УДК 378.1

В.Ю. Григорьев

Проблемы идентификации первичных данных при анализе регионального образования

В данной статье рассматривается проблема корректности статистических данных по образованию при комбинировании различных источников. Несоответствие географических названий помогло выявить неясности в предоставлении данных, особенно это относится к тем регионам, в составе которых присутствуют самостоятельные субъекты Российской Федерации. В статье приведены примеры расхождений информации по регионам, которые могут привести к недостоверности при дальнейшем анализе и прогнозировании образования. Также поднимается вопрос о регулярности сбора информации.

Ключевые слова: образование; информация об образовании; статистика образования; статистический анализ; нормализация данных.

Ключевой составляющей научной деятельности Центра экономики непрерывного образования Российской академии народного хозяйства и государственного управления при Президенте Российской Федерации (ЦЭНО РАНХиГС) является анализ сферы образования. При формировании моделей, применяемых для анализа различных образовательных уровней — дошкольного, общего, среднего профессионального и высшего образования — используется статистика, предоставляемая информационными базами Федеральной службы государственной статистики (Росстат). В ЦЭНО данные аккумулируются прежде всего с сайтов «Государственная статистика ЕМИСС» (URL: https://fedstat.ru/) и «Федеральная служба государственной статистики» (URL: http://www.gks.ru/), формируя единую базу данных, включающую показатели численности, имущества и финансирования, предоставляемые ведомством в открытом доступе.

Основная модель анализа сферы образования предполагает работу с данными в разрезе по годам, что создает определенные сложности при операциях с показателями в рамках информатизации соответствующих подходов [1]. Так, данные по численности населения доступны с 1991 г., тогда как в других показателях данные могут отсутствовать в довольно большом периоде. Например, показатель «Общая численность обучающихся в образовательных организациях, реализующих программы общего образования» ведется с 2004 г., а данные для показателя «Численность педагогических работников образовательных организаций, реализующих программы общего образования», доступны только с 2009 г. Таким образом, любые вычисления с использованием последних двух показателей могут быть осуществлены за период начиная от 2009 г.

Иногда в общей базе отсутствует статистика по одному или нескольким регионам за несколько лет. Например, показатель «Валовой региональный продукт в основных ценах» для Чеченской Республики не предоставлялся за 2004 г., как и общее значение показателя по округу до 2008 г. При этом данные по остальным регионам Южного федерального округа доступны в ЕМИСС.

Помимо описанной проблемы с предоставлением ежегодной статистики, показатели часто собираются неравномерно или разными департаментами. В качестве примера можно привести передачу сбора статистики по дошкольным образовательным организациям из Федеральной службы государственной статистики в Министерство образования России в 2014 г. — данные по ним доступны в разных разделах базы. Кроме необходимости в объединении их в один показатель, возникает также вопрос: используется ли для сбора одинаковый алгоритм или он был разработан заново? Это снижает степень достоверности предоставляемой статистики.

ЦЭНО использует деление по федеральным округам и более мелким регионам: краям, областям, республикам. Часто при этом формируется ниспадающая цепочка-иерархия: от более крупных регионов к более мелким [3]. Однако данные по регионам тоже могут быть представлены неравномерно. Аккумулирование открытой статистики выявило проблемы с идентификацией данных по регионам, используемых при анализе. Применяемая в ЦЭНО аналитическая модель подразумевает сбор данных отдельно по каждому из 85 субъектов Российской Федерации и их дальнейшее объединение по федеральным округам. Именно такая структура предоставления данных используется на сайте: URL: http://www.gks.ru/.

Однако далеко не всегда представление данных в источнике соответствует этому описанию. Так, помимо 8 федеральных округов в России существуют 12 экономических районов, которые также объединяют в себе несколько субъектов, но с несколько иными задачами. В данном случае интересно, что при объединении таких субъектов данные предоставляются как по каждому из них отдельно, так и суммарно по всему региону.

Так, например, по Восточно-Сибирскому экономическому району представлены данные за 1993 г, 1996 г., 1999–2001 гг., 2003–2004 гг. Доступны также аналогичные данные по входящим в него субъектам: Забайкальскому краю, Иркутской области, Красноярскому краю, Республике Бурятия, Республике Тыва, Республике Хакасия. Но сумма данных от этих субъектов не совпадает с итоговым значением в экономическом районе. Возникает вопрос: откуда взялось такое расхождение? К сожалению, база данных ЕМИСС не содержит информации о происхождении данных. Они могут быть предоставлены разными департаментами или организациями, которые используют разные алгоритмы для сбора и обобщения информации. Могло быть и так, что при суммировании показателей по экономическим районам какая-то часть данных была исключена по неизвестной причине. В любом случае это создает проблему верификации, а также ставит вопросы для последующего корректного и безопасного использования информации [4]. В случаях, когда нельзя получить данные напрямую из источника, придется пойти на компромисс и определить, какой показатель будет считаться истинным в ходе дальнейшего

анализа. При этом придется оговорить, что степень достоверности такого анализа будет зависеть от точности предоставляемой информации.

Описанный случай географической неопределенности и нестыковок показателей между различными регионами не является единичным. При использовании устоявшегося территориального деления «страна – федеральный округ – регион» также возникают сложности в определении субъектов, входящих в данный регион. Так, в Северо-Западном округе в составе Архангельской области выделяется Ненецкий автономный округ, для которого ведутся отдельные статистические показатели, присутствующие в общей базе данных. Аналогичная ситуация наблюдается и в других регионах, где существуют (или существовали) отдельные субъекты, входящие в их состав:

1. Тюменская область (Ханты-Мансийский автономный округ — Югра, Ямало-Ненецкий автономный округ).
2. Забайкальский край (Агинский Бурятский автономный округ).
3. Иркутская область (Усть-Ордынский Бурятский автономный округ).
4. Камчатский край (Корякский автономный округ).
5. Красноярский край (Таймырский (Долгано-Ненецкий) автономный округ).
6. Пермский край (Коми-Пермяцкий автономный округ).

На примере Тюменской области заметно, что идет сбор данных отдельно по области и по округам, что никак не обозначено в общем показателе по Тюменской области (табл. 1). То есть при формировании анализа по представленной статистике необходимо вводить виртуальные показатели «Тюменская область» и «Тюменская область без автономных округов», последний из которых представлен в статистике. Однако в показателях численности населения (например, «Численность постоянного населения — мужчин по возрасту на 1 января») существует отдельный показатель «Тюменская область (без АО)». Он не содержит данных, но его наличие ставит под вопрос достоверность имеющихся показателей округов и области, требуя подтвердить, входят ли данные по округам в общий показатель по области или они представлены отдельно.

Таблица 1

Общая численность обучающихся в образовательных организациях, реализующих программы общего образования (2016 г.)

Общая численность обучающихся в образовательных организациях, реализующих программы общего образования		**2016 г.**
Тюменская область	Всего	176 664
	Городские поселения	113 656
	Сельская местность	63 008
Ханты-Мансийский автономный округ — Югра (Тюменская область)	Всего	201 406
	Городские поселения	184 414
	Сельская местность	16 992
Ямало-Ненецкий автономный округ (Тюменская область)	Всего	71 145
	Городские поселения	57 266
	Сельская местность	13 879

Данная картина наблюдается во всех 7 регионах РФ, включающих самостоятельные субъекты, и может вызвать путаницу при формировании массива данных для статистического анализа. Сходную проблему можно увидеть в Мурманской области, где отдельно выделена статистика по населенным пунктам федерального подчинения, формирующим показатель «Сельские населенные пункты областного подчинения Мурманской области, находящиеся в ведении федеральных органов государственной власти и управления» (табл. 2).

Таблица 2

Общая численность обучающихся в образовательных организациях, реализующих программы общего образования (2004 г.)

Общая численность обучающихся в образовательных организациях, реализующих программы общего образования		**2004 г.**	**2005 г.**
Мурманская область	Всего	99 303	
	Городские поселения	92 256	
	Сельская местность	7 047	
Сельские населенные пункты областного подчинения Мурманской области, находящиеся в ведении федеральных органов государственной власти и управления	Всего		90 709
	Городские поселения		83 869
	Сельская местность		6 840

Описываемый пример характеризуется разницей в происхождении предоставляемых данных: в 2004 г. все они были отнесены к области, тогда как по населенным пунктам федерального подчинения (Александровск, Видяево, Заозерск, Островной, Североморск) информация отсутствует, а в 2005–2007 гг. наблюдается обратная ситуация. Основываясь на прослеживаемом тренде изменения данных, можно заключить, что оба показателя фактически являются одним и тем же показателем, но под разными названиями. Однако в таком случае возникает вопрос: если населенные пункты федерального подчинения выделены в отдельный показатель, то собираются ли данные по ним в отрыве от основного показателя? Включаются ли они в общий показатель по Мурманской области? При анализе региона в целом важно понимать, насколько целостны используемые данные и можно ли безболезненно исключить показатели по отдельным населенным пунктам из общего расчета. Та же ситуация наблюдается в Саратовской области, где существует показатель «Сельские населенные пункты, подчиненные Саратовской области, находящиеся в ведении федеральных органов государственной власти и управления».

Отдельно необходимо отметить отсутствие единого формата для представления географических наименований при формировании баз данных Росстата. В России официальные названия территориальных единиц содержатся в Государственном каталоге географических названий, который находится в ведении федерального бюджетного государственного учреждения «Федеральный

научно-технический центр геодезии, картографии и инфраструктуры пространственных данных». Создание и обновление данных каталога географических названий проводится в рамках исполнения Федерального закона от 18.12.1997 № 152-ФЗ «О наименованиях географических объектов» и соответствующего приказа Минэкономразвития России от 27.03.2014 № 172 «Об утверждении Порядка регистрации и учета наименований географических объектов, издания словарей и справочников наименований географических объектов, а также выполнения работ по созданию Государственного каталога географических названий». Реестры географических наименований по регионам доступны на странице Федеральной службы государственной регистрации, кадастра и картографии (Росреестр)[1].

Несмотря на существование такого каталога и служб, фактически используемые географические наименования в статистике сферы образования представлены разными форматами, имеет место использование сокращений наименований регионов и субъектов, не соответствующих каталогу. Однако основная проблема заключается не только в этом, а еще и в том, что эти наименования различаются в разных базах. То есть, при совместном анализе данных сайтов — URL: http://www.fedstat.ru и URL: http://www.gks.ru — необходимо разработать отдельную процедуру нормализации данных. Если такая процедура не реализована, при анализе массива возникают повторы, часть данных не объединяется по принадлежности к одному региону, и сформировать достоверные итоговые показатели невозможно. В итоге исследовательская группа может либо использовать данные только одной из баз при условии, что внутри нее не требуется нормализация, либо следует добавить в анализ дополнительный шаг по обработке данных. Сложность в его реализации заключается в отсутствии доступа к статистическим данным в удобном для их анализа программными средствами формате. Необходимо выгружать их с сайтов и загружать в инструментарий для анализа, а ведь в рамках небольших исследовательских проектов эта процедура часто осуществляется вручную.

Обозначенные выше проблемы существенно влияют на построение аналитических моделей для оценки и прогнозирования процессов в сфере образования, формирования системных подходов к построению информационных образовательных сред [2]. Высокий риск использования недостоверных данных или неверного определения показателей региона может привести к ложному результату, что поставит под сомнение все исследование. Уменьшить его можно только через нормализацию всех используемых данных.

Определенные сложности возникают при разработке процедуры нормализации. Представленные выше примеры демонстрируют, что при всей схожести

---

[1] Государственный каталог географических названий — Росреестр. URL: https://rosreestr.ru/site/activity/geodeziya-i-kartografiya/naimenovaniya-geograficheskikh-obektov/gosudarstvennyy-katalog-geograficheskikh-nazvaniy/ (дата обращения: 10.02.2018).

рассмотренных проблем для каждой из них потребуется собственное решение. Частично может потребоваться ручная очистка данных от «мусора» в виде пустых полей, лишних пробелов или отступов. При этом увеличивается срок обработки статистики и риск ошибки.

Компромиссным вариантом можно назвать унификацию самих баз данных первичной статистической информации, приведение собираемых показателей в общий вид, совпадающий если не с официальными источниками, как в случае с каталогом географических наименований, то хотя бы с другими статистическими базами. Однако этот процесс невозможен без тесного взаимодействия департаментов, отвечающих за подготовку и размещение данных в Росстате и Министерстве образования, а налаживание этого взаимодействия может оказаться длительным и дорогостоящим процессом. Поскольку в ближайшей перспективе никаких распоряжений по слиянию баз данных не предполагается принять, то такой вариант вообще маловероятен.

Литература

1. *Григорьев В.Ю.* Подходы к определению роли информатизации в системе показателей качества высшего образования // Вестник Российского университета дружбы народов. Серия «Информатизация образования». 2017. Т. 14. № 4. С. 418–429.

2. *Григорьев В.Ю.* Системный подход к формированию многофункциональной информационно-образовательной среды юридического вуза: постановка проблемы // Известия высших учебных заведений. Правоведение. 2005. № 3 (260). С. 204–213.

3. *Гриншкун В.В.* Теория и методика использования иерархических структур в информатизации образования // Информатика и образование. 2003. № 12. С. 117–119.

4. *Гриншкун В.В., Димов Е.Д.* Принципы отбора содержания для обучения студентов вузов технологиям защиты информации в условиях фундаментализации образования // Вестник Российского университета дружбы народов. Серия «Информатизация образования». 2012. № 3. С. 38–45.

Literatura

1. *Grigor'ev V.Yu.* Podxody' k opredeleniyu roli informatizacii v sisteme pokazatelej kachestva vy'sshego obrazovaniya // Vestnik Rossijskogo universiteta druzhby' narodov. Seriya «Informatizaciya obrazovaniya». 2017. T. 14. № 4. S. 418–429.

2. *Grigor'ev V.Yu.* Sistemny'j podxod k formirovaniyu mnogofunkcional'noj informacionno-obrazovatel'noj sredy' yuridicheskogo vuza: postanovka problemy' // Izvestiya vy'sshix uchebny'x zavedenij. Pravovedenie. 2005. № 3 (260). S. 204–213.

3. *Grinshkun V.V.* Teoriya i metodika ispol'zovaniya ierarxicheskix struktur v informatizacii obrazovaniya // Informatika i obrazovanie. 2003. № 12. S. 117–119.

4. *Grinshkun V.V., Dimov E.D.* Principy' otbora soderzhaniya dlya obucheniya studentov vuzov texnologiyam zashhity' informacii v usloviyax fundamentalizacii obrazovaniya // Vestnik Rossijskogo universiteta druzhby' narodov. Seriya «Informatizaciya obrazovaniya». 2012. № 3. S. 38–45.

V.Yu. Grigoryev

The Problems of Identifying Primary Data in the Analysis of Regional Education

In this paper, the problem of the correctness of statistical data on the education obtained by combining different sources is considered. The mismatch of geographical names helped to clarify the uncertainties in the provision of data, especially in those regions in which independent subjects of the Russian Federation are present. The article gives examples of discrepancies in information by region, which can lead to inaccuracy in the further analysis and forecasting of education. The issue of the regularity of information gathering is also raised.

Keywords: education; information about education; education statistics; statistical analysis; data normalization.