ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 050: [002:001.8]

Р.С. Гиляревский, А.Н. Либкинд, В.Г. Богоров, И.А. Либкинд

Вычисление периода полужизни научных журналов в условиях неполноты данных *Journal Citation Reports**

Решается задача определения динамики показателей периода полужизни научных журналов Cited Half-life (CdHL) и Citing Half-life (CgHL)) в условиях существенной неполноты и недостаточной точности данных, содержащихся в Journal Citation Reports (JCR). Выполнен детальный анализ этих данных для показателей CdHL и CgHL за период 1997–2018 гг. Уточнена ранее сформулированная гипотеза о подобии распределений журналов по показателям периода полужизни. Проработаны методы определения средневзвешенных значений соответствующих показателей. Проверка гипотезы подобия позволила определить временные рамки, в пределах которых ее справедливость не вызывает сомнений. Показано, что динамика значений показателей периода полужизни положительна, причем эта динамика значительно более ярко выражена для постоянно сохраняющихся журналов, т.е. таких журналов, каждый из которых присутствовал в JCR в течение всего 22-х летнего исследуемого периода.

Ключевые слова: *старение литературы, период полужизни, Cited Half-life, Citing Half-life, Journal Citation Report, распределения журналов, гипотеза подобия*

DOI: 10.36535/0548-0027-2020-11-2

ВВЕДЕНИЕ

Когда Дж. Бернал в 1958 г. высказался о желательности измерения скорости старения литературы, а Р. Бартон и Р. Кеблер в 1960 г. использовали для этого показатель периода полужизни статей, они, в первую очередь, интересовались изучением темпов развития различных областей науки. В то время для этого приходилось затрачивать много сил, поскольку источники ссылок на статьи были традиционными, и прослеживать их приходилось вручную. Данные, опубликованные в [1], до сих пор приводятся в качестве примера старения литературы в различных науках, хотя они давно уже устарели. Теперь мы располагаем оцифрованными источниками о цитировании научных статей, среди которых наиболее надежным и авторитетным служит *Journal Citation Reports (JCR)*.

Однако этот источник долгое время был ограничен ретроспективой старения в 10 лет, после чего отсутствие точных данных обозначалось текстовым выражением вида «>10». Это легко понять, если предположить, что число номеров каждого из 10 тыс. журналов, число статей в каждом номере, число ссылок в каждой статье увеличивается в 10 раз (хотя на самом деле в *JCR* эта величина выше). Тогда число ссылок, которое надо проследить за 10 лет, составит 100 млн. Правда, с 2017 г. *JCR* это ограничение снял, что позволило разработать в рамках нашего проекта методику подсчета, «восстанавливающую» полную ретроспективу значений показателей периода полужизни [2]. Тем не менее, если учитывать конечную цель таких подсчетов в вычислении периода полужизни статей в отраслях знания, то можно сказать, что трудности на этом не заканчиваются. *JCR* публикуется ежегодно в двух различных по тематике выпусках, в которых журналы частично дублируются, поскольку статьи в них могут относиться к разным тематическим категориям. Да и сами эти тематические категории в Web of Science (WoS) и JCR с течением времени изменяются, а также могут переходить из одного тематического выпуска в другой. Некоторые журналы в определенные годы могут исчезать из *JCR* из-за временного снижения импакт-фактора или по другим причинам.

Настоящая работа посвящена преодолению названных трудностей для полноценного исследования

* Работа выполнена во исполнение государственного задания ВИНИТИ РАН по теме 0003-2019-0001 и при поддержке Российского фонда фундаментальных исследований (проекты РФФИ 20-07-00014 и 20-010-00179.

динамики показателей полужизни журнальных статей – *Cited Half-life* (*CdHL*) и *Citing Half-life* (*CgHL*). Показатель *CdHL* (период полужизни цитируемых статей журнала) определяется на основе данных о годе опубликования тех статей[1] из определенного журнала, которые в заданные годы[2] были процитированы другими периодическими изданиями. Соответственно, показатель *CgHL* (период полужизни цитирующих статей журнала) вычисляется на основе данных о годе опубликования тех статей из других журналов, которые в заданные годы[3] были процитированы этим журналом. Полученные сведения помогут проследить тенденции развития отраслей знания для последующего прогнозирования перспективности тематики экспериментальных исследований.

Сама идея определения динамики показателей периода полужизни основывается на том, что для соответствующего набора журналов (например, категорий JCR и/или их совокупностей) того или иного ежегодного выпуска *JCR* необходимо вычислить некоторое обобщающее для этого распределения значение показателя. В этом качестве естественно использовать его средневзвешенное значение, которое учитывает не только все показатели в распределении, но и вес (число журналов, их долю) каждого из них. Расположив значения по годам (в таблице или на графике), мы сможем увидеть динамику соответствующего показателя.

ИСХОДНЫЕ ДАННЫЕ, ИХ ОСОБЕННОСТИ И ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА

В качестве источников исходных данных были использованы ежегодные выпуски аналитико-статистической базы данных *Journal Citation Reports* (*JCR*) за период 1997–2018 гг. Эта БД расположена на платформе *Web of Science* (*WoS*) компании *Clarivate Analytics*. *JCR* публикуется ежегодно в виде двух тематических выпусков: *Journal Citation Reports – Science Edition* (*JCR-SE*) и *Journal Citation Reports – Social Science Edition* (*JCR-SSE*). В основном *JCR* представляет результат статистической обработки двух журнальных баз данных *WoS*. А именно: *JCR-SE* – результат обработки БД *Citation Index-Expanded – SCI-E* (естественные, точные и технические науки) и *JCR-SSE* – результат обработки БД *Social Sciences Citation Index – SSCI* (общественные науки)[4]. Каждый из тематических выпусков содержит характеристики соответствующих единиц двух важнейших информационных объектов (на языке теории баз данных – сущностей):

1) информационный объект «Журнал»: в качестве единицы (экземпляра) этого информационного объекта выступает один из журналов, включенных в *JCR*;

2) информационный объект «Тематическая категория *WoS*»: в качестве единицы (экземпляра) этого информационного объекта выступает одна из категорий *WoS*, включенных в *JCR*.

Прежде чем перейти к детальному описанию каждого из рассматриваемых информационных объектов, приведем ситуации, которые были обнаружены при анализе значений показателей периода полужизни указанных информационных объектов и их единиц. Обнаруженные ситуации можно сгруппировать следующим образом:

a) *«раздвоение» отдельных единиц объектов*. Раздвоение 1-го типа – это ситуация, когда в одном и том же году журнал или категория *WoS* оказываются представленными одновременно и в *JCR-SE*, и в *JCR-SSE в одной и той же категории WoS*. Раздвоение 2-го типа – это ситуация, когда в одном и том же году журнал оказывается представленным одновременно в нескольких категориях *WoS;*

b) *миграция отдельных единиц объектов.* Это понятие целесообразно применять только для описания «поведения» категорий *WoS*: ситуация, когда некоторая категория *WoS*, находясь в определенном году в конкретном тематическом выпуске, через несколько лет перемещается в другой тематический выпуск, например из *JCR-SE* в *JCR-SSE* или наоборот;

---

[1] Если придерживаться терминологии *JCR*, то следует сказать, что помимо статей учитываются и некоторые другие публикации, а именно такие, которые в принципе могут быть процитированными (*citable items*), в частности, обзоры.

[2] Обычно это годы, следующие за годом опубликования первого ежегодного комплекта цитирующих журналов.

[3] Обычно это годы, следующие за годом опубликования первого ежегодного комплекта цитируемых журналов.

[4] Именно эти две БД, согласно *Clarivate Analytics*, являются единственными источниками исходных данных для *JCR*. Так, на сайте *Clarivate Analytics* в разделе «*Journal Citation Reports Help*» читаем: «*Journal Citation Reports aggregates the meaningful connections of citations created by the research community through the delivery of a rich array of publisher-independent data, metrics and analysis of the world's most impactful journals included in the Science Citation Index Expanded* (*SCIE*) *and Social Sciences Citation Index* (*SSCI*) …» (http://jcr.help.clarivate.com/Content/home.htm).

Однако, как показывает анализ, значительная часть журналов, отражающихся в БД *Arts & Humanities Citation Index–A&HCI* (искусство и гуманитарные науки) также представлены в *JCR*. Так, например выпуск *JCR* за 2018 г. содержит 95 журналов, соответствующих категории «*History*» (назовем эти журналы «журналами по истории»), которая согласно классификации *Web of Science* (*WoS*), включена в БД *A&HCI*. Причем более половины этих журналов (53 из 95) *WoS* относит к единственной категории, а именно, к категории «*History*» и только к этой категории. Из оставшихся 42 журналов по истории четыре относятся также к категории «*Cultural Studies*», и два – к категории «*Ethic Studies*» (обе категории соответствуют гуманитарной проблематике). Таким образом, из 95 журналов по истории более 59 (53+6), т.е. 62% представлены в *WoS* только в БД *A&HCI*. Добавим, что из 4739 наименований журналов, которые были включены хотя бы в один ежегодный выпуск *JCR-SSE* за период 1997-2018 гг., 264 журнала согласно классификации *WoS* были отнесены только к гуманитарным наукам, а общее число журналов, которые посвящены, прежде всего, гуманитарным проблемам, за это период составило 315 наименований. Следовательно, используя *JCR*, мы можем быть уверены, что в анализ попадают и чисто гуманитарные тематики и соответствующие им журналы.

ЗАМЕЧАНИЕ. Возможно, что мнение о том, что *JCR* не содержит журналы из БД *A&HCI* связано с тем, что в *JCR* для этих журналов не приводятся значения импакт-фактора. В связи с этим отметим, что в *JCR* значения остальных показателей, которые в последние годы являются «стандартными» для этого издания (и что особенно важно для целей настоящей статьи – показатели периода полужизни), как правило, приводятся.

c) *полное отсутствие данных*. Ситуация, когда в соответствующих полях *JCR* вместо числового значения показателя указывается «*Not Available*» («Данные недоступны» или «Данные отсутствуют»);

d) *неточные (недостаточно точные) данные*. Ситуация, когда в соответствующих полях *JCR* для периода полужизни (в случае, если значение показателя превышает 10 лет) вместо конкретного числового значения приводится текстовое выражение вида «>10.0».

Журналы

Число журналов в ежегодном выпуске *JCR-SE* за рассматриваемый период почти удвоилось: 4962 названий в 1997 г. и 9156 – в 2018 г. Аналогичная тенденция характерна и для *JCR-SSE*: 1672 и 3382 журнала в 1997 г. и в 2018 г. соответственно.

Журнал в годовом выпуске *JCR* может соответствовать одной или более категориям *WoS*, т.е. между журналом и категориями *WoS* в *JCR* установлено отношение вида «один-ко-многим» (раздвоение 2-го типа). Естественно, что журнал может присутствовать также и в различных тематических ежегодных выпусках *JCR* (раздвоение 1-го типа): как в *JCR-SE*, так и в *JCR-SSE*. Очевидно, что эти ситуации, т.е. раздвоение 1-го и 2-го типов, для журналов совершенно естественны и являются проявлением многоаспектности исследований, публикуемых в таких журналах. Отметим, что это не препятствует дальнейшей обработке данных.

Таблица 1

***JCR-SE*: характеристики массивов журналов с точки зрения возможностей оценки периода полужизни изданий по естественным, точным и техническим наукам**

Годы *JCR*	Все журналы	*Для Cited Half-life (CdHL)*					*Для Citing Half-life (CgHL)*				
		Все журналы, %		Абсолютно сохранившиеся журналы, %			Все журналы, %		Абсолютно сохранившиеся журналы, %		
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1997	4962	**15,7**	13,6	68,8	14,5	9,6	**30,0**	4,8	68,8	29,9	0,8
1998	5467	**15,3**	16,0	62,4	15,6	7,6	**29,3**	4,0	62,4	30,2	0,5
1999	5550	**15,2**	13,9	61,5	16,1	5,3	**29,6**	4,3	61,5	30,5	0,6
2000	5686	**15,1**	13,3	60,0	16,7	4,2	**30,6**	4,4	60,0	32,6	0,6
2001	5752	**15,1**	12,1	59,3	17,0	3,3	**30,9**	3,9	59,3	32,7	0,4
2002	5876	**15,5**	10,6	58,1	18,1	2,5	**31,5**	3,7	58,1	33,5	0,4
2003	5907	**15,9**	9,3	57,8	19,0	1,9	**31,9**	3,4	57,8	34,4	0,4
2004	5969	**16,3**	8,2	57,2	19,5	1,8	**31,2**	3,1	57,2	33,5	0,6
2005	6088	**16,3**	7,3	56,0	20,2	1,5	**31,7**	2,5	56,0	34,7	0,5
2006	6166	**16,5**	6,4	55,3	20,8	1,2	**31,6**	2,3	55,3	35,2	0,5
2007	6426	**16,1**	6,4	53,1	21,2	0,9	**31,3**	4,2	53,1	34,9	2,4
2008	6620	**16,6**	5,5	51,5	22,6	0,7	**31,5**	1,8	51,5	36,5	0,4
2009	7387	**16,3**	8,6	46,2	24,2	0,4	**31,8**	1,9	46,2	37,0	0,4
2010	8073	**15,6**	9,7	42,3	24,8	0,4	**32,7**	1,9	42,3	38,0	0,3
2011	8336	**16,1**	8,5	40,9	27,1	0,3	**32,9**	2,1	40,9	39,6	0,5
2012	8471	**16,5**	7,1	40,3	28,9	0,2	**34,1**	1,7	40,3	42,2	0,3
2013	8539	**17,4**	5,4	40,0	30,9	0,2	**36,2**	1,9	40,0	44,8	0,4
2014	8659	**18,6**	4,3	39,4	33,6	0,1	**29,6**	1,7	39,4	32,4	0,7
2015	8802	**19,5**	3,5	38,8	36,3	0,2	**30,8**	1,6	38,8	33,9	0,6
2016	8866	**21,4**	**2,3**	38,5	39,4	0,1	**30,9**	1,7	38,5	**34,5**	0,8
2017	*9015*	***22,5***	*1,9*	*37,8*	*42,4*	*0,1*	***30,1***	*1,4*	*37,8*	*34,6*	*0,9*
2018	*9156*	***23,3***	*1,4*	*37,3*	*44,3*	*0,1*	***29,7***	*1,2*	*37,3*	*34,3*	*0,9*

ПРИМЕЧАНИЕ: **1** – годы *JCR*; **2** – общее число журналов; **3** и **8** – неточные данные (общий массив журналов): доля журналов со значениями показателя полужизни, превышающими 10 лет (от общего числа журналов); **4** и **9** – полное отсутствие данных (общий массив журналов): доля журналов, данные о значениях показателя полужизни которых отсутствуют (от общего числа журналов); **5** и **10** – доля абсолютно сохранившихся журналов (от общего числа журналов); **6** и **11** – неточные данные (абсолютно сохранившееся журналы): доля журналов со значениями показателя полужизни, превышающими 10 лет (от числа абсолютно сохранившихся журналов); **7** и **12** – полное отсутствие данных (абсолютно сохранившиеся журналы): доля журналов, данные о значениях показателя полужизни которых отсутствуют (от числа абсолютно сохранившихся журналов).

Таблица 2

***JCR-SSE* : характеристики массивов журналов с точки зрения возможностей оценки периода полужизни изданий по общественным наукам**

Годы *JCR*	Общее число журналов	*Для Cited Half-life (CdHL)*					*Для Citing Half-life (CgHL)*				
		Все журналы, %	Абсолютно сохранившиеся журналы, %				Все журналы, %		Абсолютно сохранившиеся журналы, %		
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1997	1672	**12,1**	32,6	73,0	13,1	25,6	**25,3**	5,3	73,0	26,0	1,2
1998	1678	**12,3**	29,3	72,8	13,5	21,9	**26,6**	3,4	72,8	27,4	0,8
1999	1699	**13,9**	26,3	71,9	15,6	18,4	**26,3**	3,9	71,9	27,0	0,9
2000	1697	**15,0**	23,4	72,0	16,7	15,2	**30,2**	4,0	72,0	31,4	1,2
2001	1682	**16,9**	21,5	72,6	19,0	13,8	**32,4**	4,4	72,6	34,0	1,5
2002	1709	**17,9**	19,0	71,4	20,5	11,1	**33,5**	3,2	71,4	36,4	0,6
2003	1714	**19,5**	17,2	71,2	22,5	9,3	**35,5**	3,2	71,2	37,8	0,7
2004	1712	**20,6**	14,6	71,3	23,8	7,9	**37,1**	2,5	71,3	39,4	0,8
2005	1747	**21,3**	13,2	69,9	25,1	7,4	**38,9**	1,8	69,9	42,3	0,7
2006	1768	**22,1**	11,0	69,1	26,6	6,1	**39,9**	1,5	69,1	44,2	0,4
2007	1866	**22,0**	11,7	65,4	28,2	5,2	**40,1**	1,3	65,4	45,4	0,3
2008	1985	**22,9**	10,7	61,5	30,3	3,8	**41,6**	1,5	61,5	47,7	0,3
2009	2257	**25,1**	14,7	54,1	37,1	2,5	**43,3**	1,4	54,1	51,4	0,4
2010	2731	**20,3**	18,6	44,7	34,6	2,6	**41,3**	2,3	44,7	51,1	0,5
2011	2966	**20,4**	18,7	41,2	36,9	2,5	**44,1**	2,4	41,2	55,3	0,5
2012	3047	**21,3**	16,5	40,1	39,1	2,3	**45,8**	2,4	40,1	57,7	0,7
2013	3080	**23,3**	14,8	39,6	42,8	2,5	**51,6**	2,1	39,6	61,7	1,1
2014	3154	**24,8**	12,8	38,7	45,9	2,0	**46,9**	1,5	38,7	49,2	0,6
2015	3224	**27,0**	10,3	37,9	49,2	1,8	**48,1**	1,6	37,9	52,2	0,7
2016	3241	**31,3**	7,6	37,7	54,1	1,1	**50,7**	1,9	37,7	54,3	0,8
2017	*3312*	***33,9***	*5,3*	*36,9*	*59,1*	*0,8*	***46,8***	*1,1*	*36,9*	*49,4*	*0,7*
2018	*3382*	*35,3*	*4,1*	*36,1*	*60,8*	*0,9*	***48,4***	*1,2*	*36,1*	*52,7*	*1,0*

ПРИМЕЧАНИЕ: **1** – годы *JCR*; **2** – общее число журналов; **3** и **8** – неточные данные (общий массив журналов): доля журналов со значениями показателя полужизни, превышающими 10 лет (от общего числа журналов); **4** и **9** – полное отсутствие данных (общий массив журналов): доля журналов, данные о значениях показателя полужизни которых отсутствуют (от общего числа журналов); **5** и **10** – доля абсолютно сохранившихся журналов (от общего числа журналов); **6** и **11**– неточные данные (абсолютно сохранившееся журналы): доля журналов со значениями показателя полужизни, превышающими 10 лет (от числа абсолютно сохранившихся журналов); **7** и **12** – полное отсутствие данных (абсолютно сохранившиеся журналы): доля журналов, данные о значениях показателя полужизни которых отсутствуют (от числа абсолютно сохранившихся журналов)

Полное отсутствие данных в случае журналов. Сведения об этой ситуации, полученные нами в результате соответствующего анализа, приводятся для *JCR-SE* в табл. 1 (графы 4 и 9), а для *JCR-SSE* – в табл. 2 (графы 4 и 9). Для показателя *CdHL* доля таких журналов в *JCR-SE* достигает 13,6%. Для выпусков *JCR-SSE* доля журналов с полным отсутствием данных для показателя *CdHL* достигает почти трети (32,6%). Следует отметить, что с течением времени доля таких журналов заметно уменьшается. Так, если в случае *JCR-SE* для *CdHL* в выпуске 1997 г. эта доля составляет 13,6%, в 2010 г. – 9,7%, а в выпуске 2018 г. – 1,4%. В случае *JCR-SSE* в выпуске 1997 г. – 32,6%, в 2010 г. – 18,6%, а в выпуске 2018 г. – 4% (137 журнала из 3382). Аналогичная картина наблюдается для обоих тематических выпусков *JCR* и в случае другого показателя периода полужизни – *CgHL*, хотя цифры здесь значительно меньше. Так, максимальное значение доли журналов, у которых отсутствовали данные для показателя *CgHL* для случая *JCR-SE* – 4,8% (1997 г.), минимальное – 1,4%. Аналогичные цифры для *JCR-SSE* близки: 5,3% и 1,2% соответственно. Такое падение совершенно естественно: *JCR* со временем удается собрать информацию о недостающих данных.

Неточные данные в случае журналов. Во всех ежегодных выпусках за период 1997–2016 гг. при значениях того или иного показателя, превышающих

10 лет, в соответствующем поле просто указывается текстовое выражение вида «>10». Таким образом, только данные в двух ежегодных выпусках (2017 г. и 2018 г.) из 22 (1997–2018 гг.), попавших в анализ тематических выпусков, не страдают этим недостатком. Доля журналов с неточными данными в любом ежегодном выпуске всегда больше 10%, нередко составляет десятки процентов, в отдельных случаях превышает 50% и, что очень важно, со временем, как правило, увеличивается (см. графы 3 и 8 в табл. 1 и 2). Следует отметить, что тогда, когда эти показатели искусственно не ограничены числом «10», они нередко очень значительно превышают указанное ограничение в 10 лет. Так, для выпуска *JCR-SE* за 2017 г. максимальное значение *CdHL* составляет 105,1 года (!), а для *CgHL* – 113,2 (!!) года. В случае *JCR-SSE* эти цифры ниже, но также достаточно значительны: для *CdHL* – 56,9 года и для *CgHL* – 74 года.

Для настоящего исследования представляет интерес и такой признак журнала, как страна его издания. Как показал анализ, существуют ситуации, когда в разные годы для одного и того же журнала в *JCR* в поле «*Region*», указаны разные страны. Это, скорее всего, объясняется следующим. Данные о журнале и, следовательно, о стране его издания, поступают в *JCR* из заявки издателя журнала в компанию *Clarivate Analytics*. У журналов, особенно международных, иногда происходит смена издателя. Новый издатель может располагаться в стране, отличной от страны нахождения прежнего издателя. В итоге в заявке нового издателя журнала в *Clarivate Analytics* и, следовательно, в описании журнала в *JCR* появится страна, отличная от страны, указанной в предыдущих ежегодных выпусках. С тем, чтобы избежать ненужной неопределенности, в качестве признака «Страна» того или иного журнала мы приняли актуальное значение этого признака, т.е. то, которое приводится в *JCR* за 2018 г.

Тематические категории *WoS*

Ежегодные выпуски *JCR-SE* включают 172–178 тематических категорий *WoS*, а выпуски *JCR-SSE* – 54–58 категорий *WoS*[5] (см. табл. 3)**.** При этом набор категорий в ежегодных выпусках *JCR* со временем меняется, однако эти изменения незначительны.

Раздвоение категорий. В случае категорий может иметь место раздвоение 1-го типа, т.е. ситуация, когда одна и та же категория в одном и том же ежегодном выпуске находится и в *JCR-SE*, и в *JCR-SSE*. Например, в 2018 г. шесть категорий – «*Nursing*», «Psychiatry», «*History & Philosophy of Science*», «*Public, Environmental & Occupational Health*», «*Rehabilitation*», и «*Substance Abuse*» – присутствуют и в *JCR-SE*, и в *JCR-SSE*. Отметим, что с точки зрения логической структуры науки эта ситуация может быть вполне естественной. Однако при обработке таких данных возникает ряд трудностей, и мы, исходя из тематической классификации, принятой в *Web of Science (классификация WoS)*, «привязали» каждую такую категорию к тому тематическому выпуску *JCR*, которому согласно указанной классификации соответствует рассматриваемая категория: к *JCR-SE* или к *JCR-SSE*. В пользу такой привязки свидетельствует, например, следующий факт: подавляющее большинство журналов «раздваивающихся» категорий присутствует в одном и том же году и в *JCR-SE*, и в *JCR-SSE*. Например, в *JCR-SSE* категории «*History & Philosophy of Science*» (*H&PS*) в 2018 г. соответствует 46 журналов, в *JCR-SE* в этом же году – 62 журнала. При этом 37 журналов оказались общими. Таким образом, общее количество уникальных (неповторяющихся) журналов в «объединенной», теперь привязанной только к тематическому выпуску *JCR-SSE* категории *H&PS*, увеличится и составит 71 наименование (46 + (62–37)).

Миграция категорий WoS. Анализ выпусков *JCR* показал также, что встречаются случаи, когда категория *WoS*, находясь в одном году в определенном тематическом выпуске, через несколько лет его «покидает» и «мигрирует» в другой. Миграция категорий, в отличие от их раздвоения, только с большой натяжкой может быть объяснена логикой развития науки. Скорее всего, это результат определенной непоследовательности решений менеджмента *JCR*. Как и в случае с раздвоением категорий, такая мигрирующая категория, исходя из тематической классификации, принятой в *WoS*, была нами «привязана» к соответствующему тематическому выпуску *JCR*.

Полное отсутствие данных о показателях для категорий. В период 1997 – 2002 гг. у всех категорий *WoS* данные о значениях показателей периода полужизни просто отсутствуют: в соответствующих полях для каждой категории *WoS* в *JCR* указано «*Not Available*». Однако, начиная с 2003 г. и в *JCR-SE*, и в *JCR-SSE* ни для одной из категорий таких записей нет.

Неполные (неточные, приближенные) значения данных о показателях для категорий. В период 2003–2018 гг. для всех категорий, у которых значения *CdHL/CgHL* превышают 10 лет, вместо точного числового значения указывается текстовое выражение вида «>10» (напомним, что для периода 1997–2002 гг. эти данные полностью отсутствуют). Доля категорий с такими неточными данными со временем возрастает. Так, в *JCR-SE* для показателя *CdHL* эта доля в 2003 г. составляла 5,9 %, в 2018 г. – 13,5%., а для показателя *CgHL* – 10,6% и 16,3% соответственно. Для *JCR-SSE* доли категорий, которые характеризуются указанной неточностью значений показателей, оказались еще значительнее, а темпы возрастания этой доли выше, чем для *JCR-SE*: в 2003 г. в *JCR-SSE* эти цифры составляла 18,7 %, а в 2018 г. – 42,9%, а для показателя *CgHL* – 14,8% и 42,9% соответственно (см. табл. 3).

Заканчивая рассмотрение данных, которые в *JCR* приводятся для категорий *WoS*, нужно отметить, что самый существенный их недостаток – это то, что каждый ежегодный выпуск обязательно содержит категории, у которых вместо значений *CdHL/CgHL* указано «>10». Причем со временем число и доля таких категорий увеличивается. Эти обстоятельства, к сожалению, делают совершенно невозможным использование таких данных при расчете средневзвешенных показателей *CdHL/CgHL* с помощью описанного далее «Метода виртуализации распределений».

[5] 80–85% категорий *JCR-SSE* приходится на общественные науки и 15–20% – на гуманитарные науки

Таблица 3

Количество категорий *WoS*, включенных в *JCR* и их доля с неточными данными

Год *JCR*	*Journal Citation Report – Science Edition (JCR-SE)*			*Journal Citation Report – Social Science Edition (JCR-SE)*		
	Всего категорий *WoS*	Доля категорий с неточными данными (текстовые значения «>10.0»), %		Всего категорий *WoS*	Доля категорий с неточными данными (текстовые значения >10), %	
		CdHL	*CgHL*		*CdHL*	*CgHL*
2003	170	5,9	10,6	54	16,7	14,8
2004	170	4,7	10,0	54	18,5	14,8
2005	171	4,1	8,8	54	18,5	14,8
2006	172	4,1	8,7	55	18,2	16,4
2007	172	4,1	8,7	55	20,0	20,0
2008	173	4,6	0,0	56	25,0	0,0
2009	174	5,2	0,0	56	28,6	17,9
2010	175	4,6	8,0	57	22,8	22,8
2011	178	5,6	11,2	56	26,8	25,0
2012	170	6,5	0,0	56	26,8	0,0
2013	176	7,4	0,0	56	32,1	0,0
2014	176	9,1	14,2	56	32,1	37,5
2015	177	10,2	15,3	57	35,1	42,1
2016	177	12,4	16,9	57	35,1	50,9
2017	178	13,5	16,9	60	40,0	40,0
2018	178	13,5	16,3	56	42,9	42,9

Для того, чтобы обойти эту трудность, мы применили некоторый опосредованный подход. Он состоит в том, что вместо данных о категориях мы использовали данные о тех журналах, которые соответствуют той или иной категории. Это значит, что вместо некоторой категории *WoS* в качестве ее «полноправного представителя» будет выступать совокупность тех журналов, которые этой категории соответствуют. При таком подходе все вычисления и преобразования, которые необходимо выполнить для определения динамики значений некоторого показателя категории, производятся над данными ее журналов-представителей, тогда как собственно данные указанной категории при опосредованном подходе, ввиду их недостаточности, в этих целях не используются.

ФОРМИРОВАНИЕ СОВОКУПНОСТЕЙ (МНОЖЕСТВ И ПОДМНОЖЕСТВ) ЖУРНАЛОВ

Рассматриваемые множества журналов, представленных в ежегодных выпусках *JCR*, в определенном смысле являются представителями всей мировой науки. С одной стороны, это позволяет утверждать, что те данные о тенденциях значений показателей периода полужизни, которые получатся на основе анализа этих множеств, будут в максимальной степени представительными и надежными. Однако следует отметить и следующее. Эти множества представляют конгломерат журналов. Действительно, каждый ежегодный выпуск *JCR* включает журналы из многих десятков стран (более 80) и соответствует сотням (более 250) тематических категорий *WoS*. Можно предположить, что тенденции, характерные для одного направления исследований (тематической категории *WoS*), наложатся на тенденции, характерные для других направлений исследований и, возможно, взаимно исказят (затемнят) друг друга. То же можно предположить и в отношении региональной составляющей: тенденции значений показателей полупериодов жизни для подмножества журналов, соответствующего некоторой стране, могут отличаться от тенденций для соответствующего подмножества журналов другой страны. Более того, можно ожидать, что такое различие по региональному признаку (стран издания) будет наблюдаться даже внутри одной той же категории *WoS*.

Можно также выделить те подмножества журналов, которые присутствуют в *JCR* на протяжении всего 22-х летнего периода. Тенденции такого «ядра» журналов могут отличаться от тенденций, характерных для всего соответствующего множества журналов. Учитывая все это, необходимо исследовать как тенденции, характеризующие все множество журналов, так и тенденции, характерные для отдельных его подмножеств. Эти подмножества будут сформирова-

ны с использованием следующих классификационных признаков:

a) принадлежность журнала к заданному (i-му) ежегодному выпуску (году опубликования) *JCR*;

b) принадлежность журнала к определенному тематическому выпуску *JCR* (к *JCR-SE* или к *JCR-SSE);*

c) принадлежность журнала к определенной тематической категории WoS и/или к заданному набору этих категорий;

d) принадлежность журнала той или иной стране издания;

e) факт присутствия журнала во всех без исключения ежегодных выпусках *JCR* за заданный период (в нашем случае за 1997–2018 гг.). В этом случае для нас не важно, в каких тематических выпусках (в *JCR-SE* или в *JCR-SSE*) присутствует журнал: важно только, чтобы он был в *JCR* в каждом году. Эти журналы назовем согласно работам [2, 3] *«Абсолютно сохранившимися журналами»* или *«Всегда присутствующими журналами* "*Absolutely preserved journals* or *always present journals (AP Journals)" or* "*Absolutely retentive journals (AR Journals)"*. Иногда, для краткости, мы их будем называть «постоянными». В качестве исходной точки на шкале времени в этом случае примем 1997 г. Можно считать, что такие журналы являются некоторым ядром мировых научных журналов.

Таким образом, сформулированную в начале статьи задачу следует уточнить и расширить, переформулировав ее следующим образом. Необходимо определить динамику каждого из двух показателей (*CdHL* и *CgHL*) путем сопоставления их средневзвешенных значений за определенные годы. Указанное сопоставление должно быть выполнено как для всего рассматриваемого множества журналов, так и для подмножеств журналов, соответствующих наиболее крупным областям знания (естественные, точные, технические, общественные и гуманитарные науки). Кроме того, следует определить динамику показателей для отдельных разделов этих областей (более 250 тематических категорий *Web of Science*), а также для определенных объединений этих категорий: физика, химия, биология, медицина, технические и сельскохозяйственные науки, история, философия и т.п. Важно также установить влияние региональной составляющей (страны) для каждого из рассматриваемых множеств и подмножеств журналов.

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 этого показателя, в полной мере касаются и показателя *CgHL*.

МЕТОДИКА РАСЧЕТА СРЕДНЕВЗВЕШЕННЫХ ЗНАЧЕНИЙ ПОКАЗАТЕЛЕЙ ПОЛУЖИЗНИ

Перейдем к обсуждению конкретных возможностей, предоставляющих имеющиеся данные для решения поставленных задач. Очевидно, что попытки определения динамики средневзвешенных значений рассматриваемых показателей, учитывая только те журналы, для которых значения показателей в *JCR* даются в числовой форме, приведут, в лучшем случае, к существенным искажениям. Ниже предлагаются использованные нами два независимых друг от друга метода расчета соответствующих значений, позволяющие избежать таких искажений, по крайней мере, существенно их ослабить. Один из этих методов (упрощенный, более чем приближенный), который можно назвать методом приписывания численных значений каждому такому журналу, у которого вместо конкретного значения показателя в *JCR* указано «>10», дает возможность действительно лишь очень приближенно оценить средневзвешенные значения заданного показателя. Второй – метод виртуализации распределений – является сложным, однако, он обеспечивает получение значительно более точных значений показателей, и тем самым позволяет получить более достоверную картину динамики исследуемых показателей.

Метод приписывания значений показателя

Этот метод состоит в том, чтобы заменить текстовое выражение «>10» (неполные, неточные данные) на числовое, например, на «10,1»[6]. Это значит, что каждому конкретному (здесь важно подчеркнуть – конкретному) журналу, характеризующемуся такими неполными данными значения показателя, «приписывается» числовое значение. Это позволяет при соответствующих вычислениях учитывать также и те журналы, у которых рассматриваемые показатели имеют значения, превышающие 10 лет. К сожалению, этот метод, хотя и дает возможность оценить, в качестве первого приближения, тенденции изменения рассматриваемых показателей, однако при расчете их средневзвешенных значений приводит к очень большим искажениям. Так, при приписывании минимально возможных численных значений (10,1 года) величина показателя будут существенно занижена. Напротив, при выборе достаточно большого числа, например, 15,0, возникает серьезная опасность необоснованного и очень существенного завышения средневзвешенного значения соответствующего показателя. В настоящем исследовании будет использовано минимальное число 10,1. Такой выбор, несмотря на его произвольность, все же позволяет утверждать, что в этом случае, по крайней мере, не будет нарушено неравенство $10{,}1 <= x$, где x – это одно из возможных значений показателя для какого либо журнала из усеченной части распределения.

Метод виртуализации распределений

Прежде чем перейти к непосредственному изложению этого метода, необходимо ввести ряд определений, а также изложить гипотезу, на которой этот метод основывается.

Распределение журналов по значениям показателя. Назовем распределением журналов по значениям показателя *CdHL/CgHL* таблицу, в первой графе (колонке) которой последовательно в порядке возрастания приводятся значения *CdHL/CgHL*, а во второй

[6] Выбор этого минимального значения может быть оправдан тем, что, по крайней мере не будет необоснованного завышения реальных значений того или иного показателя периода полужизни.

графе против каждого значения *CdHL/CgHL* из первой колонки указано число журналов, каждый из которых имеет именно это значение. В дальнейшем для краткости вместо выражения «распределение журналов по значениям показателя» будем писать «распределение», а вместо «распределения журналов по показателю *CgHL*» – «распределение *CgHL*».

Взаимно однотипные распределения. Под взаимно однотипными распределениями подразумеваем распределения журналов, соответствующие одному и тому же показателю. Так, взаимно однотипными распределениями являются различные распределения журналов по значениям показателя *CdHL*. Взаимно однотипны также все распределения журналов по значениям показателя *CgHL*. Важно помнить, что два распределения, одно из которых соответствует показателю *CdHL*, а другое – *CgHL*, не взаимно однотипны. Сопоставление взаимно разнотипных распределений друг с другом, как правило, не корректно. Оно необходимо только тогда, когда явным образом уточняется, что ставится задача сопоставления именно разнотипных распределений. Следует иметь в виду, что не всегда корректным будет и сопоставление взаимно однотипных распределений. Дело в том, что целесообразно, как правило, сопоставлять друг с другом только те взаимно однотипные распределения, которые принадлежат к одному и тому же классу распределений.

Классы распределений. Класс распределений – множество *однотипных распределений*, полученных на таких подмножествах журналов, каждое из которых сформировано с помощью одного и того же набора классификационных признаков. Так, например, в состав одного и того же класса входят все такие распределения, которые соответствуют показателю *CdHL* и при этом сформированы на основе признака «Страна издания». Еще два класса образуют распределения, причем одни из них соответствуют показателю *CdHL*, а другие – *CgHL* и, при этом полученные на подмножествах журналов, сформированных на основе признака «Категория *WoS*». Еще один пример класса: все распределения, соответствующие показателю *CgHL* и при этом полученные на подмножествах журналов, которые сформированы на основе двух признаков: «Категории *WoS*» **и** «Страна издания» (здесь «**и**» играет роль конъюнкции). Класс может включать набор распределений по *CdHL*, сформированных путем выделения журналов, соответствующих, например, категории «*Psychiatry*» и, в свою очередь, распределённых по странам издания этих журналов. Таким образом, в этот класс входят, в частности, распределения журналов по *CdHL*, которые могут быть описаны следующим образом: «*Psychiatry – Germany*», «*Psychiatry – Russia*», «*Psychiatry – USA*» и т.д.

Группы журналов в распределении по значениям заданного показателя (группы журналов). Для наших целей группа журналов – такая совокупность журналов в некотором распределении, у каждого из которых (журналов) значения заданного показателя численно равны друг другу. Таким образом, в распределении насчитывается столько групп, сколько в этом распределении насчитывается численно различающихся значений этого показателя: от минимального до его максимального значения. Например, группу в некотором распределении *CdHL* составляют журналы, у каждого из которых значение показателя CdHL равно 7,8 года.

Важно учитывать следующее обстоятельство. При равенстве значений двух различных показателей (*CdHL* и *CgHL* соответственно) количество и состав журналов в группе из распределения по показателю *CdHL* не обязательно совпадает (точнее, обычно не совпадает) с количеством и составом журналов в группе из распределения по показателю *CgHL*.

Замечание. Результаты предыдущего этапа нашего исследования изложены в работе [2], в которой использовались понятия «основная часть распределения» и «хвост распределения», причем за «точку раздела» было принято значение показателя, равное 10 годам. При этом за основную часть распределения признавалась та его часть, значения показателя у которой меньше либо равно 10. Соответственно, хвост распределения – та его часть, значения показателя у которой больше 10. Такие определения были «подсказаны» ситуацией, согласно которой почти во всех ежегодных выпусках в качестве максимального числового значения фигурирует именно число 10, а для значений, больших 10, указывается, как мы уже неоднократно отмечали, текстовое выражение вида «>10». К сожалению, предпринятые нами в дальнейшем попытки руководствоваться этими определениями показали их недостаточность. В итоге принято решение отказаться от такого деления распределения. Вместо этого, после серии экспериментов было приняты следующие определения, которыми мы руководствовались в настоящей работе при соответствующих вычислениях.

Части распределения. Рассмотрим распределение по некоторому показателю. Отложим по оси абсцисс значения показателя, т.е. численные обозначения групп журналов. В качестве точки, которая делит распределение на основную часть и хвост, примем значение показателя, соответствующее медиане распределения. Совокупность групп, которая находится слева от медианы, вместе с группой, которая соответствует медиане, будем называть *основной частью распределения*. Совокупность групп, которые расположены справа от указанной точки назовем *хвостом распределения*.

Вычисления, выполненные с целью уточнения алгоритмов и отладки программных средств, показали, что их точность возрастает с делением на более мелкие части. В настоящей работе мы остановились на делении каждой из частей распределения еще на две части. Казалось бы, что в этой ситуации можно отказаться от терминов «основная часть распределения» и «хвост распределения». Действительно, эти термины в нашем случае представляют собой некоторый анахронизм, они иллюстрируют только развитие подходов к решению задач настоящего исследования и связаны с историй этого исследования. Тем не менее, мы не отказываемся от этой терминологии, так как она может оказаться полезной в будущем при детальном анализе исследуемых распределений с точки зрения их типологии и статистических характеристик.

Теоретическое обоснование и описание метода виртуализации

Анализ данных, выполненный ранее в рамках нашего исследования в работе [2], показал, что структуры взаимно-однотипных распределений журналов для всех ежегодных выпусков *JCR* достаточно близки. А именно, доля, которую занимает такая группа в распределении, полученном на массиве журналов некоторого ежегодного выпуска *JCR*, близка к доле соответствующих (по значению заданного показателя) групп остальных ежегодных выпусков. Это оказалось справедливым и для долей более крупных частей распределений. На основании этих и некоторых дополнительных данных в работе [2] была предложена рабочая гипотеза, которую здесь, после некоторого ее уточнения, назовем и сформулируем следующим образом.

Гипотеза подобия распределений, принадлежащих одному и тому же классу. Распределения журналов, принадлежащих одному и тому же классу, по своей структуре подобны друг другу. При этом, чем больше журналов в каждом таком распределении и чем больше доля, которую занимает та часть распределения, в которой находится заданная группа журналов, тем выше вероятность попадания журналов в эту группу, т.е. тем больше журналов окажется в этой группе. Другими словами, число и доля журналов в некоторой группе распределения (положительно) зависят от общего числа журналов в этом распределении, а также находятся в некоторой положительной зависимости от того, какую долю от общего числа журналов занимает та часть распределения, в которой находится эта группа.

На основании этой гипотезы в работе [2] был предложен относительно несложный математический аппарат и, в свою очередь, *функция преобразования неполного (усеченного) распределения в распределение полное*[7]. С помощью этой функции из исходного усеченного распределения создается некоторое виртуальное распределение, которое мы будем рассматривать в качестве полноправного представителя исходного. Такая виртуализация осуществляется с использованием соответствующих данных некоторого полного (реперного) распределения, которое должно принадлежать к тому же классу, к которому принадлежит преобразовываемое распределение. В качестве реперных используются распределения, соответствующие 2018 г., которые, как неоднократно отмечалось выше, в случае журналов всегда являются полными. В качестве исходных данных в функции преобразования используются:

a) данные, которые в самом общем виде описывают преобразуемое распределение: общее число журналов в этом распределении, число журналов в соответствующей части распределения, число журналов, значения показателя у которых больше 10 лет, число журналов с полным отсутствием данных о значениях показателя;

b) данные из реперного распределения, аналогичные тем, которые приведены в (*a*), а также данные о детальной структуре этого (реперного) распределения, а именно: количество (число) журналов, которое соответствует тому или иному значению показателя в реперном распределении.

Помимо этого, в функции преобразования учитывается также число журналов, у которых полностью отсутствуют данные о значениях соответствующего показателя как в преобразовываемом (усеченном) распределении , так и в реперном (полном) распределении, а также вычисляется разность между этими числами, которая затем делится между соответствующими частями преобразуемого распределения. Это деление производится пропорционально доле, которая занимает та или иная часть преобразуемого распределения. Это значит, что такая численная «прибавка» включается в число журналов соответствующей части усеченного распределения.

Для того, чтобы в общем виде представить указанную функцию преобразования, введем следующие обозначения:

G – реперное (полное) распределение, *I* – преобразуемое усеченное распределение, I_{virt} – виртуальное распределение, получаемое с помощью итерационного применения функции преобразования $f(n_G^{d_j} \to n_{I_{virt}}^{d_j})$ и являющееся заместителем (представителем) исходного усеченного распределения *I*;

j – некоторая часть соответствующего распределения;

$n_G^{d_j}$ – число журналов в группе d_j^G, расположенной в части j_G распределения *G*;

$n_{I_{virt}}^{d_j}$ – число журналов в группе $d_j^{I_{virt}}$ расположенной в части $j_{I_{virt}}$ виртуального распределения I_{virt}, которое (число журналов) вычисляется с помощью функции преобразования $f(n_G^{d_j} \to n_{I_{virt}}^{d_j})$ и, требуя при этом, чтобы группа $d_j^{I_{virt}}$ виртуального распределения I_{virt} была численно равна (по значению рассматриваемого показателя) группе d_j^G реперного *G*, т.е. при условии $d_j^{I_{virt}} = d_j^G$;

$\alpha_{I/G}^N$ – отношение [общего числа журналов N_I в усеченном (преобразуемом) распределении *I*] к [общему числу журналов N_G в реперном распределении *G*];

$\mathrm{Re}l_{I/G}^{n_j}$ – отношение [числа журналов n_I^j (с учетом «прибавки», указанной в предыдущем абзаце), находящихся в данной части j_I распределения *I*] к [числу журналов n_G^j в аналогичной части j_G распределения *G*].

В итоге в общем виде функции преобразования может быть записана следующим образом:

$$f(n_G^{d_j} \to n_{I_{virt}}^{d_j}) = \alpha_{I/G}^N * \mathrm{Re}l_{I/G}^{n_j} * n_G^{d_j}$$

[7] Несмотря на то, что математический аппарат, включая функцию преобразования, был разработан, исходя из неуточненной формулировки гипотезы, этот аппарат, как оказалось, полностью соответствует уточнённой формулировке этой гипотезы.

Таблица 4

Динамика средневзвешенных значений *Cited Half-life*, объединенных массивов журналов (*JCR-SE* + *JCR-SSE*)

Год выпуска *JCR*	Для всех журналов			Для абсолютно сохранившихся журналов		
	Способ полной виртуализации	Способ частичной виртуализации	Способ приписывания значений	Способ полной виртуализации	Способ частичной виртуализации	Способ приписывания значений
	Период полужизни, годы (тысячные доли)					
1	**2**	**3**	**4**	**5**	**6**	**7**
1997	7,999	7,695	6,569	10,422	8,874	6,526
1998	8,056	7,830	6,640	10,450	8,955	6,667
1999	8,025	7,769	6,652	10,457	8,952	6,750
2000	8,031	7,787	6,668	10,423	8,963	6,839
2001	8,033	7,781	6,689	10,391	8,994	6,956
2002	8,057	7,790	6,700	10,331	9,023	7,055
2003	8,025	7,737	6,697	10,325	9,034	7,124
2004	8,000	7,734	6,749	10,293	9,099	7,231
2005	8,003	7,718	6,732	10,267	9,120	7,298
2006	8,047	7,739	6,750	10,258	9,161	7,402
2007	8,031	7,762	6,758	10,235	9,207	7,501
2008	7,992	7,726	6,782	10,197	9,274	7,616
2009	7,958	7,754	6,814	10,225	9,406	7,800
2010	8,010	7,707	6,680	10,241	9,424	7,828
2011	8,002	7,698	6,677	10,265	9,538	7,933
2012	8,000	7,754	6,732	10,353	9,676	8,057
2013	7,963	7,724	6,801	10,397	9,831	8,187
2014	7,955	7,759	6,884	10,505	9,994	8,313
2015	7,903	7,776	6,975	10,589	10,160	8,439
2016	7,820	7,776	7,124	10,676	10,367	8,608
2017	8,541	8,480	7,388	10,757	10,757	8,805
2018	8,541	8,541	7,415	10,926	10,926	8,845

Уточним, что такие вычисления должны быть последовательно выполнены для каждой группы журналов, образующих реперное распределение. Если реперное распределение G состоит, например, из S групп, то для получения соответствующего виртуального распределения следует осуществить S итераций таких вычислений. В итоге вместо исходного неполного (усеченного) распределения I мы получим некоторое виртуальное, но теперь уже полное распределение I_{virt}, включающее все те группы (значения показателя), которые содержит реперное распределение G, причем для каждой из этих групп виртуального распределения будет указано соответствующее ей вычисленное число журналов. Полученное таким образом виртуальное распределение I_{virt} в дальнейших расчетах рассматривается в качестве полноправного представителя (заместителя) исходного усеченного распределения I. Средневзвешенное значение показателя для данного исходного распределения I вычисляется уже не непосредственно на основе данных этого распределения, а опосредовано, т.е. на основе данных того виртуального распределения I_{virt}, которое построено как представитель исходного.

Следует отметить, что метод виртуализации и разработанные для его реализации программные средства позволяют для одного и того же распределения осуществлять вычисления средневзвешенного значения соответствующего показателя двумя различными, но частично совпадающими способами.

А. *Способ полной виртуализации*. На основании гипотезы подобия и с использованием функции преобразования для усеченного распределения предварительно строится соответствующее ему виртуальное распределение. Это значит, что из текущего (усеченного) распределения используются только следующие данные: общее число журналов, число (доля) журналов в соответствующих частях распределения, а также число журналов, у которых полностью отсутствуют данные о значениях показателя. Важно понимать, что при этом игнорируются все значения показателя у всех журналов, которые формируют исходное распределение, даже в тех случаях, когда реальное значение показателя <=10 годам.

Таким образом, виртуальное распределение, которое теперь является «заместителем» исходного усеченного, полностью строится только с помощью перерасчета соответствующих значений реперного (полного, не усеченного) распределения, конечно, с использованием тех данных из исходного распределения, которые перечислены в предыдущем абзаце. После построения виртуального распределения вычисляется сумма произведений каждого значения показателя этого распределения на соответствующее ему число журналов, а затем полученная сумма делится на общее число журналов в распределении. Результат этого деления представляет средневзвешенное значение соответствующего показателя и рассматривается нами в качестве представителя той совокупности журналов, которые образовали исходное распределение. Например, если исходное распределение образовано из журналов выпуска *JCR* 2010 г., то полученное таким способом значение *CdHL*, которое равно 8,011 (графа 2 табл. 4) характеризует эту совокупность журналов по состоянию именно в 2010 г.

В. С*пособ частичной виртуализации*. В этом случае преобразование исходного (усеченного) распределения осуществляется только для его усеченной части, т.е. той части распределения, которой соответствуют текстовые выражения вида «>10». Что касается части, у которой показатель принимает численные значения (т.е. не превышающие 10 лет), то она преобразованию не подвергается. Из этих двух частей составляется результирующее распределение, являющееся некоторым «склеенным» (гибридным) распределением, состоящим из двух частей: одна часть – это неизменённый фрагмент исходного распределения со значениями показателя <=10, другая – результат преобразования того фрагмента этого распределения, значения показателя которого превышает 10 лет. Средневзвешенное значение соответствующего показателя в этом случае вычисляется на основе данных такого склеенного (гибридного) распределения.

ДОСТОИНСТВА И НЕДОСТАТКИ ПРЕДЛОЖЕННЫХ СПОСОБОВ ВЫЧИСЛЕНИЯ РАСПРЕДЕЛЕНИЙ

В результате применения двух предложенных методов для одного и того же распределения по некоторому заданному показателю можно получить три различных значения средневзвешенного показателя:

1) средневзвешенное значение, полученное с помощью метода приписывания значений;

2) средневзвешенное значение, полученное с помощью метода полной виртуализации;

3) средневзвешенное значение, полученное с помощью метода частичной виртуализации.

Серьёзным недостатком метода приписывания значений является то, что приписываемое значение принимается произвольно, только исходя из того, чтобы оно было больше 10 лет, т.е. как этого требует заменяемое выражение («>10»). При этом принятое в настоящей работе минимально возможное значение (10,1 года) может заметно занижать вычисляемые таким образом значения. Достоинство этого метода – относительно несложные операции вычисления, хотя в случае использования компьютеров это оказывается не столь существенным. В любом случае, при помощи метода приписывания значений можно определить тенденцию в изменениях этих значений, если таковая имеет место, либо, напротив, убедиться в отсутствии каких-либо изменений.

Достоинство метода полной виртуализации состоит, прежде всего, в том, что он позволяет сопоставить исходное и виртуальное распределение, оценить степень их близости и, тем самым, дать возможность оценить насколько справедлива гипотеза подобия. Что касается недостатков, то здесь следует отметить, что вычисление средневзвешенного значения на основании данных только виртуального распределения полностью игнорирует реальные данные исходного и, тем самым, скорее всего, может существенно искажать результат. Очевидно, что наиболее точными окажутся результаты, полученные с помощью метода частичной виртуализации, так как в этом случае, с одной стороны, – используются все реальные данные исходного распределения, а, с другой, – с помощью виртуализации восполняются недостающие данные.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Сформируем массивы журналов по следующей схеме:

a) из тематического выпуска *JCR-SE* и тематического выпуска *JCR-SSE* для каждого ежегодного выпуска *JCR* сформируем объединенный массив уникальных журналов, т.е. журналов, неповторяющихся в рассматриваемом году. Это значит, что в такой массив попадут все журналы, которые в указанном году присутствовали хотя бы в одном из двух тематических выпусков *JCR*;

b) потребуем, чтобы журнал в обязательном порядке присутствовал в каждом ежегодном массиве, сформированном на предыдущем этапе. В итоге получим ежегодные массивы журналов, которые обычно мы обозначаем как абсолютно сохранившиеся журналы.

График средневзвешенных значений показателя *CdHL* именно для сформированных таким образом массивов абсолютно сохранившихся журналов приведен на рис.1. Эти значения получены тремя различными способами. Первый (способ полной виртуализации) полностью базируется на гипотезе подобия. Второй (способ частичной виртуализации) также использует гипотезу подобия, однако только для восполнения недостающих данных. И наконец, третий способ (способ приписывания значений) никак не обращается к гипотезе подобия. Следовательно, для ответа на вопрос, справедлива ли гипотеза подобия и если справедлива, то в какой степени и в каких пределах, нам нужно рассмотреть результаты, полученные с помощью двух первых способов.

Из графика на рис. 1 и соответствующих ему граф 5–7 табл. 4 следует, что результаты, полученные с помощью метода, который полностью опирается на гипотезу подобия (способ полной виртуализации), достаточно близки к тем результатам, которые получены исходя из указанной гипотезы подобия, но при обязательном использовании реальных (неусеченных) данных. Здесь важно подчеркнуть, что доля

этих данных в усеченных распределениях очень значительна (см. графу 6 табл. 1). Следовательно, можно утверждать, что полученные результаты не противоречат гипотезе подобия, а на отрезках времени в 8 – 10 лет средневзвешенные значения показателя *CdHL*, полученные с помощью метода, полностью базирующегося на гипотезе подобия (метод полной виртуализации), достаточно хорошо описывают динамику этого показателя. Таким образом, мы можем не только констатировать, что результаты не противоречат гипотезе подобия, но и указать те временные рамки, в которых справедливость этой гипотезы не вызывает сомнений.

В качестве более точной оценки справедливости гипотезы подобия можно было бы использовать разность между средневзвешенным значением заданного показателя (*CdHL* или *CgHL*), вычисленным для данного полного распределения (например, некоторого годового выпуска *JCR*) – с одной стороны, и значением аналогичного показателя, рассчитанного для этого же распределения, но на основе гипотезы подобия – с другой. К сожалению, возможности такой оценки в настоящее время ограничены. Действительно, единственный год выпуска *JCR*, когда можно выполнить такие сопоставительные вычисления – это 2017 г.: только в выпуске 2017 г. (как и в реперном выпуске 2018 г.) отсутствует ситуация, когда вместо конкретного значения показателя указано «>10». Выполненные нами вычисления для указанного годового выпуска *JCR* дали следующие значения: 10,725 (по реальным значениям) и 10,757 (способ полной виртуализации). Разность между этими вычислениями составила – 0,032 года (0,29%). Такая разность соответствует временному расстоянию между реперным (2018) и заданным (2017) годом, которое равно одному году. Если предположить, что указанная разность линейно зависит от величины временного расстояния, то это отклонение, например, для 2007 г. составит – 0,32 года (–0,032*10= – 0,32), что примерно соответствует данным на графике (рис.1). Этот график позволяет также оценить точность и достоверность результатов, полученных с помощью способа приписывания значений. Как и следовало ожидать, и как следует из графика, способ приписывания значений существенно занижает средневзвешенные значения показателя *Cited Half-life*. Тем не менее, полученные с помощью этого метода результаты достаточно хорошо отражают тенденцию изменения во времени этого показателя: с определенной степенью огрубления можно утверждать, что кривая, соответствующая этому способу на графике, эквидистантна кривой, соответствующей способу частичной виртуализации. Анализ графика, продемонстрированного на рис. 1, позволяет заключить, что наиболее точную картину о динамике показателя можно получить с помощью способа частичной виртуализации. Именно этим методом мы будем пользоваться в дальнейших исследованиях.

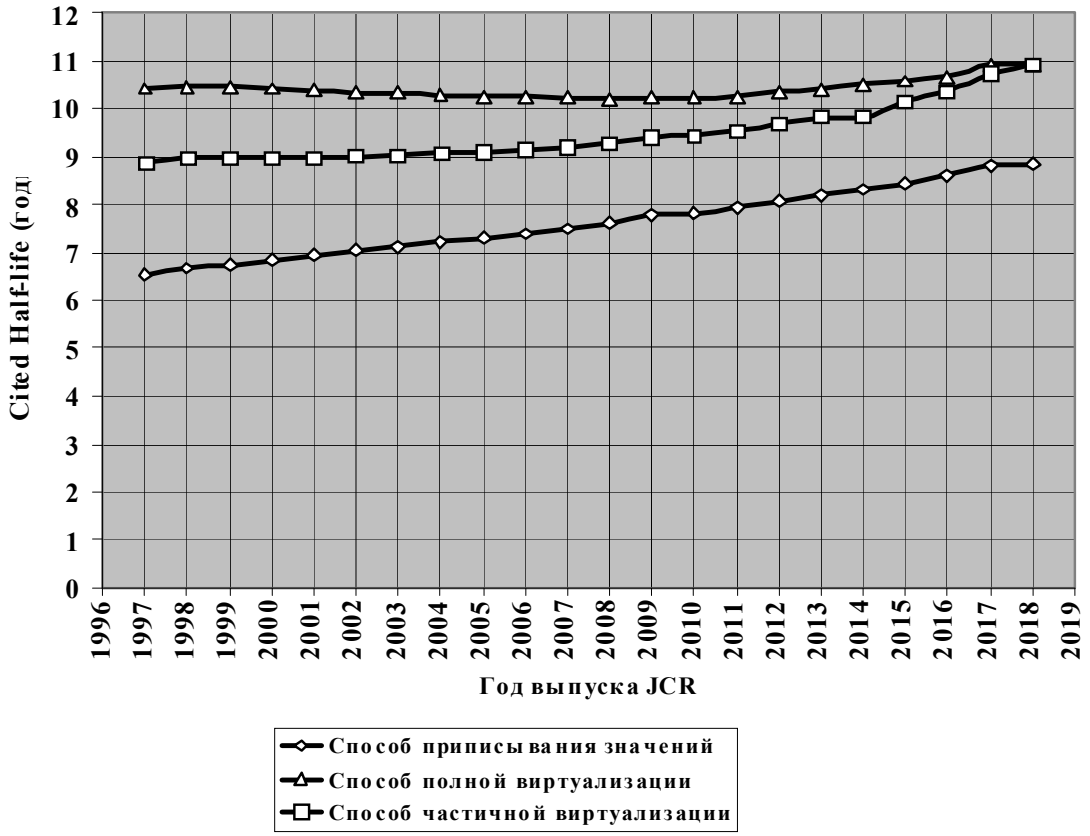

Рис. 1. Сопоставление значений показателя *Cited Half-Life*, полученных с помощью различных способов расчета (абсолютно сохраняющиеся журналы *JCR*)

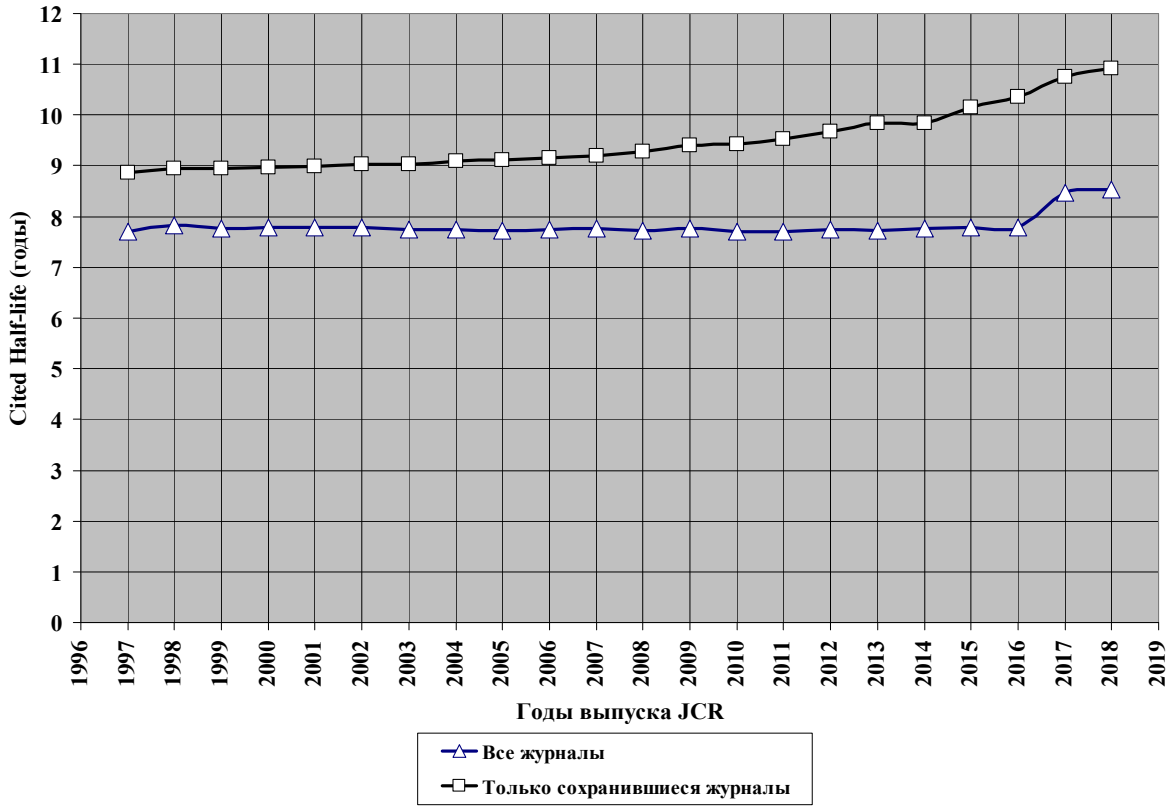

Рис. 2. Сравнение значений показателя *Cited Half-life* для массивов журналов *JCR* и их подмассивов, содержащих только абсолютно сохранившиеся журналы (результаты, получены с помощью способа частичной виртуализации)

Несколько важных выводов позволяют также сделать табл. 4 и соответствующий ей график, показанный на рис. 2. Средневзвешенные значения показателя *CdHL* для массивов, состоящих из «абсолютно сохраняющихся журналов», т.е. массивов, сформированных с учетом требований (*a*) и (*b*) всегда больше, чем соответствующие значения для массивов, сформированных только с учетом требования (*a*) – «все журналы». Кроме того, следует отметить, что изменения значений этого показателя в случае «абсолютно сохраняющихся журналов» происходят более динамично, чем соответствующие изменения в случае «все журналы». Более того, начиная с 1999 г. и до 2011 г. в случае «все журналы» динамика имеет очень слабую (едва заметную) отрицательную тенденцию. Затем тенденция медленно меняется на противоположную, и только в последние два года значение показателя начинает быстро расти.

ЗАКЛЮЧЕНИЕ

Уточнена ранее сформулированная нами гипотеза подобия распределений журналов по значениям показателей Cited Half-life Citing Half-life. На большом эмпирическом материале осуществлена проверка этой гипотезы, которая подтвердила ее справедливость. Установлено, что гипотеза выполняется для тех пар взаимно-однотипных распределений, которые принадлежат одному и тому же классу (подклассу) распределений. Установлено, что в пределах 8-10 лет гипотеза выполняется с большой точностью, однако и в интервале 15-20 лет ее использование для оценки динамики показателей *CdHL* и *CgHL* оказывается достаточно эффективным.

Скорректированы понятия основной части и хвоста распределения. Предложено и применено дополнительное деление этих частей распределения, что позволяет вычислять средневзвешенные значения показателей *CdHL и CgHL* с существенно большей точностью. Детально проработаны методы определения средневзвешенных значений соответствующих показателей. Разработаны средства, с помощью которых сформированы многие десятки классов взаимно-однотипных распределений, суммарно включающих тысячи распределений. Эти средства также позволяют реализовать предложенные методы на указанных распределениях.

Установлено, что для случаев «абсолютно сохранившихся журналов» значения показателя *CdHL* характеризуются достаточно ярко выраженной положительной динамикой, тогда как для случаев «все журналы» о положительной динамике можно говорить только в пределах временного отрезка, соответствующего последним трем годам.

Таким образом, созданы алгоритмы и методы вычисления, на основе которых работают программы, позволяющие наиболее точно определять период по-

лужизни научных журналов и их совокупностей, объединенных тематическими категориями *JCR*. В конечном счете это дает возможность сравнивать динамику развития областей науки, отраслей знания и отдельных научных дисциплин. Поскольку среди способов планирования и прогнозирования, применяемых в самых различных областях науки, самым традиционным и проверенным является экстраполяция тенденций развития, период полужизни научной литературы будет служить для этой цели важным подспорьем. Он также может применяться и при отслеживании долговечности идей выдающихся ученых [4]. Особенно важно, что предложенные нами способы впервые позволяют проследить эти тенденции по периоду полужизни не только цитируемых, но и цитирующих статей. Между этими показателями существует определенная, но пока еще не изученная корреляция, которая заслуживает отдельного анализа.

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. Либкинд А.Н., Маркусова В.А., Либкинд И.А. К вопросу определения динамики показателей периода полужизни журналов по *Journal Citation Reports* // Научно-техническая информация. Сер. 2. – 2020. – № 5. – С. 29-38; Libkind A.N., Markusova V.A., Libkind I.A. Approach for Using Journal Citation Reports in Determining the Dynamics of Half-Life Indicators of Journals //Automatic Documentation and Mathematical Linguistics. – 2020. – Vol. 54, № 3. – P. 174-184.
3. 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.
4. Розенберг Г.С. «Хиршивость» науки и период полураспада цитируемости научных идей // Биосфера. – 2018. – Т. 10, № 1. – С. 52-64.
5. Москалева О.В. Использование наукометрических показателей для оценки научной деятельности // Науковедческие исследования. – 2013. – С. 85-109.
6. Liang G., Hou H., Chen Q. et al. Diffusion and adoption: an explanatory model of “question mark” and “rising star” articles // Scientometrics. – 2020. – Vol. 124. – P. 219-232.

Материал поступил в редакцию 05.07.2020

Сведения об авторах

ГИЛЯРЕВСКИЙ Руджеро Сергеевич – доктор филологических наук, профессор, главный научный сотрудник, заведующий отделением ВИНИТИ РАН; профессор факультета журналистики Московского государственного университета им. М.В. Ломоносова
e-mail: ruggero29@gmail.com

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

БОГОРОВ Валентин Григорьевич – руководитель отдела образовательных программ *Clarivate Analytic*s, Москва
e-mail: valentin.bogorov@clarivate.com

ЛИБКИНД Илья Александрович – ведущий аналитик, ООО Сервисное бюро ВИП, Москва
e-mail: Libkind_Ilya@hotmail.com