Организация Объединенных Наций ECE/CES/GE.41/2016/4

Экономический и Социальный Совет

Distr.: General
7 July 2016
Russian
Original: English

Европейская экономическая комиссия

Конференция европейских статистиков

Группа экспертов по переписям населения и жилищного фонда

Восемнадцатое совещание
Женева, 28−30 сентября 2016 года
Пункт 5 предварительной повестки дня
Методы оценки качества и пригодности для использования регистров и административных источников

В каких случаях административных данных достаточно для использования вместо статистической информации? Показатель качества, основанный на сопоставлении переписей

Записка Статистического управления Португалии (СУП)[1]

Резюме

Статистическое управление Португалии рассматривает возможность использования административных данных для целей переписи 2021 года. С тем чтобы решить эту проблему, путем сопоставления административных данных с переписной информацией производится проверка качества имеющихся административных наборов данных. Задача состоит в том, чтобы оценить риски замены части собранной в ходе переписи информации сведениями, полученными из административных источников. Были применены методы увязки результатов переписи 2011 года и административных данных. Полученные результаты свидетельствуют об очень высоких показателях соответствия сопоставляемой информации для каждой подобранной пары записей как для географических, так и для демографических переменных. При сопоставлении социально-экономических переменных результаты оказываются менее однородными:

[1] Анабела Дельгаду*, Сандра Лагарту*, Паула Паулину*, Жоау Капелу (Группа по вопросам переписи 2021 года).

GE.16-11660 (R) 050816 050816

1611660

идентичная информация распределяется менее равномерно между переписью и административными источниками. Кроме того, с учетом того, что некоторые статистические данные могут быть получены из других источников, некоторые микроданные переписи (по экономическим и образовательным характеристикам населения) были сравнены с данными национального обследования рабочей силы. Эти результаты совпадают с общими результатами сопоставления, проведенного в рамках данного мероприятия. Наконец, для проверки достоверности результатов сопоставления были использованы результаты контрольного обследования переписи 2011 года.

I. Справочная информация

1. В стратегии переписи 2021 года Португалии предусмотрено использование административных данных для предоставления информации по некоторым конкретным вопросам переписи – согласно наблюдаемой в странах ЕС и ЕЭК ООН общей тенденции к использованию более эффективного метода переписи – при высоких стандартах качества, но при меньшей нагрузке на респондентов и меньших затратах для государства. Статистическое управление Португалии (СУП) в настоящее время готовит для переписи населения и жилищного фонда 2021 года технико-экономическое обоснование новой модели оценки пригодности имеющихся административных данных для статистических целей.

2. Один из элементов этого исследования предусматривает сравнение характеристик набора основывающихся на регистрах данных о населении с соответствующими характеристиками результатов национальной переписи населения 2011 года. В результате этого мероприятия будет определено, насколько административные данные, полученные из нескольких источников, близки к данным, собранным в ходе переписи, а также будут выявлены расхождения.

3. С тем чтобы подтвердить результаты оценки некоторых экономических и образовательных характеристик населения, мы также проводим сопоставление микроданных, полученных в ходе переписи 2011 года, с результатами обследования рабочей силы Португалии за первый квартал 2011 года (ОРС). Кроме того, для подтверждения результатов мы используем индекс последовательности переписи 2011 года (ИПП), рассчитанный в ходе контрольного обследования (КО).

II. Выбор административных источников и переменных

4. С учетом технико-экономического обоснования для переписи 2021 года правовую основу, в соответствии с которой Статистическое управление Португалии получило доступ к административным данным, образуют Закон № 22/2008 о национальной статистической системе от 13 мая 2008 года и Обсуждение в Национальной комиссии по защите данных № 929/2014 от 11 июня 2014 годы (были зашифрованы цифровые идентификаторы и ограничен доступ как к именам и фамилиям, так и адресам).

5. Для нынешнего обследования с учетом возможного использования административных данных в целях переписи были отобраны девять источников данных (см. таблицу 1). По отобранным административным источникам данных были определены следующие 15 целевых переменных, ранее полученных в ходе национальной переписи 2011 года: 7 – по географическим и демографиче-

ским характеристикам и 8 – по экономическим и образовательным характеристикам (см. таблицу 2).

Таблица 1
Источники административных наборов данных для сопоставления с микроданными переписи 2011 года

Административные источники	Год	Число записей	Описание	Название
Институт регистрации и юридического оформления (ИРО)	2010 год	11 565 714	Регистр населения	БДИК
Иммиграционная и пограничная служба (ИПС)	2011 год	434 708	Регистр иностранцев	ИПС
Институт социального обеспечения (ИСС)	2011 год	7 209 027	Регистр социального обеспечения	ИСС
Управление по вопросам стратегии и планирования (УСП)	2011 год	2 736 659	Регистр занятости («Бюллетень трудовых ресурсов и занятости»)	КП («Квартальный статистический бюллетень Португалии»)
Институт по вопросам занятости и профессиональной подготовки (ИЗПП) и Региональное управление статистики Мадейры (РУСМ)	2011 год	702 215	Регистр безработных	ИЗПП
Главное управление статистики образования и науки (ГУСОН) и Региональный секретариат по вопросам образования и людских ресурсов автономного региона Мадейра (РСО)	2011 год	1 965 842	Регистр учащихся	ГУСОН
Общий пенсионный фонд (ОПФ)	2010 год	1 103 980	Регистр Пенсионного фонда государственных служащих	ОПФ

Таблица 2
Отдельные административные вопросы для сопоставления с переменными переписи 2011 года

Наборы административных данных	Имеющаяся информация по демографическим вопросам
БДИК	Место жительства (муниципалитет), пол, дата рождения, юридическое брачное состояние, страна рождения, страна гражданства
ИПС	Страна рождения, страна гражданства, текущий статус активности, род занятий
ИСС	Текущий статус активности, место работы, статус занятости
КП	Место работы, род занятия, отрасль (предприятие), положение в занятии, число лиц, работающих на предприятии, продолжительность рабочего дня, образовательный уровень
ИЗПП	Текущий статус активности
ОПФ	Текущий статус активности
ГУСОН	Посещение учебных заведений

III. Методологические аспекты

6. Цель этого мероприятия состоит в том, чтобы сопоставить для каждого лица точное значение целевой переменной по административным наборам данных, которая больше всего соответствует статистической концепции и определению, с микроданными переписи 2011 года.

7. Сравниваемые данные о населении получены на основе результатов предыдущего сопоставления микроданных переписи 2011 года и административных записей, поэтапно отобранных из нескольких источников (с использованием сочетаний имеющейся информации – пол/имя и фамилия/дата рождения/семейное положение/страна гражданства/муниципалитет обычного жительства – для последовательной увязки микроданных переписи по каждому набору административных данных). Подготовка данных (включая регистрацию) и их стандартизация были осуществлены ранее. Опущенные характеристики в регистры не добавлялись, и было сочтено, что данные обновлены.

8. Удалось сопоставить 9 949 599 записей переписи с административными записями из выбранных источников, что покрывало 94% постоянного населения в 2011 году, при этом показатель ложных позитивных распознаваний составил 6% (это значение представляет собой общее число записей переписи, сопоставленных по меньшей мере с одним набором административных данных).

9. С учетом сопоставленных записей основная цель этого мероприятия заключается в том, чтобы оценить для отобранных переменных, получим ли мы информацию из наборов административных данных о физических лицах, аналогичную той, которая была собрана в ходе переписи 2011 года. Только после

анализа этих результатов мы могли бы рассмотреть возможность использования административных данных для замены статистической информации, собранной в ходе переписи.

10. Показатель соответствия оценивался на основе сопоставления точной информации по каждой паре записей, поддающейся увязке. В отношении этих записей, которые касались одного и того же физического лица, наша гипотеза состоит в том, что если соответствие подтверждается, то мы можем рассчитывать на административную информацию для статистических целей. В обоснование такого решения мы можем использовать два дополнительных критерия: результаты контрольного обследования переписи 2011 года (ИПП), а также результаты сопоставления данных переписи 2011 года и микроданных обследования рабочей силы за первый квартал 2011 года.

IV. Результаты и обсуждение

11. В таблице 3 приводятся результаты, полученные в ходе сопоставления, для набора отобранных переменных переписи с указанием в качестве сравнения имеющейся административной информации. Мы приводим данные о численности населения, количестве имеющихся административных записей и фактическом числе административных записей в сравнении с микроданными переписи (полученными в результате процесса сопоставления). Мы также приводим значения глобального индекса последовательности (ИПП) по результатам контрольного обследования (КО) переписи 2011 года (СУП, 2013 год).

12. Прежде чем представить результаты, два замечания: первое касается категорийных переменных, а второе – переменных с различными уровнями детализации информации. В настоящем документе мы приводим только результаты по всем категориям и сводную информацию, однако проведенное исследование носило исчерпывающий характер, было подробным в плане сопоставления и позволило получить разнообразные результаты.

13. Первое замечание касается того, что все категорийные переменные были также подвергнуты сопоставлению по группам. Если мы возьмем, например, текущий статус активности, то в таблице 3 для сравнения микроданных переписи с отдельными регистрами ИСС по всем категориям коэффициент соответствия составит примерно 81%. В этом случае в рамках групп сопоставления могут несколько различаться. Снова обратимся к текущему статусу активности: 92% физических лиц, указавших в переписном листе, что они *работают*, в системе социального обеспечения Португалии зарегистрированы в качестве *занятых*.

14. Второе замечание касается рассмотрения переменных с разными уровнями детализации информации. Если мы возьмем, например, занятость, то в таблице 3 для сопоставления микроданных переписи с индивидуальными регистрами КП коэффициент соответствия составит примерно 63%. Эта величина означает более высокий уровень агрегирования информации, т.е. до одного знака. Для этого вида переменных общая тенденция заключается в том, что чем выше уровень дезагрегирования, тем ниже оценочный коэффициент соответствия.

15. Перейдем теперь к анализу результатов глобального сопоставления в таблице 3. Сопоставление результатов по демографическим переменным свидетельствует о высоком коэффициенте соответствия – 90–99% – для даты рождения, пола, страны рождения, страны гражданства и юридического брачного со-

стояния. Кроме того, весьма высокий коэффициент соответствия имеет показатель места обычного жительства: информация полностью совпадает примерно для 95% всех включенных в регистры пар, подвергнутых сравнению.

Таблица 3
Результаты сопоставления микроданных переписи 2011 года и административных записей

Переменная	Число записей, подлежащих сравнению, согласно переписи населения 2011 года	Число административных записей, подлежащих сравнению с переписью 2011 года, с разбивкой по источникам		Число пар, подвергнутых сравнению	Коэффициент соответствия для пар, подвергнутых сравнению (%)	ИПП[2] (%)
Место жительства (муниципалитет)	10 562 178	БДИК	11 565 714	9 308 384	94,6	**97,7**
Пол	10 562 178	БДИК	11 565 714	9 308 384	99,9	**99,0**
Дата рождения	10 562 178	БДИК	11 565 714	9 308 384	92,6	**95,7**
Юридическое брачное состояние	10 562 178	БДИК	11 565 714	9 308 384	95,3	**97,4**
Страна рождения	10 562 178	БДИК	11 565 714	9 308 384	94,7	**84,0**
		ИПП	434 708	107 136	91,3	**84,0**
Страна гражданства	10 562 178	БДИК	11 565 714	9 308 384	99,4	**97,8**
		ИПП	434 708	107 136	90,3	**97,8**
Текущий статус активности	8 989 849	ИСС	7 066 838	4 910 073	81,2	
		ИПП	379 965	107 136	27,1	
		ОПФ	1 103 980	716 264	92,1	
		ИЗПП	702 215	454 479	42,1	
Место работы (муниципалитет)	4 361 187	ИСС	4 107 425	2 788 758	56,6	**77,6**
		КП	2 736 659	2 045 476	81,6	**77,6**
Род занятий	4 361 187	КП	2 736 659	2 045 476	61,9	
		ИПП	124 721	171 370	52,9	
Отрасль	4 361 187	КП	2 736 659	2 045 476	74,1	
Положение в занятии	4 361 187	КП	2 736 659	2 045 476	93,0	**82,2**
		ИСС	4 107 425	2 788 758	85,5	**82,2**

[2] ИПП измеряет ошибки содержания; он представляет собой долю статистических единиц (постоянное население) в пределах одной и той же классификация для переписи 2011 года и КО переписи по всем общим единицам для двух статистических операций.

Переменная	Число записей, подлежащих сравнению, согласно переписи населения 2011 года	Число административных записей, подлежащих сравнению с переписью 2011 года, с разбивкой по источникам		Число пар, подвергнутых сравнению	Коэффициент соответствия для пар, подвергнутых сравнению (%)	ИПП[2] (%)
Число лиц, работающих на предприятии	4 361 187	КП	2 736 659	2 045 476	54,4	**51,6**
Продолжительность рабочего дня	4 361 187	КП	2 736 659	2 045 476	56,8	
Уровень образования	10 445 093	КП	2 736 659	2 210 930	59,5	
Посещение учебных заведений	10 445 093	ГУСОН	1 965 842	1 359 916	82,2	**69,8**

16. Что касается социально-экономических переменных, то результаты являются менее однородными. Мы определяем три ситуации:

– высокий коэффициент соответствия в отношении некоторых переменных для всех имеющихся источников информации; например: положение в занятии, коэффициент соответствия для которого по данным переписи составляет около 86% в случае ИСС и 93% в случае КП;

– коэффициент соответствия со значительными различиями в зависимости от источника: такие переменные, как профессия, отрасль и текущий статус активности (в последнем случае примерно 92% для сведений, полученных через ОПФ, в то время как для сведений ИЗПП этот показатель снижается до 42%);

– коэффициент соответствия оценивается в результате сравнения с одним источником: от 50% для числа лиц, работающих на предприятии, и продолжительности рабочего дня до более чем 80% в случае показателя посещения учебных заведений.

17. Для подтверждения результатов сопоставления данных переписи и наборов административных данных мы решили воспользоваться результатами, полученными на основе показателя качества КО переписи населения 2011 года – ИПП. К нашему удивлению, оценочный коэффициент соответствия и показатели ИПП оказались весьма близкими по большинству отобранных переменных (несмотря на то, что в случае некоторых переменных применяемые концепции хотя и являются близкими, но не совпадают полностью). Этот факт подтверждает результаты, полученные в ходе сопоставления, и повышает доверие к использованию административных данных.

18. И наконец, чтобы иметь дополнительный показатель для подтверждения результатов, мы также провели сопоставление микроданных переписи 2011 года

и ОРС[3] за первый квартал 2011 года. Размер выборки ОРС составил 39 884 физических лица. Для проведения этого мероприятия необходимо было применить правило соответствия (пол/имя и фамилия/дата рождения/семейное положение/муниципалитет обычного жительства) к переписным данным. Мы получили 17 732 пар данных для сопоставления с микроданными переписи 2011 года (6 995 физических лиц в возрасте 15 лет и старше).

19. В таблице 4 указаны соответствующие результаты сопоставления – микроданные переписи в сравнении с административной информацией и микроданные переписи в сравнении с микроданными ОРС – по восьми переменным рабочей силы и образовательным переменным.

Таблица 4
Результаты сопоставления микроданных переписи 2011 года и ОРС

Переменные	Соответствие: перепись населения – ОРС (%)	% соответствия: перепись – административные записи по отдельным источникам административных данных	
Статус в составе рабочей силы	**84,3**	81,2	ИСС
Род занятий	**67,8**	61,9	КП
Промышленность	**77,6**	74,1	КП
Положение в занятии	**86,5**	93,0	КП
Число лиц, работающих на предприятии	**60,6**	54,4	КП
Продолжительность рабочего дня	**72,6**	56,8	КП
Уровень образования	**80,2**	59,5	КП
Посещение учебных заведений	**86,5**	87,4	ГУСОН

20. В этом случае, если имеется несколько административных источников для целевой переменной, мы используем самые высокие показатели коэффициента соответствия, полученные в ходе сопоставления (из таблицы 3), – микроданные переписи в сравнении с административной информацией.

21. За исключением образовательных характеристик значения коэффициента соответствия по обоим сопоставлениям для отобранных переменных являются аналогичными. Мы считаем, что эти результаты свидетельствуют о более высоком общем уровне соответствия результатов сопоставления микроданных переписи 2011 года и административных записей.

22. И наконец, замечание по вопросу об охвате объектов наблюдения. Из таблицы 3 явствует, что некоторые переменные не в полной мере охвачены административными данными Португалии, имеющимися для технико-экономического обоснования переписи 2021 года. По сути, на основе первона-

[3] Португальское ОРС, которое проводится по всей стране, представляет собой выборочное обследование, дающее ежеквартальные результаты (в последнее время – ежемесячные). В 2011 году оно позволило собрать информацию о рынке труда по примерно 40 000 физических лиц.

чального анализа диагностических информационных потребностей мы знаем, что некоторые ключевые для переписи населения и жилищного фонда вопросы (например, вопросы, связанные с домашними хозяйствами или образованием) не охватываются административными данными Португалии в полной мере или хотя бы частично. Это вопрос не является предметом рассмотрения в ходе нынешнего мероприятия, так же как и расхождения между источниками (для этого случая был подготовлен соответствующий свод правил).

V. Выводы

23. Оценка качества административных данных для статистических целей может оказаться колоссальной задачей. Одним из шагов в рамках этого процесса оценки – после рассмотрения вопросов концепций, классификаций, своевременности, обработки данных и обращения с ними, увязки и согласования данных и других вопросов – является проверка (помимо вопросов охвата) того, насколько информация, которую мы получаем из источников административных данных, представляет собой то, что нам необходимо для статистических данных переписи, иными словами, насколько она действительна и точна.

24. Вполне очевидно, что непросто достичь компромисса между тем, что мы имеем, и тем, что нам необходимо, особенно в тех случаях, когда данный процесс предусматривает использование ресурсов, которые не находятся в нашем распоряжении или под нашим контролем, как это имеет место в случае наборов административных данных. При решении этой конкретной задачи многие страны, которым предстоит переход от традиционных моделей переписи к моделям, основанным на регистрах, сталкиваются с теми же проблемами, что и Португалия. Для Статистического управления Португалии это мероприятие по простому сопоставлению является частью сложного проекта, который осуществляется в настоящее время и реализация которого будет продолжена и после переписи 2021 года.

25. Мы считаем, что эти результаты могут стать основой для обсуждения вопроса о том, должны ли административные данные заменить данные переписи или же их следует использовать в дополнение к указанным данным. В настоящее время мы приводим следующие некоторые выводы/соображения относительно полученных результатов:

– результаты свидетельствуют о том, что в подавляющем числе случаев административные данные и микроданные переписи 2011 года согласуются между собой;

– мы провели сравнение записей по отдельным лицам на основе административных данных для семи демографических переменных переписи 2011 года (все они были задействованы в рамках мероприятия с использованием правила сопоставления). Полученные коэффициенты соответствия оказались очень высокими (информация, содержащаяся в подвергнутых сравнению парах записей, на 90% совпадает);

– мы также сравнили характеристики, связанные с рабочей силой, и образовательные характеристики по восьми отобранным переменным переписи 2011 года и для некоторых переменных рынка труда получили коэффициент соответствия, превышающий 80%;

– при сопоставлении административных данных с микроданными переписи 2011 года было установлено, что данные «Квартального статистиче-

ского бюллетеня Португалии» (КП) являются наиболее последовательным – и, в глобальном масштабе, имеют самый высокий коэффициент соответствия – для всех наборов переменных, по которым имеется информация;

– показатели сопоставимости указывают на неравенство только в случае неравных значений (различия не обусловлены невозможностью преобразования данных или отсутствием описания); таким образом, мы считаем, что, несмотря на наличие явной проблемы в плане охвата, административные данные могут быть использованы в дополнение к информации, собранной в ходе переписи, или заменить ее;

– некоторые различия в результатах сопоставления можно объяснить временными интервалами между наборами данных, а также концептуальными вопросами; кроме того, в настоящее время мы связываемся с субъектами, отвечающими за источники данных, на предмет получения новых массивов данных, и мы считаем, что некоторые из обусловливающих несоответствие вопросов могут быть решены на основе более поздних представлений;

– надежность использования административных данных для статистических целей была подтверждена с использованием дополнительных критериев качества информации на основе КО и результатов сопоставления переписи 2011 года и ОРС 2011 года;

– для целей будущей деятельности изучаются правила взаимного сопоставления и иерархической градации источников административной информации.

VI. Справочная документация

НСИ – Национальный статистический институт (2013 год), «Inquérito de Qualidade dos Censos 2011 – Metodologia e resultados» («Обследование качества переписи 2011 года – методология и результаты»), Национальный статистический институт, Лиссабон.