер рассматриваемых нами совокупностей журналов и соответствующих им показателей позволяет классифицировать эти совокупности по следующим основаниям (признакам):

1) принадлежность журнала к заданному (i-му) ежегодному выпуску (году опубликования) *JCR-SE/ JCR-SSE*;

2) принадлежность журнала к данной тематической категории *Web of Science* (*subject category WoS – WoS Category*) и/или к заданному набору *WoS Categories*;

3) принадлежность журнала той или иной стране (страна издания журнала);

4) факт присутствия журнала в i-м выпуске *JCR-SE/JCR-SSE* и в следующем $i+1$ выпуске. Такие журналы будем называть *«Относительно сохранившимися журналами»* или *«Относительно постоянными журналами»* – "*Relatively regular journals (RR Journals)"* or *"Relatively preserved journals" (RP Journals)* or *"Relatively retentive journals (RR Journals)"*;

5) факт присутствия журнала во всех без исключения ежегодных выпусках *JCR-SE/ JCR-SSE* за заданный период (в нашем случае за 1997-2018 гг.). Эти журналы назовем *«Абсолютно сохранившимися журналами»* или *«Всегда присутствующими журналами»* – "*Absolutely preserved journals* or *always present journals (AP Journals)"* or "*Absolutely retentive journals (AR Journals)"* [7]. В качестве исходной точки на шкале времени в этом случае будет принят 1997 г.;

6) факт присутствия журнала в i+1 выпуске при обязательном отсутствии этого журнала в i-м выпуске. Такие журналы будем называть относительно новыми – *relatively new journals (RN Journals)*;

7) факт присутствия журнала в i-м выпуске *JCR-SE/JCR-SSE* при обязательном его отсутствии во всех предшествующих выпусках. Такие журналы назовем абсолютно новыми – *absolutely new journals (AN Journals).*

В настоящей работе будут рассмотрены те совокупности журналов и категорий *WoS*, которые могут быть сформированы с помощью признаков, указанных в пунктах (1)–(3) и (5). Для описания и сопоставления различных распределений журналов по *CdHL* и *CgHL* воспользуемся понятиями, применяемыми для аналогичных целей в статистике: среднее значение распределения (*mv – mean value*); мода распределения (*mode – Md*); медиана распределения (*median – Mn*); полная ширина на уровне половины максимального значения (*FWHM – full width at half maximum*); асимметрия распределения[3].

[2] Число и доля таких категорий со временем возрастает. Так, если в 2003 г. в выпуске в *JCR-SE* доли категорий со значением *Aggregate Cited Half-life (AgrCdHL)* и *Aggregate Citing Half-life (AgrCgHL),* составляли 4,7% и 11,2 (%), то в 2018 г. эти показатели достигли 13,4% и 16,3% соответственно. Аналогичная, но еще более ярко выраженная картина наблюдается и для выпуска *JCR-SSE*: если в 2003г. доля таких категорий для *AgrCdHL* и для *AgrCgHL* составляла 14,8% и 16,7% соответственно, то в 2018 г. эти значения достигли 43,1% и 41,3%.

[3] *Среднее значение i-го распределения* (*mv – mean value*) – сумма значений всех элементов *i*-го распределения, деленная на число этих элементов – является аналогом математического ожидания в теории вероятностей. В нашем случае элементы – это значения *CdHL/CgHL* (годы), а значения элементов – это число журналов, соответствующих конкретному году *CdHL /CgHL*. *Мода распределения (mode – Md)* – значение, которое в *i*-м распределении встречается наиболее часто. В нашем случае мода – это то значение *CdHL/CgHL* (в годах), которому соответствует наибольшее число журналов в этом распределении. *Медиана распределения (median - Mn)* – такое значение в рас-

Замечание. В дальнейшем в целях упрощения изложения в тех случаях, когда это не будет вызывать путаницы, мы будем упоминать только показатель *CdHL*, полагая при этом, что все соображения, определения и вычисления, которые приводятся далее в отношении показателя *CdHL*, в полной мере касаются и показателя *CgHL*.

Определение 2. Те журналы, каждый из которых в данном выпуске *JCR* охарактеризован значением *CdHL*, не превышающем 10 лет, будем называть *основными журналами распределения*, а ту часть распределения, которая соответствуют этим журналам, – *«Основной частью распределения»*.

Определение 3. Журналы, каждый из которых в выпуске *JCR* охарактеризован значением *CdHL*, превышающем 10 лет, будем называть *журналами хвоста распределения*, а ту часть распределения, которая соответствует этим журналам, – *«Хвостом распределения журналов» или просто «Хвостом распределения»*. Это определение в значительной степени совпадает с понятием «хвост распределения», которое обычно применяется в статистике в теории распределений. При этом предложенное здесь *ad hoc* определение хвоста распределения журналов, несмотря на частный характер, имеет свои достоинства – с его помощью мы всегда однозначно можем указать на начало хвоста распределения.

Определение 4. Распределение, у которого каждому журналу хвоста распределения в *JCR* для показателя *CdHL* приводится числовое значение, будем называть полным распределением, и, соответственно, распределение, у которого журналам хвоста в *JCR* приводятся текстовые выражения вида «>10.0», – усеченным (неполным).

Важное замечание. Все предлагаемые далее приемы и методы ориентированы, прежде всего, на превращение усеченных распределений в полные. Следующий шаг – вычисление средневзвешенных значений *CdHL* для такого превращенного распределения. При этом мы не ставим перед собой задачу определения действительных числовых значений *CdHL* для тех журналов, у которых в *JCR* в качестве значения показателя *CdHL* указано текстовое выражение «>10.0». Еще раз подчеркнем: в данном случае речь идет не о журналах как таковых, а о совокупностях журналов и соответствующих им распределениях.

ВОЗМОЖНЫЕ ПРИЕМЫ ОЦЕНКИ ДИНАМИКИ ПЕРИОДА ПОЛУЖИЗНИ

Вернемся к особенностям исходных данных. Как уже отмечалось, при построении распределений, а также при вычислении средних/средневзвешенных значений *CdHL* и *CgHL*, неполные, по сути, текстовые выражения вида «>10.0» не удается корректно использовать для численной обработки, по крайней мере, без применения некоторых искусственных приемов.

Прием 1(способ). Каждому журналу и категории *WoS*, у которых для *CdHL*/*CgHL* в *JCR* вместо числовых значений указано «>10.0», припишем значения 10,1, что позволит при соответствующих вычислениях учитывать также те журналы, у которых рассматриваемые показатели имеют значения, превышающие 10 лет. Насколько такой прием смещает (естественно, в сторону уменьшения) вычисляемые значения показателей, можно оценить сравнив вычисленные двумя различающимися способами средневзвешенные значения[4] *CdHL* для одного и того же набора журналов, который, к тому же, соответствует одному и тому ежегодному выпуску *JCR*. В качестве такого набора используем выпуск *JCR* за 2018 г., где каждому журналу, в том числе и журналу с *CdHL*, превышающему 10 лет, приписано числовое значение.

При вычислении средневзвешенного значения *CdHL* этим способом каждому журналу, у которого *CdHL* превышает 10 лет, вместо реальных числовых значений припишем значение 10,1. При другом способе будем использовать реальные числовые значения *CdHL*, в том числе и для тех журналов, у которых *CdHL* превышает 10 лет[5]. Понятно, что чем больше доля журналов, для которых *JCR* приводит текстовые значения вида «>10.0», тем менее точными оказываются получаемые средневзвешенные значения *CdHL*/*CgHL*. Однако при отсутствии точных исходных данных и учитывая, что этот прием применяется к каждому ежегодному выпуску *JCR*, все же можно будет с определенной точностью определить тенденции изменения значений этих показателей. Действительно, в случае применения к каждому годовому выпуску *JCR* этого приема мы при вычислении вносим если не неизменную от выпуска к выпуску, то близкую по величине систематическую ошибку. При сопоставлении показателей в динамике на эту ошибку можно делать поправку, а в некоторых случаях её можно просто игнорировать. Остается добавить, что вычисленное значение систематической ошибки (0,91)[6] в случае выпусков *JCR-SE* является, скорее всего, её верхней оценкой. Дело в том, что именно выпуск 2018 г., для которого была осуществлена эта

пределении, что ровно половина из значений l_j в распределении больше или равна этому значению ($l_j \le Md$), а другая половина меньше или равна этому значению. *Полная ширина на уровне половины максимального значения* (*FWHM – full width at half maximum*) – разность между правой и левой координатами на оси абсцисс, при условии, что на оси ординат эти координаты соответствуют половине максимального значения в распределении. *Асимметрия распределения* – характеризует степень его отклонения от распределения симметричного. Если правый хвост распределения длиннее левого то говорят, что распределение характеризуется положительной (правой) асимметрией (*right asymmetry/positive asymmetry*), если левый хвост длиннее правого, то будем говорить, что распределение характеризуется отрицательной (левой) асимметрией *(left asymmetry/negative asymmetry*).

[4] Средневзвешенное значение *CdHL* вычислялось путем деления {суммы [произведений (числа журналов в данной группе *CdHL*) на (значение *CdHL* этой группы)]} на {общее число журналов в распределении}.

[5] В первом случае средневзвешенное значение CdHL = 8,94, во втором – этот показатель оказывается равным 9,85. Указанный прием приводит к тому, что вычисленное с его помощью средневзвешенное значение получилось заметно ниже того, которое вычислено с использованием реальных значений: 9,85–8,94=0,91.

[6] См. сноску 5.

оценка, характеризуется максимальным значением доли журналов, у каждого из которых *CdHL* больше 10 лет. Аналогичная ситуация характерна и для средневзвешенных значений *CgHL*.

К сожалению, при построении распределений журналов по показателям *CdHL/CgHL*, применение описанного приема не решает проблему с неточными данными. Действительно, непосредственное включение в распределение журналов, для каждого из которых указано «>10.0», оказывается невозможным. Такие журналы из соответствующих распределений приходится исключать. В противном случае на графике после значения «10» возникнет почти вертикальный отрезок с длиной, равной доле этих журналов. Таким образом, предложный *Прием 1* при построении распределений оказывается недостаточным, а также не отличается и особой точностью при вычислении средневзвешенных значений соответствующих показателей. Очевидно, что для более полноценного решения этих задач необходимо разработать более эффективный способ.

Прием 2 (способ). Попытаемся «преобразовать» усеченное распределение в полное. Усеченным будем называть распределение, из которого исключены журналы со значениями *CdHL/CgHL*, превышающими 10 лет, а полным распределение – из которого такие журналы не исключены, при условии, что каждому журналу сопоставлено числовое, а не текстовое значение указанных показателей. Для предполагаемого преобразования некоторого усеченного распределения в полное попытаемся использовать данные другого (реперного) распределения при условии, что оно само является полным. Для реализации этого приема попытаемся сформулировать рабочую гипотезу, предварительно введя некоторые необходимые определения и выполнив необходимый анализ ряда распределений журналов по значениям показателей *CdHL* и *CgHL*.

Определение 5. Группой журналов в данном распределении будем называть такую совокупность журналов, которые имеют одинаковые (совпадающие, равные) значения данного показателя, в частности, – равные значения *CdHL* или *CgHL*. Для краткости группу журналов просто группой. Например, группой являются все журналы, значения *CdHL* каждого из которых составляют 7,5 лет.

На основе этого определения и используя данные ежегодных выпусков *JCR*, по значениям *CdHL* и *CgHL* были построены соответствующие графики распределений журналов по этим показателям: на оси абсцисс отложены группы журналов (значения *CdHL* или *CgHL*), а по оси ординат – число (доля) тех журналов, каждый из которых характеризуется значением *CdHL/CgHL*, соответствующим этой группе.

Определение 6. Однотипными (распределениями, принадлежащими определенному классу распределений) будем называть распределения, соответствующие одному и тому же показателю. Так, однотипными распределениями являются все распределения журналов по значениям показателя *CdHL* – их отнесем к одному классу, а другой класс однотипных распределений представляют распределения журналов по значениям показателя *CgHL*.

Особенности распределений *CdHL* и *CgHL* и их динамику иллюстрируют графики на рис. 1 и 2. Для того чтобы на одном рисунке можно было совместить и сопоставить несколько различных графиков распределений, шкала значений по оси ординат приводится в долях числа журналов от их общего числа. На рис. 1 представлены распределения, соответствующие и *CdHL,* и *CgHL*, т.е. распределения, принадлежащие двум различным классам. Рассмотрим распределения по *CdHL*. Несмотря на то, что эти распределения соответствуют разным тематическим выпускам *JCR* (*JCR-SE и JCR-SSE*) и различным ежегодным (2017 г. и 2018 г.) выпускам этого ресурса, по своей форме они близки. Ещё более ярко выражено сходство для распределений по *CgHL*, которые хотя и относятся к одному и тому же выпуску *JCR-SE* (2018 г.), однако соответствуют разным наборам журналов: одно из этих распределений построено с учетом всех журналов, а другое – с учетом только постоянно сохраняющихся журналов, т.е. тех, которые присутствовали в каждом выпуске *JCR-SE* в течение всего периода наблюдений (1997–2018 гг.).

Чтобы не затемнять изображение, на рис. 2 приведены лишь линии трендов, а не конкретные значения каждого из распределений по *CdHL/CgHL*. Судя по значениям R^2 (коэффициент детерминации – *determination coefficient*) линии трендов достаточно хорошо аппроксимируют реальные распределения (значения R^2 близки к 1). Из рис. 2 видно, что значения *Cited Half-life* и *Citing Half-life* со временем увеличиваются. В частности, заметно увеличение значений моды соответствующих распределений. Так, для случая распределений по *CdHL*: 2003 г. мода равна 5,8 лет; 2010 г. – 6,1 лет; 2017 г. – 6,4 года. Соответственно, для случая распределений по *CgHL* мода составляет: 2003 г. – 7,7 лет; 2017 г. – 8,3 года. Кроме того, наблюдается увеличение доли журналов (приведены вверху рис. 2 после значений R^2), у которых значения соответствующих полупериодов жизни превышает 10 лет. Для случая *CdHL* этот показатель составлял: в 2003 г. – 16,1%; в 2010 г. – 17,0%; в 2017 г. – 22,6%. В случае *CgHL* – в 2003 г. – 31,2%; в 2017 г. – 32,4%[7].

Суммируя результаты анализа, и сопоставляя графики, часть из которых представлена на рис. 1 и 2, можно заключить:

- формы графиков распределений по *CdHL* близки друг к другу;
- формы графиков распределений по *CgHL* близки друг к другу
- распределения журналов по *CdHL* и по *CgHL* характеризуются положительной (правой) асимметрией (*right asymmetry*);

[7] При рассмотрении графиков, изображенных на рис. 2, может сложиться ошибочное впечатление, что на долю журналов, у которых значения соответствующих полупериодов жизни превышает 10 лет, приходится всего несколько процентов. Это связано с тем, что на оси ординат отложены значения долей (%), рассчитанные исходя из общего числа только тех журналов, у которых значения соответствующих показателей не превышают 10 лет. Это значит, что журналы, у которых значения полупериода жизни превышает 10 лет, в распределениях, представленных на рис. 2 просто не учитывались.

- значения моды для распределения по *CgHL* заметно больше значений моды соответствующих распределений по *CdHL* (распределение по *CgHL* смещено вправо по отношению к соответствующему распределению по *CdHL*);
- величина *FWHM* в случае *CgHL* всегда меньше, чем в случае *CdHL*. Визуально это выглядит следующим образом: распределение по *CgHL* в заданном году всегда выше и уже распределения по *CdHL* в этом же году.

Помимо подобия графиков распределений, принадлежащих одному и тому же классу, можно отметить и различия между распределениями, принадлежащих к разным классам. Так, на рис. 2 видно, что значения моды у распределений *CgHL* существенно больше, чем в случае распределений *CdHL*, а значение *FWHM* на графиках *CgHL* меньше, чем этот параметр для графиков *CdHL*.

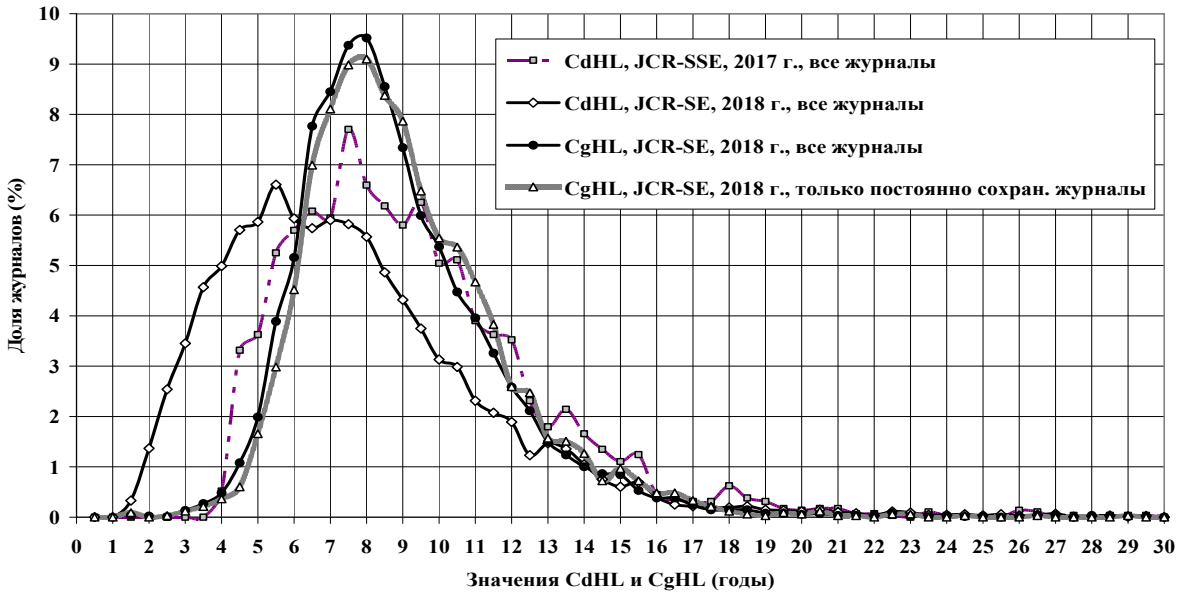

Рис. 1. Полные (не усеченные) распределения журналов по *Cited Half-life и Citing Half-life*
Две кривые, которым соответствуют первая и вторая строка легенды (считая сверху) соответствуют распределениям журналов по *CdHL,* другие две кривые (третья и четвертая строка легенды) – распределениям по CgHL.

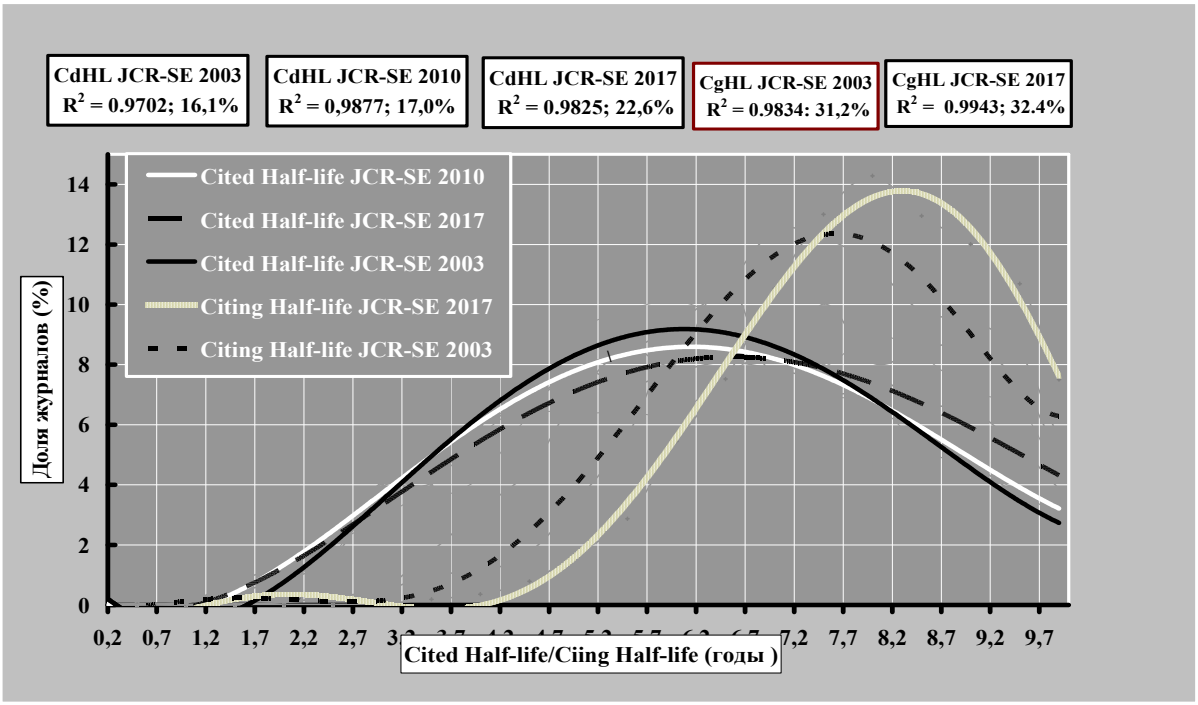

Рис. 2. Примеры усеченных распределений *Cited Half-life* и *Citing Half-life* за 2003, 2010, 2017 гг.

Исходя из вышеизложенного, можно сформулировать следующие утверждения.

Утверждение 1. Однотипные распределения, даже если они получены за различные моменты времени и на различающихся массивах журналов, по своей форме и основным характеристикам достаточно близки друг к другу.

Утверждение 2. Рассмотрим два распределения: *C* и *D*, и пусть наборы значений *CdHL* (группы журналов) в этих распределениях полностью совпадают. Далее, пусть распределение *C* содержит больше журналов, чем распределение *D*. Можно предположить, что в распределении, которое характеризуется бо́льшим числом журналов (распределение *C),* в некоторую группу попадет больше журналов, чем в такую же группу распределения, которое имеет меньшее числом журналов (распределение *D).*

Утверждение 3. Рассмотрим два распределения: *F* и *H*. Пусть общее число журналов распределения *F* равно общему числу журналов распределения *H,* а число журналов в какой-либо части (в основной части или в хвосте) распределения *F* больше, чем число журналов в соответствующей части распределения *H*. И пусть некоторая группа m_F находится в заданной части распределения *F*, а некоторая группа m_H – в такой же части распределения *H*. При этом потребуем, чтобы по значению показателя *CdHL* эти группы совпадали, т.е. выполнялось условие: $m_F = m_H$. Можно предположить, что при выполнении всех сформулированных нами условий и требований в том распределении, у которого в заданной части содержится бо́льшее число журналов, в группу m_F попадет больше журналов, чем в равную ей группу m_H того распределения, которое содержит меньшее число журналов в той же части (в распределении *H*).

Утверждение 4. Суммируя Утверждения 2 и 3, можно заключить, что чем больше журналов в некотором распределении и чем бо́льшая доля данной части (основной или хвоста) распределения, тем выше вероятность попадания журналов в ту или иную группу, находящуюся в этой части распределения. Другими словами – число журналов в некоторой группе (положительно) зависит от общего числа журналов в распределении, а также находится в некоторой положительной зависимости от того, какую долю от общего числа журналов занимает та часть распределения, в которой находится данная группа.

Рабочая гипотеза. Утверждения 1 и 4, которые, по сути, являются взаимно дополняющими, будем рассматривать в качестве рабочей гипотезы, которая в более компактном виде может быть сформулирована следующим образом: чем больше журналов, распределенных по значениям показателя *CdHL/CgHL,* и чем большая доля той части распределения, в которой находится заданная группа журналов с определенным значением соответствующего показателя, тем выше вероятность попадания журналов в эту группу. Другими словами: число журналов в некоторой группе (положительно) зависит от общего числа журналов в распределении, а также находится в некоторой положительной зависимости от того, какую долю от общего числа журналов занимает та часть распределения, в которой находится эта группа. Проверка и использование этой гипотезы (в случае подтверждения её справедливости), как мы надеемся, позволит с достаточной степенью точности, преобразовывать усеченные распределения в полные. Это, в свою очередь, даст возможность «восстанавливать» данные о числе (доле) журналов в группах, находящихся в хвосте соответствующих распределений (именно эта часть распределения нас, прежде всего, интересует, хотя, в принципе, положения рабочей гипотезы в равной мере распространяются и на основную часть распределений). Напомним, что под группой журналов мы понимаем численное значение (измеряемое в годах) показателя *CdHL/CgHL*, а под числом журналов в данной группе – количество журналов, каждый из которых характеризуется значением показателя *CdHL/CgHL*, соответствующим этой группе.

Определение 7. Реперным (базовым) будем называть распределение, которое является полным и с помощью которого предполагается дополнить распределение, являющееся неполным (усеченным). Основное требование к реперному распределению – все его группы должны иметь числовое значение, т.е. отсутствует группа с текстовым значением вида «>10.0».

Попытаемся формализовать рабочую гипотезу, чтобы с её помощью можно было рассчитывать число журналов в соответствующих группах. Рассмотрим пару, образованную двумя распределениями журналов по *CdHL*, каждое из которых соответствует одному из ежегодных выпусков *JCR*: распределение *G*, соответствующее выпуску *JCR* за год *g*, и распределение *I*, соответствующее выпуску *JCR* за год *i*. При этом потребуем, чтобы в качестве первого элемента в любой такой паре всегда выступало реперное (полное) распределение, а в качестве второго – усеченное, неполное[8], т.е. из рассмотрения будут исключены все пары, состоящие из двух неполных распределений. Причем формируя очередную пару распределений, мы каждый раз будем выбирать в качестве реперного одно и то же распределение – соответствующее 2018 г.[9]. Такое постоянство в выборе реперного распределения будет способствовать бо́льшей сопоставимости результатов.

Оценим общее число пар, которое, возможно, потребуется обработать для решения задач, поставленных в настоящей статье. Исходя из имеющихся у нас данных и, несмотря на сформулированные ограничения на состав пар, общее число таких пар может насчитывать многие сотни и даже тысячи. Поясним, на чем основывается эта оценка. Анализируемый нами период охватывает 1997-2018 гг., следовательно, обработке и анализу могут быть подвергнуты 22 ежегодных выпуска *JCR*. Получаем 21 пару. Каждый выпуск *JCR* состоит из двух изданий: *JCR-SE* (охватывает тематику по естественным, точным и техническим наукам) и *JCR-SSE* (охватывает тематику по обществен-

[8] Последнее требование не является обязательным: возможна пара из распределений, соответствующих выпускам *JCR* за 2018 г. и 2017 г. Последнее также является полным.

[9] При этом у нас сохраняется возможность выбрать в качестве реперного другое распределение, а именно то, которое соответствует 2017 г., (см. сноску 8).

ным наукам) – это значит, что полученное число удвоится и составит 42 пары. Поскольку речь идет о двух показателях (*CdHL* и *CgHL*), то это число снова удвоится и составит уже 84 пары. В каждом из ежегодных тематических изданий *JCR* мы выделяем две совокупности журналов: (1) совокупность всех журналов, содержащихся в данном годовом тематическом выпуске; (2) совокупность только тех журналов, которые присутствуют во всех без исключения соответствующих ежегодных тематических изданиях в течение всего 22-х летнего периода. В итоге получаем 168 пар, и это только те пары, которые можно сформировать, рассматривая совокупности журналов без учета того, каким тематическим категориям *WoS* соответствует тот или иной журнал. Если же сформировать совокупности журналов, соответствующие тем или иным наборам категорий *WoS*, а также совокупности журналов, соответствующие конкретным странам, то число пар возрастет на порядки. Столь обширный экспериментальный материал позволяет надеяться на получение достаточно надежных результатов, которые с высокой степенью вероятности позволит принять или отбросить рабочую гипотезу.

Преобразуем распределение G в распределение *I*.

a. Выберем в качестве реперного распределения G, а именно, то распределение, которое соответствует выпуску *JCR* за 2018 г. Этот выпуск содержит все (полные) численные данные о значениях интересующего нас параметра для каждого журнала. Нам известно общее число журналов N_g в распределении G и число тех журналов n_g, каждый из которых в этом распределении имеет значение *CdHL*, превышающее 10 лет (хвост распределения), а также известно число журналов l_g, образующих основную часть распределения, т.е. число журналов у которых значение *CdHL* не превышает 10 лет.

b. Выберем в качестве второго элемента пары такое распределение *I*, которое соответствует выпуску *JCR*, содержащему неполные данные (т.е. одному из ежегодных выпусков за период 1997–2016 гг.). Это значит, что в распределении *I* приводятся численные данные о значении *CdHL* только для тех журналов, у которых *CdHL*≤10, а против каждого журнала, у которого *CdHL*>10, указано только «*CdHL*>10» и не более того. Несмотря на указанную неполноту данных в распределении *I*, нам все же известно (как и в случае распределения *G*) общее количество журналов N_i и число тех журналов n_i, каждый из которых характеризуется значением *CdHL*>10 (хвост распределения), а также известно число журналов l_g, образующих основную часть распределения, т.е. число журналов у которых значение *CdHL* не превышает 10 лет.

c. Вычислим отношение общего числа журналов в распределении *I* к общего числа журналов в распределении *G*. Обозначим это отношение через $\alpha_{i/g}$:

$$\alpha_{i/g} = \frac{N_i}{N_g} \qquad (1)$$

Определим также соответствующие доли хвостов (β_g и β_i) в распределениях *G* и *I*:

$$\beta_g = \frac{n_g}{N_g} \qquad (2)$$

$$\beta_i = \frac{n_i}{N_i} \ . \qquad (3)$$

Следовательно, доля основных частей этих распределений составит: для G – $(1-\beta_g)$, а для I – $(1-\beta_i)$.

d. Вычислим коэффициент $k_{g\to i}$, который назовем коэффициентом преобразования (трансформации) распределения *G* в распределение *I*.

Для хвоста этот коэффициент будет выглядеть следующим образом:

$$k_{g_t\to i_t} = \alpha_{i/g} * \frac{\beta_i}{\beta_g} \ , \qquad (4)$$

где $k_{g_t\to i_t}$ – коэффициент преобразования в случае хвоста распределения, индекс t при символах g и i в выражении $k_{g_t\to i_t}$ указывает на то, что речь идет именно о хвосте распределения.

Для основной части коэффициент преобразования соответственно выглядит следующим образом:

$$k_{g_b\to i_b} = \alpha_{i/g} * \frac{1-\beta_i}{1-\beta_g} \ , \qquad (5)$$

где $k_{g_b\to i_b}$ – коэффициент преобразования в случае основной части распределения, индекс b при символах g и i в выражении $k_{g_b\to i_b}$ указывает на то, что речь идет именно об основной части распределения.

e. Вычислим функцию преобразования для каждой группы m_g распределения *G* в соответствующую группу m_i распределения *I*, т.е. для каждой группы m_i распределения *I* вычислим число журналов r_{m_i}, которое ей соответствует (в ней содержится). В общем виде эта функция выглядит следующим образом:

$$f(r_{m_g} \to r_{m_i}) = k_{g\to i} * r_{m_g} \ , \qquad (6)$$

где: r_{m_g} – число журналов в группе m_g распределения *G*;

r_{m_i} – число журналов в группе m_i распределения *I*;

$f(r_{m_g} \to r_{m_i})$ – функция преобразования числа журналов r_{m_g} в группе m_g распределения *G* в число журналов r_{m_i} в группе m_i распределения *I*, при усло-

вии, что обе эти группы по значению *CdHL* равны друг другу, т.е. $m_i = m_g$.

h. Для групп, принадлежащих хвосту распределения, функция преобразования выглядит:

$$f_t(r_{m_g} \to r_{m_i}) = \alpha_{i/g} * \frac{\beta_i}{\beta_g} * r_{m_g} , \qquad (7)$$

а для группы, принадлежащей основной части распределения, функция преобразования примет вид:

$$f_b(r_{m_g} \to r_{m_i}) = \alpha_{i/g} * \frac{1-\beta_i}{1-\beta_g} * r_{m_g} . \qquad (8)$$

Упрощая формулу (7), получим, что число журналов r_{m_i} в заданной группе m_i, находящейся в хвосте распределения *I*, рассчитывается по формуле:

$$r_{m_i} = \alpha_{i/g} * \frac{\beta_i}{\beta_g} * r_{m_g} . \qquad (9)$$

Аналогично, после упрощения формулы (8), для основной части получим:

$$r_{m_i} = \alpha_{i/g} * \frac{1-\beta_i}{1-\beta_g} * r_{m_g} . \qquad (10)$$

Приведенные вычисления необходимо сделать последовательно столько раз, сколько групп p_g содержит распределение *G*. Причем каждый раз в формуле (9) и, соответственно, в формуле (10) указывается то число журналов r_{m_g}, которое соответствует текущей группе m_g распределения *G*.

Для определения числа итераций для преобразования усеченных распределений в полные введем следующие обозначения: p_g – число групп в реперном распределении *G;* p_{g_t} – число групп в хвосте распределения *G*.

Отметим, что если в распределении *G* содержится p_g групп, то количество итераций вычисления по формуле (9) необходимо выполнить p_{g_t} раз, а по формуле (10) необходимо выполнить $p_g - p_{g_t}$ раз, а, в целом, для всего распределения такие вычисления должны быть выполнены p_g раз.

В результате этих вычислений вместо исходного усеченного распределения *I* мы должны получить теоретическое распределение I_{theor}, близкое исходному, при этом не совсем с исходным совпадающее. Основное отличие полученного распределения I_{theor} от исходного *I*, состоит в том, что I_{theor} является полным распределением, тогда как исходное – усеченным, т.е. эти распределения различаются своими хвостовыми частями. Однако и основные части этих распределений также будут различаться: расчетные значения далеко не всегда могут совпадать со значениями, соответствующими исходному распределению. Забегая вперед, отметим, что именно степень различия (совпадения) между основными частями распределений *I* и I_{theor} может служить одним из критериев для оценки степени справедливости принятой рабочей гипотезы. В общем, расчетное (теоретическое) распределения I_{theor} в некотором смысле является виртуальным. Действительно, при всех приведенных выше расчетах, мы требуем, чтобы и количество групп, и сами группы (значения *CdHL*) в распределении I_{theor} были равны числу групп и соответствующим им значениям *CdHL* реперного распределения *G*. Это требование вытекает из сформулированной нами рабочей гипотезы, точнее из той ее части, в которой утверждается подобие (близость) форм графиков различных (но однотипных) распределений. При этом сопоставляемые графики соответствуют различающимся распределениям, т. е. распределениям, полученным из разных ежегодных выпусков *JCR*.

Различие между реперным распределением *G* и распределением I_{theor}, «восстановленным» с помощью реперного и в результате использования формул (9) и (10), будет состоять только в различии числа (доли) журналов в группах.

Анализ результатов этих вычислений и построений должен подтвердить или опровергнуть рабочую гипотезу. В случае ее подтверждения нам останется рассчитать средневзвешенные значения *CdHL/CgHL* для соответствующих совокупностей журналов, полученных (совокупностей) по данным различающихся по времени опубликования того или иного ежегодного выпуска *JCR*. Средневзвешенные значения можно будет вычислить по следующей несложной схеме:

(a) умножить значение данной группы (значение показателя) на число журналов в этой группе;

(b) просуммировать полученные произведения (сумму взять по всему распределению);

(c) разделить результат действия *(a)* на результат действия *(b)*. Результат действия *(c)* и будет искомым средневзвешенным значением показателя *CdHL/CgHL*, которое характеризует данную совокупность журналов в данный момент времени.

Формализуя эту схему, получим:

$$W\_CHL_i = (\sum_1^p m_i * r_{m_i}) / \sum_1^p r_{m_i} , \qquad (11)$$

где W_CHL_i – средневзвешенное значение *CdHL/CgHL*, характеризующее совокупность журналов, соответствующую распределению *I*.

Полученные значения для каждого распределения журналов следует нанести на график, на оси ординат которого будут отложены средневзвешенные значения *CdHL/CgHL*, а на оси абсцисс – годы опубликования выпусков *JCR*.

ЗАКЛЮЧЕНИЕ

В настоящей работе поставлена задача разработки методики и способов определения динамики значения показателей периода полужизни *Cited Half-life* (*CdHL*) и *Citing Half-life* (*CgHL*). С этой целью по данным ежегодных выпусков «*Journal Citation Reports – Science Edition» (JCR-SE)* и «*Journal Citation Reports – Social Science Edition» (JCR-SSE)* за 22–х летний период (1997-2018 гг.) сформированы и соответствующим образом проанализированы массивы мировых журналов. Установлено, что для первых 20 ежегодных выпусков JCR (1997-2016 гг.) существует проблема неполноты (неточности) данных. Для этих выпусков каждого из журналов, у которого значение показателя CdHL (CgHL) превышает 10 лет, в JCR приводится не конкретное численное значение соответствующего показателя, а текстовое выражение вида «>10.0».

Для решения проблемы неполноты данных нами было предложено два метода, один из которых потребовал разработки рабочей гипотезы, состоящей в следующем. Чем больше журналов, представленных в виде распределения по значениям показателя *CdHL/CgHL,* и чем большая доля той части распределения, в которой находится некоторая группа журналов с определенным значением соответствующего показателя, тем выше вероятность попадания журналов в эту группу. Другими словами: число журналов в некоторой группе (положительно) зависит от общего числа журналов в распределении, а также находится в некоторой положительной зависимости от того, какую долю от общего числа журналов занимает та часть распределения, в которой находится эта группа. Напомним, что под группой журналов мы понимаем численное значение (измеряемое в годах) показателя *CdHL/CgHL*, а под числом журналов в данной группе – количество журналов, каждый из которых характеризуется значением показателя *CdHL/CgHL*, соответствующим этой группе. Предложенная гипотеза основывается на ряде полученных в настоящем исследовании результатов. В частности, на наблюдении, согласно которому однотипные распределения, даже если они получены за разные моменты времени и на различающихся массивах журналов, по своей форме и основным характеристикам достаточно близки друг к другу.

Последующая формализация этой гипотезы привела к разработке несложного математического аппарата, который в случае подтверждения справедливости предложенной гипотезы позволит «восстанавливать» недостающие данные и с помощью этих данных рассчитывать средневзвешенные значения показателей *CdHL* и *CgHL*. Что, в свою очередь, даст возможность определять тенденции в изменении этих значений для заданного набора журналов в заданные интервалы времени.

* * *

Авторы выражают глубокую признательность профессору Р.С. Гиляревскому за ряд ценных советов и критических замечаний, касающихся методики настоящего исследования, истории рассматриваемых показателей и их истолкования. Авторы также благодарны сотруднику компании *Clarivate Analytics* В.Г. Богорову за активное участие в обсуждении рассматриваемых в статье проблем и консультации, касающиеся структуры данных *JCR*.

СПИСОК ЛИТЕРАТУРЫ

1. Burton R.E., Kebler R.W. The “half-life” of some scientific and technical literature // American Documentation. – 1960. – Vol. 11, № 1. – P. 18-22.
2. Price D. General theory of bibliometrical and other cumulative processes // Journal of the American Society for Information Science. – 1976. – Vol. 29, № 5. – P. 292-206.
3. Line M.B. The “half-life” of periodical literature: apparent and real obsolescence / note by B.C. Vickery // Journal of Documentation. – 1970. – Vol. 26, № 1. – P. 46–54.
4. Мотылев В.М. Старение научно-технической литературы. – Л.: Наука, 1986. – 159 с.
5. Михайлов А.И., Черный А.И., Гиляревский Р.С. Основы информатики / изд. 2-е перераб. и доп. – М.: Наука, 1968. – С. 97–98.
6. Михайлов А.И., Черный А.И., Гиляревский Р.С. Научные коммуникации и информатика. – М.: Наука, 1976. – С. 173–176.
7. Либкинд А.Н, Маркусова В.А., Либкинд И.А., Янц М., Иванов К.Н. Моделирование динамики процесса сохранения журналов в качестве наиболее авторитетных научных изданий // Научно-техническая информация. Сер. 2. – 2013. – № 3. – С. 9-34; Libkind A.N., Markusova V.A., Libkind, I.A. Jansz M., Ivanov K.N. Modeling the dynamics of the retentivity process of journals among the most authoritative scientific serials // Automatic Documentation and Mathematical Linguistics – 2013 – Vol. 47, № 2. – P. 69–92

Материал поступил в редакцию 27.02.20.

Сведения об авторах

ЛИБКИНД Александр Наумович – кандидат технических наук**,** ведущий научный сотрудник ВИНИТИ РАН, Москва
e-mail: anliberty@mail.ru

МАРКУСОВА Валентина Александровна – доктор педагогических наук, ведущий научный сотрудник ВИНИТИ РАН, Москва
e-mail: markusova@viniti.ru

ЛИБКИНД Илья Александрович – Ведущий программист, PerformIT, Москва
e-mail: anliberty@mail.ru