Методы извлечения данных из веб-страниц для сбора информации о ценах на товары потребительской электроники и на авиаперевозки для построения ГИПЦ в Италии

Федерико Полидоро (Federico Polidoro[*]), Рикардо Джаннини (Riccardo Giannini), Розанна Ло Конте (Rosanna Lo Conte), Стефано Моска (Stefano Mosca) и Франческа Розетти (Francesca Rossetti)[1]
ИСТАТ, Итальянский институт статистики, Рим Италия

Перевод: Статкомитет СНГ

Аннотация. Эта работа посвящена результатам тестирования методов извлечения данных из веб-страниц для целей обследования потребительских цен по двум продуктам: потребительская электроника (товары) и авиаперевозки (услуги). Эта статья опирается на работу, проводимую ИСТАТом в рамках европейского проекта «Многоцелевая статистика цен» (МСЦ). Среди многих тем, охваченных МСЦ, присутствуют модернизация сбора данных и использование извлечения данных из веб-страниц. Рассмотрены также вопросы качества (в смысле эффективности и уменьшения ошибок) и приведены некоторые предварительные замечания относительно использования «больших данных» для статистических целей. Общие цели работы указаны во Вступлении (раздел 1). В разделе 2 объясняется выбор продуктов для тестирования сбора данных методом извлечения информации из веб-страниц. В разделах 3 и 4 после описания обследования для таких позиций, как потребительская электроника и авиаперевозки, приводятся и обсуждаются результаты/вопросы тестирования методов извлечения данных из интернета. В разделе 5 отмечены возможности для улучшения качества данных для измерения инфляции в результате использования извлечения данных из веб-страниц. В Заключении (раздел 6) внимание уделено вопросам использования «больших данных». В двух рамках приведена информация о централизованном сборе данных о потребительских ценах в Италии и о технологических решениях для извлечения данных из веб-страниц.

Ключевые слова: многоцелевая статистика цен, извлечение данных из веб-страниц, интернет как источник данных

---

[*] Автор-корреспондент: Федерико Полидоро (Federico Polidoro), ИСТАТ, Итальянский статистичекий институт, Рим, Италия. Тел.: +39 0646734157; Fax: +39 0646734173; E-mail: polidoro@istat.it

[1] Эта статья представляет результаты работы пяти авторов, в частности, Рикадро Джаннини, является автором текста в п. 2 и в Рамке 1; Стефано Моска – автор пп.4.1. и 4.2; Франческа Розетти – автор п.3; Федерико Полидоро – автор аннотации, пп.1, 5 , 6, п.4.3 и супервайзер статьи.

1. Вступление

Модернизация инструментария сбора данных для улучшения качества гармонизированного индекса потребительских цен является одной из важнейших составляющих проекта «Многоцелевая статистика цен», запущенного Европейской комиссией. Модернизация сбора данных включает три основных элемента: более широкое использование электронных устройств для сбора данных о ценах, использование данных сканирования как источника информации для оценки инфляции и расширение извлечения данных из веб-страниц для получения информации из интернета для целей составления гармонизированного индекса потребительских цен.

Национальный статистический институт Италии (ИСТАТ) активно работает по всем трем направлениям модернизации сбора данных. В частности, команда статистиков и специалистов по информационным технологиям протестировала методы извлечения данных из веб-страниц для обследования потребительских цен, уделяя особое внимание двум группам продуктов: потребительская электроника (товары) и авиаперевозки (услуги). Процедуры для извлечения данных из веб-страниц были разработаны и протестированы для обоих этих видов продуктов.

Тестирование было проведено в рамках текущей ситуации с обследованием потребительских цен в Италии, где данные частично обираются централизованно ИСТАТом (более 21% корзины товаров в смысле весов), а частично – из интернета.

Таким образом цель этих новых усилий была двоякой: с одной стороны, понимание качества в смысле сокращения ошибок измерений и в смысле эффективности и затрат обследования потребительских цен; с другой стороны, исследование и анализ проблем, возникающих при извлечении информации из веб-страниц, привели к некоторым мыслям и замечаниям по поводу использования этих методов для доступа к «большим данным» в интернете для измерения инфляции.

Несмотря на то, что работа выполнялась в течение ограниченного времени, это позволило нам, как будет видно дальше, достичь важных, хотя и не окончательных результатов по повышению эффективности. Гораздо больше еще необходимо сделать. Например, мы только начали анализировать возможные улучшения качества обследования, сделав некоторые предварительные замечания об использовании «больших данных» и о последствиях этого для дизайна обследования (Scheveningen Memorandum, 2013).

2. Определение продуктов для тестирования сбора данных методом извлечения информации из веб-страниц

В рамках широкого списка продуктов, потребительские цены на которые централизованно собирает ИСТАТ, специалисты ИСТАТА собрали информацию о ценах в Интернете (частично для групп B и D и полностью для группы C) (Рамка 1). Для этих продуктов – до начала проекта МСЦ (MPS) - сбор данных в сети проводился, в основном, в ручном режиме с использованием метода «копирования и вставки».

При постановке задач для разработки, тестирования и внедрения методов извлечения информации из веб-страниц в рамках европейских проектов сначала следовало выбрать продукты или группы продуктов:

i. которые представляли бы и товары, и услуги;

ii. для которых интернет является очень важных каналом розничной торговли;

iii. для которых этап сбора данных занимает продолжительное время;

iv. для которых важно расширение охвата выборки, как во времени, так и в пространстве, и преодоление ограничений, связанных с ручным сбором данных.

Что касается критерия (ii), то в соответствии с данными обследования ИСТАТА «Аспекты повседневной жизни», которое поставляет информацию о поведении домохозяйств и соответствующих аспектах их повседневной жизни, в 2012 году доля домохозяйств, владеющих персональными компьютерами, составляла 59,3% (в 2013 г. она увеличилась до 62,8%), а доля домохозяйств, имеющих выход в Интернет, составляла 55.5% (60.7% в 2013 г.). Ориентировочный расчет объемов электронной коммерции был проведен на основании данных о том, что в 2012 году 28,2% лиц в возрасте 14 лет и старше использовали Интернет в течение последних 12 месяцев и покупали или заказывали товары или услуги для личного пользования через Интернет (в 2011 г. их было 26,3%).

Рамка 1: ИСТАТ: централизованный сбор данных о потребительских ценах

В 2013 и 2014 гг. централизованный сбор данных о потребительских ценах для расчета ГИПЦ проводился для более чем 21% (по весу) корзины продуктов. Он проводился ИСТАТом и был разделен на четыре основные группы:

A. Приобретение полных внешних баз данных (лекарства, школьные учебники, взносы домохозяйств в Национальную службу здравоохранения) Эта первая группа составляет примерно 0,6% от корзины.

B. Централизованный сбор данных, так как это наиболее эффективный путь для сбора данных, необходимых для построения индексов. Эта вторая группа составляет примерно 11,6% от корзины и включает:

i. Объявленные цены, которые могут отличаться от фактических цен покупки (например, туристические путевки);

ii. Фактические цены покупки для приобретения товаров он-лайн и в реальных магазинах (например, оплата он-лайн абонентской платы за ТВ);

iii. Фактические цены покупки на товары, которые потребители не могут приобрести через интернет (например, плата за паспорт, платные дороги).

C. Приобретение цен, касающихся реальных покупок в интернете. Эта группа составляет примерно 2,3% от корзины и включает:

i. Фактические цены покупок, собранные путем моделирования покупок в интернете (например, авиабилеты, потребительская электроника и электронные книги);

ii. Фактические цены покупок, собранные путем моделирования покупок в интернете + объявленные цены (например, тарифы на морской транспорт).

D. Другие цены, собираемые централизованно. Эта последняя группа составляет около 7% корзины, включая:

i. Единые цены на всей территории Италии (например, табак, сигареты);

ii. Данные, собранные из разных источников, таких как журналы, перечни цен в интернете, информация, полученная по электронной почте (например, цены на автомобили, тарифы на региональных железнодорожный транспорт);

iii. Данные из других обследований ИСТАТа и используемые в качестве замещающих переменных для фактических цен (например, почасовая оплата по контракту как замена фактической зарплате).

В Таблице 1 показано (в процентах) ранжирование групп продуктов, которые были куплены или заказаны через Интернет лицами в возрасте 14 лет и

старше, использовавшими интернет в течение последних 12 месяцев и покупавшими или заказывавшими товары и услуги для личного пользования в Интернете. Основными областями, в которых совершались покупки через интернет в 2012 году, были: бронирование проживания во время отпуска и другие товары и услуги для путешествий, тогда как товары потребительской электроники занимали в списке только шестое место[2].

Таблица 1

Электронная торговля. Лица в возрасте 14 лет и старше, которые пользовались интернетом за последние 12 месяцев и которые купили или заказали товары и услуги для личного пользования через интернет (по группам купленных или заказанных продуктов).
2012, *Проценты*

Ночевки на каникулах (отели, пансионы)	35.5
Другие расходы на путешествия (ж/д и авиа билеты, и пр.)	33.5
Одежда и обувь	28.9
Книги, газеты, журналы, включая электронные книги	25.1
Билеты на представления и шоу	19.7
Товары потребительской электроники	18.6
Предметы для дома, мебель, игрушки	17.9
Другое	15.1
Фильмы, музыка	14.4
Телекоммуникационные услуги	14.0
ПО для компьютеров и их обновления (кроме видеоигр)	11.5
Оборудование для компьютеров	8.4
Видеоигры и их обновления	8.0
Финансовые и страховые услуги	6.0
Продукты питания	5.6
Материалы для электронного обучения	2.8
Азартные игры	1.2
Лекарства	0.8

Источник: Обследование ИСТАТа «Аспекты ежедневной жизни»

Принимая во внимание критерии (iii) и (iv), а также имея в виду, что было бы предпочтительно протестировать методы извлечения данных из интернета для тех продуктов, по которым сбор данных уже проводится централизованно или через интернет, были, наконец, выбраны две группы продуктов: потребительская электроника (товары) и цены на авиаперевозки (услуги).

[2] Рост покупок в интернете для некоторых продуктов, например, одежды и обуви, книг, газет, журналов, включая электронные книги, билетов на представления, товаров для дома, мебели, игрушек, говорит о том, что интернет играет все большую роль как канал розничной торговли для продуктов, по которым сбор данных проводится в поле или по ссылкам не на цены, предлагаемые в интернете (в 2013 г. для Одежды и обуви процент в таблице составлял 31.5%, на 2.6 процентных пункта больше, чем в 2012).

3. Проверка и внедрение методов извлечения данных из интернета для «потребительской электроники»

3.1. Обследование цен на потребительскую электронику и применение методов извлечения данных из интернета

Набор товаров потребительской электроники, для которых ИСТАТ регулярно собирает цены, состоит из следующих позиций: a) мобильные телефоны, b) смартфоны, c) ноутбуки, d) настольные ПК, e) планшеты, f) мониторы, g) принтеры i) беспроводные или проводные телефоны, l) цифровые фотоаппараты, m) видео камеры.

Дизайн обследования одинаков для всех перечисленных выше продуктов, и его можно описать следующим образом:

Этап 1. Выбор брендов и магазинов (ежегодно); около 18 магазинов (в среднем) для каждого продукта на национальном уровне.

Этап 2. Сегментация рынка на основе технических спецификаций и производительности (фиксируется ежегодно).

Пример 1 – цифровые фотоаппараты: seg1= 'компактный' фотоаппарат; seg2= 'псевдозеркальный фотоаппарат; seg3= беззеркальный фотоаппарат; seg4= 'зеркальный фотоаппарат

- Пример 2 – PC мониторы: seg1 =размер экрана 19-20 дюймов; seg2= размер экрана 21-22 дюймов;
- Пример 3 – мобильные телефоны: seg1 =мобильные телефоны с простыми функциями; seg2= мобильные телефоны со сложными функциями;
- Пример 4 – Настольный ПК: seg1 = настольный ПК; seg2= все в одном;

Этап 3. Определение минимальных требований (фиксируется ежегодно).

- Пример 1 – Настольный ПК: ОС как минимум Windows 7, HD – 160 Gb или больше, RAM как минимум 2 Gb, и пр.

Этап 4. Ежемесячный сбор данных по всей линейке моделей в смысле коммерческих названий и основных технических характеристик, предлагаемых на рынке основными брендами в рамках сегментов, определенных на этапе 2, и удовлетворяющих минимальным требованиям, определенным на этапе 3 (ежемесячные наблюдения). На этапе 4 формируется выборка для конкретного месяца (постоянно обновляемая выборка с автоматической заменой моделей, важность которых на рынке снижается).

Пример осуществления этапа 4 может быть рассмотрен в отношении планшетов. Для проведения эффективной сегментации для построения индекса, собирают информацию об основных характеристиках планшетов, предлагаемых ведущими компаниями на итальянском рынке. Для каждой новой модели сообщаются следующие характеристики: спецификации экрана, память, операционная система, CPU, подключение, GPS, трансформатор (Табл. 2)

Таблица 2

Пример результатов Этапа 4 обследования, касающегося потребительских цен на планшеты

Код	Бренд	Тип	Память	ОС	Cpu	Подклю-чение	Gps	Экран	Транс-форм
T_Ace029	Acer	ICONIA A211 – HT.HA8ET.001	16	Android	nVidia Tegra T30L Quad-core	3G	1	10.1	0
T_Ace041	Acer	ICONIA A211 – HT.HADET.001	16	Android Ice Crea	nVidia Tegra T30L Quad-core	3G	1	10.1	0
T_Ace035	Acer	ICONIAW511-27602G0iss HT.HA4EE.006	32	Android Ice Crea	nVidia Tegra T30S Quad-core	3G	1	10.1	0
T_Ace037	Acer	ICONIAW511-	64	Windows	Atom ™ Z2760	3G	0	10.0	1

		27602G06iss NT.LONET.004		8 Pro	(1.80 GHz Intel®) Burs				
T_Ace036	Acer	ICONIAW511-27602G06iss NT.LONET.004	64	Windows 8 Pro	Atom ™ Z2760 (1.80 GHz Intel®) Burs	3G	0	10.0	1
T_Ace045	Acer	Iconica W511-27602G06iss NT.LONET.004	64	Windows 8 P -	Intel® Atom ™ Z2760 (1 MB Cache,1)	3G	0	10.0	1
T_Ace032	Acer	Ipad display retine 16 gb wi fi +cellular	16	Mac: OS X v10.6	A6X dual-core	+c	1	9.7	0
T_Ace033	Acer	Ipad display retine 32 gb wi fi +cellular	32	Mac: OS X v10.6	A6X dual-core	+c	1	9.7	0
T_Ace034	Acer	Ipad display retine 64 gb wi fi +cellular	64	Mac: OS X v10.6	A6X dual-core	+c	1	9.7	0

Таблица 3

Первоначальная нагрузка для разработки макросов для извлечения данных из веб-страниц (указывать веб-сайты и извлекать данные)

Вебсайты магазинов	Кол-во продуктов	Кол-во указа-тельных макросов	Всего время (мин) для указа-тельных макросов	Кол-во макросов для извлечения данных	Время на разработку первого макроса для извлечения данных	Время для следующих макросов (включая тестирование)	Время на оптими-зацию макросов
www.compushop.it	10	12	5 x 12 = 60	24	60	5 x 23 = 115	-
www.ekey.it	6	11	5 *x* 11 = 55	22	120	5 *x* 21 = 105	-
www.keyteckpoint.it	9	16	5 *x* 16 = 90	16	30	5 *x* 15 = 75	-
www.misco.it	10	11	5 *x* 11 = 55	22	45	5 *x* 21 = 105	-
www.pmistore.it	7	10	5 *x* 10 = 50	10	30	5 *x* 9 = 45	-
www.softprice.it	10	22	5 *x* 22 = 110	46	45	5 *x* 45 = 225	-
www.syspack.it	8	14	5 *x* 14 = 70	14	30	5 *x* 13 = 65	-
Время всего		8 час		6 час	12 час	8 час	

Этап 5. На этом этапе собирают данные о ценах для всех моделей, включенных в выборку, с каждого вебсайта магазинов, рассматриваемых в обследовании (ежемесячные наблюдения). До начала эксперимента и внедрения методов по извлечению данных из интернета в рамках проекта Евростата, сбор данных осуществлялся двумя путями:

- регистрация вручную – для нескольких магазинов (девяти) сборщики цен вручную просматривали соответствующие вебсайты и регистрировали цены во внешних файлах и базах данных.

- полуавтоматическая регистрация – еще для девяти магазинов, списки цен были загружены вручную (копирование и вставка), а затем отформатированы и подвергнуты процедурам SAS, которые увязывали (автоматически) коды продуктов в выборке (этап 4) с кодами в списках из каждого магазина.

Этап 6. Настройка базы данных для расчета индексов потребительских цен и объединения данных, зарегистрированных вручную и в полуавтоматическом режиме.

Этап 7. Анализ репрезентативности каждой модели и контроль выбивающихся значений (для того, чтобы участвовать в расчете индекса, каждая модель должна иметь минимальное количество элементарных котировок, и каждый сегмент должен быть представлен минимальным числом продуктов).

Этап 8. Расчет средней цены для каждой модели - среднее геометрическое или медиана (когда имеется немного наблюдений, даже хотя есть требуемый минимум, используется медиана).

Этап 9. Каждая страта (сегмент/бренд) представлен самой дешевой моделью, так что используется минимальная цена для представления страты и для построения микро-индексов.

Этап 10. Агрегирование микро-индексов с использованием среднего геометрического (элементарный уровень) и средневзвешенной арифметической (верхние уровни). Веса (где имеются) пропорциональны доле рынка для каждого бренда и каждого сегмента.

Этан 4 и Этап 5 требуют больше всего времени для своего осуществления. В начале проекта было решено обратить особое внимание на эти два этапа и, в особенности, на этап 5 и на полуавтоматическую регистрацию цен, для которой оказалось проще использовать методы извлечения данных из интернета. На самом деле для этих цен целью макроса для извлечения данных из интернета было заменить выгрузку списков цен, проводившуюся вручную путем копирования и вставки, на автоматическую выгрузку (извлеченный из интернета список цен). Поэтому оценка полученных результатов учитывала и количество выгруженных котировок цен в списках, и количество котировок цен, которые можно было автоматически связать для каждого магазина (через процедуры SAS) с кодами продуктов в выборке, отобранными на Этапе 4. Проверка и осуществление процедур извлечения данных из интернета для замены ручной регистрации цен (имеется в виду, что для некоторых магазинов сборщики цен вручную просматривают вебсайты и заносят каждую цену во внешние файлы или базы данных) поставили вопросы, которые в соответствии с предварительным анализом оказались слишком сложными для решения на данном этапе этого проекта.

Таким образом, интернет-магазины, выбранные для анализа, были теми девятью магазинами, для которых использовалась полуавтоматическая процедура регистрации цен, а эксперимент с регистрацией цен при помощи процедуры извлечения данных из интернета проводился с использованием бесплатной версии программного обеспечения iMacros (ПО для разработки процедур излечения данных из интернета – см. Рамку 2).

Рамка 2: ПО для осуществления процедур извлечения данных из интернета

Программа автоматически извлекает информацию из веб-страниц и ее делает узнаваемой, записывает ее в локальные базы данных / хранилища данных / файлы, устраняя использование метода «копирования и вставки».

Прежде чем выбрать программное обеспечение для тестирования были рассмотрены разные инструменты и программы. Внимание, в основном, уделялось HTQL, IRobotSoft, iMacros. HTQL (Hyper-Text Query Language) в основном используется для извлечения HTML контента из веб-страниц, построения структур таблиц на основе веб-страниц или автоматической модификации веб-страниц. IRobotSoft - средство визуальной автоматизации и извлечения данных из веб-страниц с применением HTQL. Это самое подходящее ПО для исследователей рынка, которым необходимо часто собирать данные из открытых источников, например, для риелторов, собирающих информацию о рынке жилья, или исследователей, постоянно отслеживающих определенные темы в сети. Окончательный выбор был сделан в пользу iMacros, и обоснованием данного выбора может служить следующее: iMacros – это программное решение для веб-автоматизации и веб-тестирования. iMacros позволяет пользователям многократно воспроизводить записанные действия, такие как: тестирование веб-форм, загрузку и скачивание изображений и осуществлять импорт-экспорт данных из веб-приложений с помощью файлов CSV & XML, баз данных и любых других источников.

При помощи этого продукта можно ускорить приобретение текстовой информации в сети; его можно использовать с языками программирования и скриптами (например, Java, JavaScript). Задачи iMacros могут быть выполнены в рамках наиболее популярных браузеров. Описание продукта приведено на http://wiki.imacros.net/iMacros_for_Firefox и на форумах (например, http://forum.iopus.com/viewforum.php), где приводятся примеры кодов и задачи, которые уже рассматривались и решались другими, что поможет ускорить разработку макроса. Можно воспользоваться результатами других проектов (например, http://sourceforge.net/projects/jacob-project/) с использованием языка Java, что предоставляет пользователям большие возможности для интерфейса и интеграции с другими программными продуктами и средами.

Принятый подход состоял в использовании двух разных макросов – указательного и извлекающего. Указательный макрос применяется с целью нахождения страницы, на которой имеются собираемые данные. Обычно этот макрос просто построить, и сборщик данных может справиться с этим самостоятельно. Макрос для извлечения данных выполняет реальную работу по сбору данных и записи их в плоский файл.

Этот подход продемонстрировал хорошие результаты с некоторыми важными преимуществами, но и с определенными недостатками. Основное преимущество состоит в простоте обслуживания. Во всех случаях, когда возникают проблемы с указательным макросом, нет необходимости в привлечении программиста, поскольку сборщик данных может легко восстановить указательный макрос самостоятельно. Это очень важное преимущество, так как позволяет легко справляться с проблемой нестабильности веб-сайтов (но не с проблемой доступа, как обсуждается в пункте 4.3). Основной недостаток состоит в относительно низкой степени используемости, так как сборщики данных вынуждены применять два макроса вместо одного. Подход с использованием указательного и извлекающего макросов был применен для теста с ценами на потребительскую электронику. Для тех интернет магазинов, где проводился эксперимент с извлечением данных из веб-страниц, недостатки были с лихвой компенсированы выгодами и улучшениями.

3.2. Улучшение охвата, точности и эффективности и принятие процедуры извлечения данных из веб-страниц в рамках текущего обследования цен на потребительскую электронику

В Табл. №№ 3, 4 ,5 показаны результаты, и количественная оценка экономии времени. Следует отметить, что в таблицах 3 и 4 интернет-магазины, для которых используются макросы для извлечения данных из веб-страниц, сокращены с девяти в начале эксперимента до шести для обычного оборота единиц для сбора данных; с января 2014 г. количество интернет-магазинов, для которых применяются процедуры извлечения данных из веб-страниц, равно семи (для этих магазинов см. данные в Табл.6).

В Табл. 3 приведена оценка времени для разработки макросов для извлечения данных из веб-страниц (всего 34 часа). Можно считать, что это время необходимо для внедрения макросов для ежегодно меняющейся базы, когда также выборка единиц по сбору данных пересматривается (а затем также выборка интернет магазинов).

В Табл. 4 показано сравнение между текущей (месячной) нагрузкой при полуавтоматической регистрации цен («копирование и вставка») и при извлечении данных из веб-страниц.

Наконец в Табл. 5 приведено сравнение нагрузки при использовании методов полуавтоматической регистрации цен и при извлечении данных из веб-страниц. Преимущества, вызванные использованием методов извлечения данных из веб-страниц, для выбранных (шести) магазинов очевидны, и их можно суммировать следующим образом: на годовой основе затраты времени,

необходимые для проведения обследования, сокращаются с 23 рабочих дней до 16 рабочих дней. Это означает, что использование методов извлечения данных из веб-страниц для этой подвыборки магазинов сберегает более 30% времени и может позволить увеличить количество котировок цен, которые можно использовать для расчета индексов путем автоматической увязки при помощи процедур SAS.

Таблица 4

Текущая нагрузка по сбору цен. Сравнение между нагрузкой при полуавтоматическом регистрировании и при извлечении данных из веб-страниц

Веб-сайты интернет-магазинов	Количество товаров	Полуавтоматическая регистрация: навигация, копирование и вставка	Полуавтоматическая регистрация: стандартизация формата (мин)	iMacros загрузка: выполнение макроса (мин)	iMacros загрузка: форматирование результатов (мин)
www.compushop.it	10	50	80	15	50
www.ekey.it	6	30	20	15	70
www.misco.it	10	60	90	10	45
www.pmistore.it	7	40	90	15	20
www.softprice.it	10	90	180	25	40
www.syspack.it	8	45	90	20	45
Время, всего		5 час 15 мин	9 час 10 мин	1 час 40 мин	4 час 30 мин

Источник: ИСТАТ

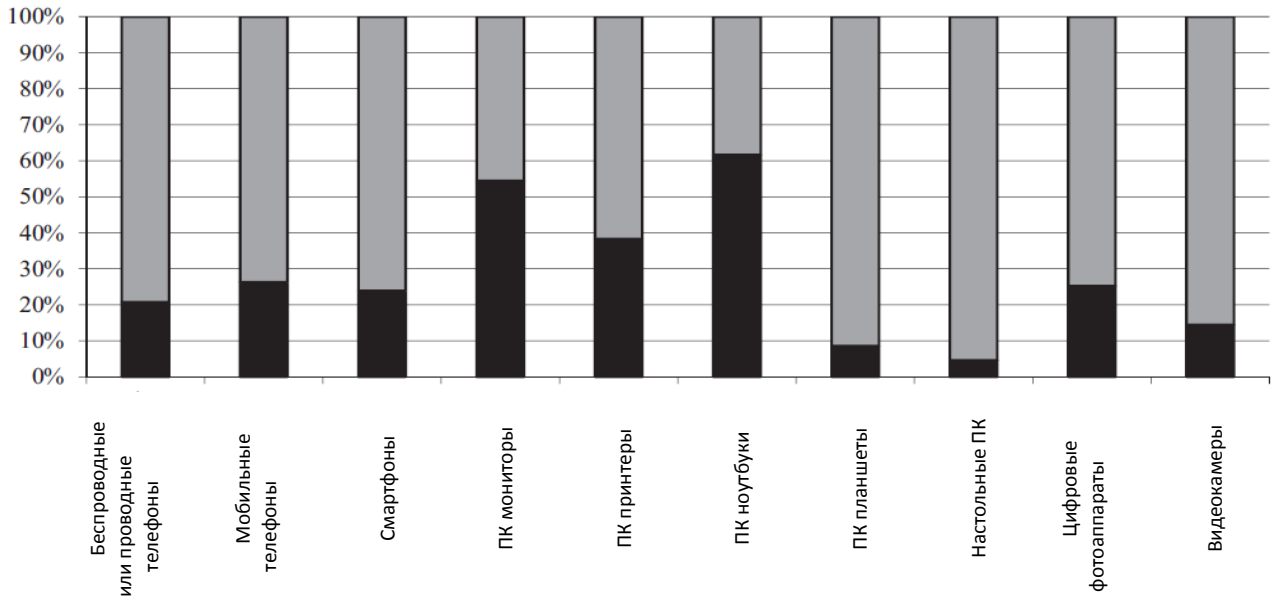

Рис. 1. Доля цен, собранных при помощи извлечения данных из веб-страниц, в текущем месячном обследовании цен на потребительскую электронику, в среднем, начиная с марта 2013

Полученные результаты побудили ИСТАТ, начиная с марта 2013 г. принять процедуры по извлечению данных из веб-страниц для всех магазинов, где это возможно. Навигация по этим сайтам и сбор информации (номер модели, описание бренда и цена) автоматически регистрируются с использованием макросов, построенных при помощи iMacros для интернет магазинов, для которых ранее были приняты методы полуавтоматической регистрации.

Итак, начиная с марта 2013 г. обследование товаров потребительской электроники проводится двумя способами: (i) ручная регистрация, (ii) скачивание с использованием извлечения данных из веб-страниц. Скачивание с использованием методов извлечения данных из веб-страниц заменило методы полуавтоматической регистрации (Рис. 1).

Таблица 5

Годовые затраты времени на половину магазинов, отобранных для сбора данных о ценах на товары потребительской электроники. Сравнение между полуавтоматической регистрацией и извлечением данных из веб-страниц (час)

	Ручная регистрация	Извлечение данных из веб-страниц
Начальная нагрузка (годовая изменяющаяся база)	0	34
Текущее ведение	0	12
Текущий сбор данных	173	74
Всего рабочих часов	173	120

Источник: ИСТАТ

В Табл. 6 для января показана ситуация с обследованием в 2014 г для разных товаров потребительской электроники, для которых в настоящее время используются методы извлечения данных о ценах (семь интернет магазинов против девяти магазинов с использованием ручной регистрации). В первом столбце указано количество моделей, отобранных в выборку на этапе 4. Во втором столбце указано количество элементарных котировок цен, извлеченных из веб-страниц. В третьем столбце указано количество элементарных котировок цен, которые было возможно увязать с кодами моделей, отобранных на этапе 4; в последнем столбце приведены проценты извлеченных из веб-страниц и увязанных цен (действительно, подходящих для построения индекса).

Таблица 6

Выборка моделей, котировки цен на товары потребительской электроники, извлеченные из веб-страниц, для расчета ГИПЦ в Италии. Январь 2014. *Единицы и проценты*

Обследование	Количество моделей в выборке	Количество котировок цен, извлеченных из веб-страниц	Количество котировок цен, собранных и увязанных с выборкой	Котировки цен увязанные/извлеченные (%)
Беспроводные или проводные телефоны	190	844	224	26,5
Мобильные телефоны	63	2024	108	5,3
Смартфоны	131	2396	187	7,8
ПК мониторы	352	2642	400	15,1
ПК принтеры	37	1837	81	4,4
ПК ноутбуки	273	2734	299	10,9
ПК планшеты	179	3597	288	8,0
Настольные ПК	143	5887	370	6,3
Цифровые фотоаппараты	179	1824	42	2,3
Видеокамеры	152	560	56	10,0
ВСЕГО	1699	24345	2055	8,4

Источник: ИСТАТ

Результаты ясно показывают, что методы извлечения данных из веб-страниц имеют потенциальные преимущества в смысле объема получаемой информации и в смысле повышения эффективности процесса производства данных в отношении обследования товаров потребительской электроники для построения Гармонизированного индекса потребительских цен (ГИПЦ) в Италии. В то же время возникают важные вопросы: возможно ли использовать эти «большие данные», извлеченные из веб-страниц, для повышения качества обследования потребительских цен? Как можно объединить этот перспективный канал (если возможно) с другими каналами получения данных для целей оценки инфляции (традиционные – через сборщиков данных и новые как сканирование данных)? Ниже этот вопрос будет обсуждаться с учетом результатов предварительного тестирования методов извлечения данных из веб-страниц при сборе информации о ценах на авиабилеты.

4. Тестирование методов извлечения данных из веб-страниц в обследовании цен на авиаперевозки

4.1. Обследование цен на авиаперевозки

ИСТАТ проводит централизованное обследование цен на авиаперевозки в течение долгого времени. Для этого имеются практические основания, поскольку сбор данных о ценах на авиаперевозки очень неэффективен, если его не проводить централизованно; обследование можно оптимизировать, используя возможности интернета. Существуют явные преимущества, общие для всех централизованных обследований: прямой контроль над всем процессом – от стратификации и выборочных процедур через обработку данных до построения индексов; возможность принять очень четкий план обследования и быстро адаптировать методы и процедуры; прямой контроль над правилами, законами и регулированием, которые могут влиять на цены; хороший охват продуктов; участие небольшого числа узкоспециализированных сотрудников.

Совокупность единиц для обследования цен на авиаперевозки состоит из пассажиров, перевезенных коммерческими авиаперевозчиками, прибывающими в итальянские аэропорты и убывающими из них. Чартерные рейсы исключаются, и принимаются во внимание только путешествия на каникулы/отдых как традиционными, так и низкобюджетными перевозчиками.

Класс по Классификатору индивидуального потребления по целям (КИПЦ) (пассажирский авиатранспорт, вес в корзине продуктов для расчета ГИПЦ составлял 0,85% в 2013 г.) сосредоточен в трех потребительских сегментах, для которых ежемесячно рассчитываются индексы: внутренние перелеты, европейские перелеты, межконтинентальные перелеты. Три сегмента далее стратифицируются по типу вектора, назначению и маршруту (назначения межконтинентальных перелетов разбиваются по признакам континента, субконтинента и мест назначения за пределами Европы). Месячные цены собирают в соответствии со следующими определениями продукта: один билет, экономический класс, взрослый, по фиксированному маршруту, соединяющему два города или мегаполиса, перелет туда-обратно, с фиксированными датами, конечная цена, включая аэропортные или агентские сборы.

В 2013 году выборка состояла из 208 маршрутов (аэропорты Италии): 47 национальных маршрутов, 97 европейских маршрутов, 64 межконтинентальных маршрута, при этом 81 маршрут – традиционные перевозчики и 127 маршрутов приходится на низкобюджетных перевозчиков.

Сбор данных по позиции «пассажирский авиатранспорт» имеет свои особенности: данные о ценах собирают путем моделирования покупок в Интернете в соответствии с фиксированным календарем. Для большинства маршрутов/векторов, данные собираются ежемесячно обычно в первый вторник месяца (день X). День вылета при моделировании покупки авиабилет считается (A) = X+10 дней и (B) = X+1 месяц, при этом считается при покупке билета туда-обратно, что время пребывания составит одну неделю для внутренних и европейских перелетов и две недели для межконтинентальных перелетов. Для некоторых маршрутов/векторов сбор данных проводится два раза в месяц (в первый и второй вторник, то есть 'дата X+ 7 дней' в каждом месяце).

На Рис. 2 приведен пример календаря для сбора данных о ценах на авиаперелеты, который проводится два раза в месяц.

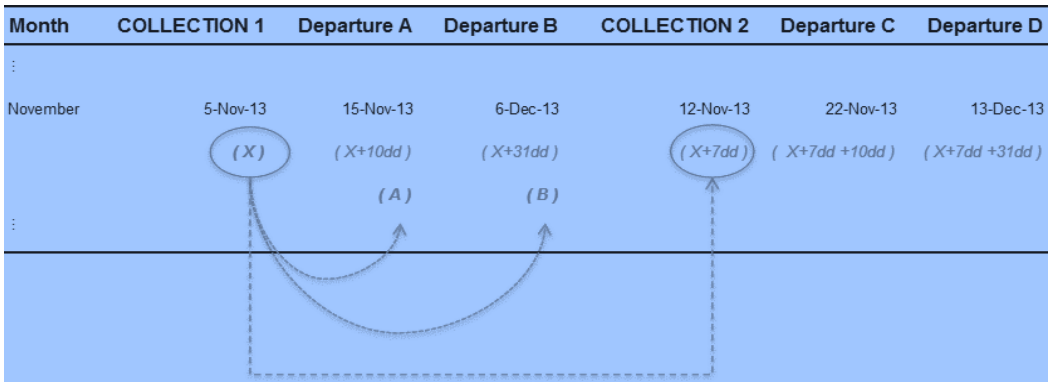

Рис.2 Пример календаря для сбора данных для построения ИПЦ/ГИПЦ в Италии. 2013 г.

В 2013 г. сбор данных проводился на веб-сайтах 10-ти низкобюджетных авиаперевозчиков и трех интернет агентств, продающих авиабилеты (Opodo, Travelprice и Edreams), где собирались данные только по ценам традиционных авиаперевозчиков. Ежемесячно регистрировалось более 960 элементарных котировок цен, которые соответствовали самому дешевому тарифу экономического класса, имевшемуся в наличии на момент бронирования, для данных дат и маршрута, включая налоги и обязательные платежи; низкобюджетные авиаперевозчики в выборке были зафиксированы.

Два специалиста, каждый из которых работает примерно 15 часов в месяц в течение трех дней, осуществляют процесс сбора данных о ценах на авиаперевозки в Италии.

4.2. Тестирование методов для извлечения данных из веб-страниц для обследования цен на авиабилеты. Предварительные результаты

Во второй половине 2013 г. и начале 2014 г. основное внимание было уделено разработке процедур для использования извлечения данных веб-страниц для регистрации цен на услуги воздушного транспорта.

Цель тестирования методов извлечения данных о ценах на авиаперевозки из веб-страниц состояла, в основном, в проверке возможности повышения эффективности обследования (аналогично потребительской электронике).

Характеристики и особенности обследования цен на пассажирские перевозки воздушным транспортом (как указано в п.4.1) приводят к определенному поведению при покупках через интернет, моделированию дат, аэропортов, авиакомпаний и определению стоимости. Учитывая это, мероприятия по тестированию методов извлечения данных из веб-страниц для сбора информации о тарифах на авиаперевозки требовали разработки и сборки макросов для

извлечения данных наряду с реализацией множества элементов логического контроля, использования мощного интерфейса сценариев (Scripting Interface), который обеспечивает обмен сообщениями между iMacros и каждым языком сценариев Windows (Windows Scripting language) или языком программирования, использованными на рассматриваемых веб-сайтах.

Первыми были рассмотрены такие низкобюджетные перевозчики, как EasyJet, Ryanair, и Meridiana. Методы извлечения данных из веб-страниц были применены к традиционным авиакомпаниям с использованием интернет агентства Opodo.it.

Что касается низкобюджетных перевозчиков, то для сайта каждой компании были характерны свои проблемы. Сайт компании EasyJet не позволял нам собирать данные о ценах, используя традиционную ссылку www.easyjet.com/it/, здесь требовалось описание конкретного аэропорта (отличное от простых кодов аэропортов в системе IATA). Компания Ryanair в самом начале тестирования, применяла тест CAPTCHA - компьютерный тест, используемый для того, чтобы определить, кем является пользователь системы - человеком или компьютером, таким образом, это остановило нас в разработке макроса для извлечения данных.

На вебсайте компании Meridiana при ответе на конкретный запрос были показаны дополнительные страницы, где предлагались опциональные услуги или запрашивалась информация о путешественниках до демонстрации конечной цены, таким образом, нам приходилось разрабатывать другой и более сложный макрос для извлечения данных о ценах.

Наконец, мы сконцентрировали внимание на компании EasyJet, и разработанные макросы обеспечили очень хорошие результаты, правильно повторяя ручной сбор данных, однако экономия времени оказалась довольно незначительной. Это объяснялось затратами времени на подготовку входящих файлов, используемых макросами для правильной идентификации маршрутов и дат, для которых происходит сбор цен, и на подготовку результатов, которые можно было использовать для построения индексов; кроме того, принимая во внимание ограниченное количество элементарных котировок цен (60), могла быть достигнута только незначительная экономия времени при использовании методов извлечения данных о ценах из веб-страниц, несмотря на то, что это было мощным инструментом для получения больших объемов элементарных данных эффективным способом.

Методы извлечения данных из веб-страниц для цен на авиаперевозки, предлагаемые традиционными авиаперевозчиками, были применены для сбора данных интернет агентства Opodo (www.opodo.it), которые включали 160 месячных котировок цен. Результаты применения макроса были оценены как с точки зрения улучшения эффективности, так и с точки зрения согласованности с данными, скаченными вручную. Что касается эффективности, то и в этом случае результаты оказались довольно скромными. В последнем тесте ежемесячного сбора данных, выгрузка 160-ти элементарных котировок цен при помощи макроса для Opodo заняла 1 час 48 мин., тогда как вручную эта выгрузка было проведена примерно за 2 с половиной часа. Для работы с Opodo необходимо подготовить входной файл для запуска макроса для поиска правильной выборки маршрутов и – вдобавок к макросу для EasyJet – для проведения различия между традиционными и низкобюджетными авиаперевозчиками. Поэтому продолжительность времени, необходимого для автоматической регистрации цен, не очень сильно отличается от времени, требуемого для ручной регистрации, принимая во внимание, что необходимо дополнительное время для обновления макроса. Но следует понимать, что в отличие от случая с компанией EasyJet, где количество

элементарных котировок цен было ограничено (60), если макрос для Opodo работает правильно и необходимы только небольшие проверки, то поскольку сбор данных касается 160-ти элементарных цен, два часа ручной работы может быть сэкономлено и посвящено другим фазам процесса производства или улучшению качества и охвата обследования.

4.3. Тестирование методов для извлечения данных из веб-страниц для обследования цен на авиабилеты. Возникшие вопросы

При использовании макроса для Opodo возникло два разных вопроса.

Первый вопрос статистического характера и касается выявления и исключение низкобюджетных авиаперевозчиков (как это делается при ручном сборе данных). В частности, когда для конкретных маршрутов и дат, макрос для Opodo встречает низкобюджетного перевозчика, который показывает самую низкую цену, на странице, где традиционный перевозчик также предлагает перелет по такой же цене, он правильно исключает низкобюджетного перевозчика, но в то же время, он уходит со страницы, не обнаружив цену традиционного перевозчика (которую следует зарегистрировать в соответствии с правилами, установленными для проведения обследования).

Второй вопрос имеет юридический характер. Вебсайт Opodo, так же как сайт eDreams (эти два сайта связаны) выставляют барьеры для автоматической процедуры ИСТАТА, как только она определяется как таковая. Поэтому ИСТАТ подготовил обращение, имеющее юридическую значимость, к администраторам обоих веб-сайтов с тем, чтобы автоматическим процедурам ИСТАТа по извлечению данных из веб-страниц было позволено беспрепятственно автоматически регистрировать информацию о ценах с их сайтов. Это обращение было сформулировано следующим образом: «обследование проводится с использованием автоматических процедур и с участием операторов, которые сгружают информацию о ценах напрямую из веб-страниц, используя методы и процедуры, разработанные данным институтом. Автоматические процедуры, разработанные ИСТАТом для сбора данных о ценах, также как и действия операторов, будут выполняться при помощи отправки запросов на сайт www.opodo/edreams.it».

После этого обращения и Opodo, и eDreams вновь открыли доступ к своему сайту, но как только были запущены новые роботы для регистрации цен на авиаперевозки, и Opodo, и eDreams автоматически заблокировали доступ. Новое официальное сообщение позволило ИСТАТу вновь получить доступ к сайтам; фактически, оказалось, что такой инструмент (официальное обращение НСС с юридической значимостью) является необходимым, но необходимо принять другие инструменты для решения проблемы получения стабильного доступа к информации на веб-сайтах при помощи автоматических процедур, используемых НСС. Это – очень важно, поскольку сбору данных для целей официальной статистики никогда нельзя препятствовать. В этом отношении ИСТАТ рассматривает пути для заключения технических соглашений с основными провайдерами интернет услуг, включенными в выборку для обследования потребительских цен.

При рассмотрении результатов, полученных при использовании методов извлечения данных из веб-страниц для EasyJet и Opodo, становится понятно, что работа по автоматическому получению данных о ценах на авиаперевозки из вебсайтов для текущего обследования потребительских цен еще не завершена. Основные вопросы, которые необходимо решить в будущем таковы:

- Использование методов извлечения данных из веб-страниц только как инструмента для повышения эффективности текущего обследования дает небольшие преимущества в смысле экономии времени, если не расширять сбор данных. Тем не менее, совершенствование и поддержание конкретных макросов для более широкого использования методов извлечения данных из веб-страниц для всего сбора данных о ценах на авиаперевозки в сети может улучшить качество (более автоматизированный и контролируемый процесс производства) и позволить сотрудникам ИСТАТа, занимающихся сбором данных, заняться другими делами, даже если это будет небольшое количество часов.

- Задачей на будущее является рассмотрение теста по использованию извлечения данных о ценах на авиаперевозки из веб-страниц в рамках подхода к использованию «больших данных», подбор технологических методов извлечения данных (отличных от iMacros), которые позволяют сгружать большое количество веб-страниц для извлечения (офлайн) информации о ценах (Cavallo, 2013). Следует провести тестирование методов для ежедневного скачивания всех цен по широкому кругу дат для выбранных маршрутов. Это может позволить изучить возможность расширения временной выборки (важнейший аспект для измерения изменения цен во времени, то есть инфляции) и сравнения полученных результатов с уже имеющимися при использовании текущего подхода. Необходимо обсуждать и решать статистические вопросы в отношении текущего дизайна обследования, который может быть пересмотрен, даже если он находится в полном соответствии с европейским регулированием (Eurostat, 2013).

- Потребуется значительный реинжиниринг (то есть обеспечение взаимодействия iMacros или другого ПО с базой данных Oracle) и реорганизация (перемещение людских ресурсов от сбора данных к проверке и анализу данных) производственного процесса при широком использовании макросов для извлечения данных из веб-страниц для обследования цен на авиаперевозки.

- Поиск решения такой важной проблемы, как наличие стабильного доступа к информации в сети.

5. Методы извлечения данных из веб-страниц и качество обследования потребительских цен

Улучшение качества обследования потребительских цен для повышения качества ГИПЦ является важнейшей целью проекта Евростата по многоцелевой статистике цен: результаты тестирования методов извлечения данных о ценах на потребительскую электронику и пассажирские авиаперевозки поставили много важных вопросов.

Влияние использования методов извлечения данных из веб-страниц на качество статистического обследования (обследования потребительских цен) может быть проанализировано по четырем основным аспектам качества [3] в смысле уменьшения ошибки при принятии логики подхода для оценки общей ошибки обследования [9].

Первый аспект касается ошибки охвата. Это ошибка связана с перечнем для отбора, который не точно представляет совокупность по тем характеристикам, на измерение которых направлено обследование. Использование методов извлечения данных из веб-страниц не должно оказать влияния на этот тип ошибки обследования. Тем не менее, использование интернета в качестве источника данных ставит вопрос о перечне для формирования выборки более сложным образом, чем это было в прошлом. Если перечень для формирования выборки единиц для сбора информации о потребительских ценах должен быть списком

всех предприятий/местных единиц, которые продают товары и услуги, на которые производятся расходы домохозяйств, этот список должен содержать некоторую информацию, касающуюся величины оборота, реализуемого этими предприятиями/местными единицами по разным каналам розничной торговли, проводя различия между физическими торговыми точками и интернет-магазинами. Выбор методов сбора данных (в поле, в сети, другие) должен определяться на основе характеристики совокупности, но использование методов извлечения данных из веб-страниц не поможет решить проблему ошибки охвата. Это в принципе зависит от того, насколько представительным является список соответствующей совокупности.

Второй аспект – это ошибка выборки. Эта ошибка присуща природе выборочного обследования, которое имеет целью измерение характеристик генеральной совокупности на основе подмножества. Принимая во внимание цель обследования потребительских цен (измерение инфляции), время является размерностью выборки, которое следует учитывать наряду с единицей выборки, которая является сочетанием предприятий и продуктов. Методы извлечения данных из веб-страниц могут позволить улучшить временной аспект выборки, который сейчас ограничен используемыми традиционными технологиями. В отношении этого тесты, проведенные в рамках проекта и описанные в данной работе, являются неадекватными. На следующих этапах проекта важность этого аспекта отбора для измерения инфляции следует тщательно проанализировать и сравнить результаты, полученные на основе более частых наблюдений, чем один раз в месяц или в два месяца, учитывая оценку изменения потребительских цен во времени.

Третий аспект связан с ошибкой, вызванной неответами, что вызывает искажения в оценках, когда единицы выборки, не представившие ответов, сильно отличаются от ответивших с точки зрения обследуемых характеристик. Это важный момент для принятия методов извлечения данных из веб-страниц для обследования потребительских цен. Если некоторые группы предприятий/продуктов обследуются только с помощью извлечения данных из веб-страниц, то предполагается, что извлеченные цены являются ценами операций и представляют потребительские цены и в физических, и в виртуальных магазинах. Если НСС выбирает фирму, которая продает товары потребительской электроники, используя оба этих канала (физические магазины и интернет), то следует собирать данные о ценах операций в обоих каналах. Выбор только метода извлечения данных о ценах из интернета может потенциально привнести ошибку, связанную с неответами, если изменения цен в физических магазинах сильно отличается от зарегистрированных изменений в интернет магазинах. Следует подчеркнуть, что a) для НСС очень дорого параллельно собирать данные о ценах в поле и в интернете и b) ошибка, связанная с неответами, уже проистекает из самого выбора интернета как единственного источника данных для некоторых продуктов, где традиционный сбор данных очень дорог. Переход от использования «копирования и вставки» к извлечению данных из веб-страниц не ухудшает проблему неответов, но позволяет выяснить быстрее, если изменение во времени цен в интернет магазинах представительно для изменения цен в физических магазинах. Для завершения анализа проблемы неответов следует упомянуть о снижении процента ответов в национальных обследованиях домохозяйств и предприятий, характерном для настоящего времени [8] и о том, как методы извлечения данных из интернета могут помочь справиться с этой проблемой.

Последний аспект касается измерения ошибки, вызванной неправильными ответами на вопросы обследования. При прочих равных условиях (если принять

интернет в качестве источника данных), использование методов извлечения данных из веб-страниц должно значительно сократить один из основных источников ошибок измерения, а именно - ошибки сборщиков данных. Кроме того, тесты должны быть расширены для того, чтобы также извлекать из веб-страниц информацию о характеристиках товаров (бренды, модели), для которых собирают информацию о ценах. Это очень важно для решения одного из основных статистических вопросов при проведении обследования потребительских цен и построении ГИПЦ, которым является вопрос о замене товаров, отобранных в выборку в начале цикла обследования.

Таким образом, a) в отношении ошибки охвата, извлечение данных из веб-страниц, по существу, нейтрально b) в отношении ошибок, связанных с неответами, оно рискованно, но таковым же является использование интернета как единственного источника данных при традиционном подходе «копирования и вставки» c) что касается ошибок выборки и ошибок измерения, то использование методов извлечения данных из веб-страниц должно гарантировать улучшения.

В заключение можно сказать, что если оценивать эффективность использования этих методов, документально подтвержденную тестами ИСТАТа, то в целом можно сказать, что при использовании методов извлечения данных из веб-страниц преимущества преобладают над недостатками, хотя и улучшения и недостатки следует изучить более тщательно.

6. Возможные направления развития и заключительные замечания

Разработка и тестирование процедур для извлечения данных из веб-страниц при сборе данных в рамках обследования потребительских цен в Италии подтвердили большой потенциал автоматического обнаружения информации о ценах в интернете (и связанной информации) с точки зрения повышения эффективности и качества. Эти результаты ставят четкие задачи для статистиков.

В том, что касается эффективности производственного процесса, проведенные тесты для потребительской электроники и авиаперевозок, ясно показали, что можно достичь важных положительных результатов. Сейчас эти улучшения очевидны, когда поиск данных ведется на нескольких веб-сайтах с большим объемом информации, но это может измениться, если нужно будет собирать больше данных с нескольких разных веб-сайтов. Рассматривая возможности использования методов извлечения данных из веб-страниц для целей расчета паритетов покупательной способности (ППП) или средних цен на уровне первичных групп в рамках международных сопоставлений, можно сказать, что, видимо, их использование следует ограничить суб-национальными пространственными сопоставлениями потребительских цен, для которых будет необходимо проводить сбор данных на определенном количестве веб-сайтов.

Что касается качества, то в пункте 5 было отмечено, что плюсы здесь превышают минусы. Если говорить о задачах статистиков (и в частности официальных статистиков), то они возникают в рамках использования «больших данных» для целей официальной статистики. Эти задачи уже были рассмотрены в исследовании, проведенном экономистами-исследователями в Массачусетском технологическом институте (МТИ) в рамках проекта «The Billion Prices Project @ MIT», цель которого состояла в мониторинге ежедневных колебаний цен в розничных интернет-магазинах по всему миру.

Во-первых, методы извлечения данных из веб-страниц как инструмент использования «больших данных» для измерения инфляции имеет прямое отношение к одной из трех определяющих характеристик «больших

данных» - скорости (одно из трех «V»- Velocity), которая очень важна для описания явления, которое характеризуется изменениями во времени.

Во-вторых, методы извлечения данных из веб-страниц для целей статистики потребительских цен широкодоступны, и поэтому национальные статистические службы могут утратить конкурентоспособность и потерять свое монопольное положение в отношении данных и информации, существующие в настоящее время благодаря их официальному статусу.

Последнее, но, тем не менее, важное соображение состоит в том, что методы извлечения данных из веб-страниц могут предоставить доступ к большим объемам данных по сравнению с текущими методами сбора статистических данных, следовательно, возможно улучшение оценки инфляции. Этот вопрос был кратко рассмотрен в разделах, посвященных двум анализируемым продуктам, но в действительности использование этих методов в перспективе предполагает обсуждение дизайна выборки, который часто не позволяет совсем или только частично позволяет использовать методы «больших данных» в рамках существующих схем отбора. Поэтому понятно, что в полной мере использовать потенциал извлечения данных из веб-страниц (и виртуально «больших данных») можно после проведения в ближайшем будущем более глубоких исследований, включая изучения возможностей методов извлечения данных из веб-страниц, отличных от iMacros. В любом случае нам следует решать задачи, связанные с использованием «больших данных» и интеграцией новых возможностей в существующие статистические обследования, поскольку они были задуманы и до настоящего времени организованы в соответствии с традиционными подходами.

Литература

[1] A. Cavallo, *Scraped Data and Sticky Prices.* MIT Sloan Work-ing Paper No. 4976-12. 2013

[2] DGINS, *Scheveningen Memorandum: Big Data and Official* Statistics. 2013

[3] D.A. Dillman, J.D. Smyth and L.M. Christian, *Internet, Phone, Mail and Mixed-Mode surveys, the Tailored Design Method.* Wiley, 4th edition. 201

[4] Eurostat - *Compendium of HICP reference documents,* Methodologies and Working papers. 2013

[5] Eurostat - *Draft ofHICP manual.* 2013

[6] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Rox-burgh and A.H. Byer, *Big data: The next frontier for innovation, competition, andproductivity.* McKinsey Global Institute.2011

[7] United Nations, *Big data and modernization of statistical systems. Report ofthe Secretary-General.* 2014

[8] United Nations *Big Data for Development: Challenges & Op-portunities.* 2012

[9] H.F. Weisberg, *The Total Survey Error Approach - A Guide to the New Science of Survey Research.* London, University of Chicago Press. 2005