



Европейская экономическая комиссия

Конференция статистиков Европы

Заседание экспертной группы ЕЭК ООН – Евростат по проведению переписи с использованием регистров

Женева, 22-23 мая 2012 г.

Сессия 1: Опыт использования регистров при проведении переписи

Общий подход к важности и использованию регистров при проведении переписи в Испании

Записка Национального института статистики Испании¹

Резюме

Перепись 2011 г. в Испании основывалась на смешанном подходе, включавшем обследование крупной выборки, равной 12% населения (3 млн. жилищ и 5,8 млн. человек), и использование административных регистров.

Со статистической точки зрения такая модель ставит несколько сложных задач. Во-первых, необходимо провести глубокое исследование различных регистров и проанализировать их характеристики, такие как полнота, точность и охват. С другой стороны, необходимо определить разные стратегии в целях привязки информации из регистров с информацией, полученной по результатам обследований. Кроме того, необходимо учесть другие аспекты, такие как несоответствия между разными источниками данных и то, насколько детальной является распространяемая нами информация.

В данном документе мы постараемся обобщить все эти вопросы, подчеркивая важность сильных и слабых сторон применяемой методологии, опираясь на наш опыт.

¹ Авторы - Carmen Teijeiro Breijo, Jorge L. Vega Valle и Miguel A. Martínez Vidal.

I. Регистры в Испании

1. В отличие от других стран, где имеется одна организация, которая объединяет все существующие в стране регистры, в Испании действует несколько организаций (министерств, институтов, советов, ...), которые администрируют разные виды регистров.
2. Если рассматривать различные источники, которые использовались при проведении переписи 2011 г. в Испании, следует вначале отметить два наиболее важных источника: административный регистр населения (PADRON) и перепись населения, жилого фонда и зданий 2001 г.
3. PADRON является административным регистром, имеющим собственные правовые нормы, регулирующие его управление. Операционная часть основывается на центральной базе данных в INE, куда ежемесячно все муниципалитеты (ответственные за проведение регистрации своих жителей) направляют изменения, произошедшие в местных регистрах. Помимо этого INE, совместно с муниципалитетами и другими местными органами, ведет работу по повышению качества PADRON посредством исключения из регистра дублирующих записей или сообщения об изменениях в местожительстве между муниципалитетами.
4. С другой стороны, одним из основных ограничений PADRON является отсутствие в нем демографических характеристик. В регистре населения указывается только базовая информация, такая как имя, фамилия, дата рождения, место рождения и гражданство².
5. Тем не менее, PADRON имеет ряд проблем, если мы хотим использовать его в качестве прямого источника статистических данных. Он является административным регистром и регистрация в нем обеспечивает людям основные права, такие как избирательное право, доступ к услугам здравоохранения и образования и т.д. Поэтому мы не можем принять решение о регистрации человека или его исключении из регистрации, учитывая только статистические критерии. В связи с этим, а также в целях обеспечения соответствия законодательству, на то, чтобы исключить какого-либо человека из регистрации, который выехал за рубеж без уведомления муниципалитета, может потребоваться несколько месяцев.
6. PADRON является основным регистром для получения демографических характеристик, но, несмотря на то, что он содержит территориальную информацию о местожительстве людей, эта информация не является организованной с вычислительной точки зрения. Другими словами, PADRON является регистром населения, но не их местожительств. Вследствие этого, еще одним важным оплотом переписи 2011 г. были сведения о зданиях, полученные в рамках переписи 2001 г.
7. Одним из важных результатов переписи населения, жилого фонда и зданий 2001 г. было то, что каждый человек имел постоянный код, присвоенный тому месту, где он проживал, и таким образом была обеспечена возможность привязки демографической и территориальной информации.
8. Мы воспользовались информацией, полученной по результатам переписи 2001 г., и после ее проведения обновляли информацию по местожительствам, опираясь на информацию PADRON по всем людям, которые прибывали или переезжали в

² Он также содержит вопрос, связанный с обучением, но эта информация не очень точная.

жилище, чей код отсутствовал в нашем регистре. Для переписи 2011 г. мы начали с основы, которая содержала:

- Информацию по жилищам, которая имелаась в переписи населения и жилого фонда 2001 г.
- Информацию по жилищам, которые были заселены, как минимум, единожды в период между 2001 и 2011 г.

9. Тем не менее, наша основа нуждается в обновлении для включения тех жилищ, которые были построены после 2001 и с того времени не были заселены. Вследствие этого, а также в целях сбора всех географических координат зданий нами была проведена практически исчерпывающая³ перепись зданий 2011 г.

10. Другими важными регистрами для переписи 2011 г. являются регистр национальных удостоверений личности и карт постоянного проживания (они содержат важную информацию для установления личности проживающего), книги записей актов гражданского состояния (они позволяют проверять информацию о рождениях, смертях и заключении браков), налоговый орган и органы социального обеспечения.

11. Однако для того, чтобы регистры можно было использовать в целях переписи, они должны удовлетворять нескольким условиям. Информация регистров должна часто обновляться, регистры должны обеспечивать хороший уровень охвата, регистры должны относиться к тем же понятиям, которые запрашиваются в рамках переписи, разбивка информации, используемой в регистрах, должна быть достаточной для распространения переписи, а также должна быть возможность получить доступ к регистру с технической и юридической точки зрения.

12. Далее в таблице приведены регистры, которые были пригодны для подготовки переписи 2011 г. в Испании:

Наименование регистра	Администратор регистра
Регистр населения	INE
Перепись 2001 г.	INE
Национальные удостоверения личности и вид на жительство	Министерство внутренних дел
Кадастр: Регистр жилищ	Казначейство
Налоговое управление	Казначейство
Социальное обеспечение	Министерство труда и социальных вопросов
Книги записей актов гражданского состояния: рождения, смерти и заключение брака	INE

³ В 20% зон действия системы регистрации информация была получена непосредственно из нашей обновленной основы по жилищам, а координаты этих жилищ были получены из информации кадастра.

Перспективы на будущее

13. Несмотря на то, что INE имеет некоторый опыт по использованию регистров, все еще имеются некоторые трудности для проведения переписи полностью на основе регистров. Тем не менее, некоторая информация из различных регистров может быть использована при небольших усилиях со стороны INE и органа, ответственного за ведение регистра.

14. Например, информация о потреблении электроэнергии может быть очень полезной в случае необходимости определения проживают ли в доме в течение всего года, в течение некоторого периода или стоит ли дом незаселенным. Далее информацию из других регистров необходимо усовершенствовать и проанализировать более детально для того, чтобы использовать большинство таких регистров. В этом случае мы можем найти различные регистры, которые содержат полезную информацию, такие как регистры, связанные с уровнем образования (как на уровне ВУЗов, так и на уровне школ), или регистры, содержащие информацию о занятых и безработных, или также регистры, администрируемые компаниями взаимного страхования или органами социального обеспечения.

Наименование регистра	Администратор
Потребление электроэнергии	Компании электроснабжения
Уровень образования (ВУЗ)	Министерство образования
Уровень образования (школа)	Министерство образования
Занятые/Безработные	Министерство труда и социальных вопросов

II. Методология в Испании: комбинационный подход

14. Переписи являются самой сложной задачей, стоящей перед статистическими органами. В действительности, перепись является не только сбором данных, но и логистической операцией. Для того, чтобы доказать это, сообщим, что бюджет переписи 2001 г. в Испании составлял более 200 млн. евро (сегодня это было бы свыше 300 млн. евро), в проведении переписи было задействовано более 40 тыс. человек, включая счетчиков, руководителей групп и т.д., и было подписано более 70 тыс. контрактов в целях замены людей в случае, если кто-то решал уйти.

15. Впервые имеется европейская директива по проведению переписи населения и жилого фонда. Данная директива содержит описание различных источников данных, приемлемых в целях переписи.

16. Одним из основных результатов переписи является численность населения. Кроме того, также важны другие виды информации в зависимости от того, что представляет интерес для различных пользователей: демографические и социальные переменные, информация о жилье и зданиях и многое другое.

17. В отношении первого пункта, учет численности населения не должен делаться только на основании административного регистра населения. Как отмечалось ранее, PADRON не должен использоваться без проверки содержащейся в нем информации. В связи с этим данные PADRON привязываются к другим административным

источникам, например, агентством по налогам и социальному обеспечению, для сбора доказательств реального проживания в стране. При таком подходе мы обнаружили, что более 95% населения зарегистрировано правильно и в то же время имеется некоторая информация (исходя из пола, места и даты рождения, местожительства и гражданства), которая свидетельствует о том, что люди не зарегистрированы надлежащим образом.

18. В связи с этим нами было принято решение о том, что не направлять работу в рамках переписи на местах для подсчета людей, в отношении проживания которых в стране у нас не было сомнений, и сконцентрировать свои усилия на тех лицах, в отношении которых были сомнения о том, проживают ли они в стране. Это явилось основной причиной почему не проводили исчерпывающий учет населения.

19. Кроме того, имеются некоторые административные регистры, которые могут обеспечить сведения о жилье и зданиях. Как отмечалось ранее, PADRON содержит информацию только о заселенных домах, а кадастр содержит настолько отличную структуру данных, что наши попытки по привязке этих двух регистров не дали приемлемых результатов. В связи с этим нами было принято решение провести исчерпывающий учет зданий и жилья.

20. С учетом всего этого перепись 2011 г. в Испании основывается на трех основных компонентах:

- Предпереписной файл, который является результатом процесса привязки разных административных регистров. Он будет применен для:
 - •принятия решения о том, подсчет какого населения будет проводиться (не проводиться).
 - •выявления групп записей, в отношении которых мы не знаем нужно ли проводить подсчет.
 - •в качестве первоначальной основы переписи зданий.
 - •предоставления данных по некоторым основным характеристикам населения: пол, возраст и место рождения.
- Большое обследование (около 12% населения) для сбора данных по вопросам народонаселения и жилого фонда. Эта операция будет полезна для:
 - •оценки размера групп записей, в отношении которых мы не знаем нужно ли проводить подсчет.
 - •предоставления информации по вопросам, которые не включены в предпереписной файл.
- Исчерпывающий подсчет зданий, который будет применяться для:
 - •получения полной основы учета зданий (основная необходимость для будущих обследований).
 - •получения географических координат зданий.

21. Информация переписи 2011 г. собиралась с применением нескольких каналов: Интернета, почты и **CAPÍ**. Прежде всего, в заселенные жилища направлялось письмо с предоставлением возможности ответить через Интернет. Тем, кто не воспользовался этой возможностью, направлялось второе письмо, включающее анкету в бумажном виде, и предоставлялось два варианта: ответить через Интернет или направить анкету по почте. И, наконец, те, кто не представили ответы, были

посещены и данные были собраны при помощи счетчиков, оснащенных карманными компьютерами.

Преимущества и недостатки методологии

22. Если сравнивать подход, выбранный Испанией в 2011 г., с традиционной переписью, то можно выявить несколько преимуществ.

23. Количество человек, необходимых для сбора всех данных (перепись зданий, жилого фонда и населения), составило менее 5 тыс. Чем меньше человек нам необходимо на местах, тем легче нам найти этих людей и тем более квалифицированными они являются. Кроме того, контроль всего персонала, принявшего участие в проекте, был гораздо проще, чем в предыдущие годы переписи, а процесс обучения был более эффективным.

24. Многоканальный подход увеличивает возможности домохозяйства ответить на вопросы переписи. Интернет наиболее всего подходит для молодого поколения и для людей, готовых сотрудничать в любое время и в любом месте. Бумажные анкеты лучше подходят для людей, не имеющих дома подключения к Интернету, и для людей, которым не нравятся новые технологии. И, наконец, счетчики с карманными устройствами подходят для тех домохозяйств, с которыми сложнее сотрудничать, и это позволяет обеспечить сплошной охват в рамках работы на местах.

25. Использование регистров позволяет существенно повысить качество данных переписи. Иногда регистры будут использоваться вместо того, чтобы задавать вопросы людям. В других случаях регистры будут использоваться для подтверждения ответов, а в третьих случаях регистры будут использоваться на этапах редактирования и вменения, которые обеспечивают более точную информацию.

26. Еще одним моментом, который следует учитывать, является то, что с экономической точки зрения эта модель гораздо менее дорогостоящая по сравнению с традиционной переписью. При проведении такой работы практически 70% затрат связаны со счетчиками, поэтому сокращение числа счетчиков означает меньшие затраты.

27. С другой стороны, другой сильной стороной переписи 2011 г. в Испании является сбор координат зданий. Эта информация является очень важной в силу ряда причин:

- Появляется возможность распространить информацию в более наглядном формате, таким как персонализированные карты.
- Пользователи могут задавать вопросы системе в привязке не только к административным границам, но и к географическим пунктам. Можно будет нарисовать неправильный многоугольник (который охватывает части некоторых административных территорий), и система ответит на запрос в соответствии с таким многоугольником.

28. Однако модель, подобная этой, не является идеальной и приводит к некоторым неудобствам. Далее представлен небольшой перечень таких проблем.

29. Работать с несколькими каналами (Интернет, бумажная форма, счетчики) не легко. Может оказаться так, что посредством разных каналов от одного и того же жилья будут получены разные ответы. Еще более усложняет ситуацию возможность того, что эти ответы не будут согласовываться между собой. Для того, чтобы

определить какой ответ является наиболее подходящим, необходимо реализовать некий алгоритм.

30. Помимо этого, вследствие погрешности выборки эта модель имеет сложности с выполнением детальных запросов (которые включают несколько вопросов переписи) на очень низком географическом уровне. Результаты будут настолько неточными, что их публикация не имеет смысла.

31. Кроме того, время на сбор данных было достаточно продолжительным (с сентября 2011 г. по апрель 2012 г.), а базовая дата приходилась на середину этого периода. Для того, чтобы обеспечить согласованность данных с базовой датой, пришлось провести различные корректировки.

32. И, наконец, подобная модель никогда ранее не разрабатывалась в нашей стране и с методологической точки зрения ее разработка являлась сложной задачей. Для подготовки последовательных и точных данных переписи будут применены выборочные методы и калибровка.

III. Связь между информацией переписи 2011 г. и информацией регистров

33. Прежде всего, до того, как начать говорить о связи между информацией переписи 2011 г. и информацией регистров, необходимо отметить как привязана информация из регистров.

34. С учетом своего качества PADRON играет роль основы, а информация из остальных регистров будет привязываться к PADRON и будет помогать в его проверке. PADRON содержит идентификационные данные по всем лицам, но эта информация хранится в трех разных разделах, а не в одном: номер национального удостоверения личности (DNI), номер паспорта и номер вида на жительство. У каждого человека в PADRON личные данные будут содержаться в одном из этих разделов.

35. В PADRON содержится информация об личном идентификационном номере каждого человека, включенного в него. Несмотря на то, что это происходит не так часто, но иногда мы можем обнаружить, что два или более человека имеют один номер, или у них неправильный номер или даже что у человека отсутствует идентификационный номер. Для решения этих проблем и сравнения информации в PADRON необходимо сделать привязку между информацией в PADRON и информацией Министерства внутренних дел⁴.

36. Для того, чтобы принять привязку между PADRON и информацией Министерства внутренних дел, требуется (на первом шаге) полное совпадение в этих двух источниках индивидуального идентификационного номера, имени, фамилии, пола, даты и места рождения. По тем записям, где такое совпадение не является идеальным (возможно, информация в одном из двух источников является неверной), проводятся вероятностные алгоритмы с внутренним порогом или контрольным параметром, который сравнивает сходные элементы между людьми.

37. Такого же типа процессы применяются к другим административным источникам для того, чтобы убедиться в наличии качественных идентификационных

⁴ Министерство внутренних дел является органом, ответственным за выдачу национальных удостоверений личности, паспортов и вида на жительство.

номеров во всех задействованных регистрах и в целях повышения процента привязанных записей.

38. С другой стороны, в переписном листе переписи 2011 г. мы не задаем людям вопрос об их индивидуальном идентификационном номере (номер удостоверения личности, паспорта или вида на жительство), поэтому привязка между данными переписи 2011 г. и информацией из регистров проводилась иным образом.

Различные методики

39. Прежде всего, для всех лиц, ответивших на вопросы переписи через Интернет или при помощи счетчика, было разработано специальное программное обеспечение, которое проверяет имеется ли такое лицо в наших регистрах по этому адресу. В большинстве регионов Испании такой алгоритм позволяет достичь процента привязки на уровне почти 80%. Лицо может оказаться не в привязке в силу разных причин: возможно, что адрес, по которому человек представляет ответы на вопросы переписи, отличается от адреса его регистрации, возможно, в связи с тем, что личные данные, которые человек заполняет в Интернет-форме переписного листа немного отличаются от информации, которая имеется в регистрах, или, возможно вследствие того, что между датой, когда формируется основа с информацией, и датой, когда человек отвечает на вопросы переписи, этот человек переехал в другое жилье.

40. Если человек ответил на вопросы переписи через Интернет или при содействии счетчика, но его запись не была привязана в ходе первого шага, или если человек заполнил переписной лист в бумажном виде, применяются различные методики в целях повышения количества привязок между информацией переписи и информацией из регистров.

41. Один из методов заключается в сравнении данных переписи с информацией из регистров на разных географических уровнях: тот же адрес проживания, та же зона действия системы регистрации или тот же район. Для того, чтобы привязка была принята успешной, мы требуем совпадения или близкой схожести информации по всем вопросам, которые мы сравниваем: имя, фамилия, дата рождения, гражданство... Если по одному из вопросов информация не совпадает, привязка не принимается. Поскольку привязка может быть принята при наличии схожей, но не точно совпадающей информации в обоих источниках (перепись и регистры) этот метод занимает много времени.

42. Другой метод, который был применен, заключается в построении «сверхкода» с некоторыми вопросами из переписи и регистров (код имени + код фамилии + дата рождения + код места рождения +...). В случае идеального совпадения⁵ между информацией переписи и информацией из регистров, привязка принимается. Преимущество данного метода заключается в том, что его очень легко реализовать и достаточно быстро, в связи с чем он применяется ко всем географическим уровням, включая всю страну.

43. В завершение в отношении регистров, которые еще не привязаны, был применен вероятностный подход, основывающийся на отклонении между двумя текстовыми полями, в рамках которого сравнивается информация переписи 2011 г. и информация из PADRON и других регистров.

⁵ По этой характеристике этот метод отличается от предыдущего.

44. Результаты весьма обнадеживающие и, хотя вероятностный подход еще не применялся, мы можем осуществить привязку около 97% записей, которые отвечали на переписной лист через Интернет, и 92% тех, кто отвечал в бумажной форме.

IV. Как мы получаем численность населения

45. Как отмечалось ранее, предпереписной файл содержит записи из PADRON, с привязкой к другим административным регистрам. Для получения численности населения мы присваиваем каждой записи **счетный фактор**.

46. Прежде всего, те записи, которые имеются в PADRON и также были найдены в других регистрах, должны считаться без каких-либо сомнений. В этом случае мы присваиваем таким записям счетный фактор, равный 1.

47. Во-вторых, если запись имеет привязку к какому-либо регистру смертей, предоставленному регистром актов гражданского состояния, то мы присваиваем такой записи счетный фактор, равный 0.

48. И, наконец, если запись не была привязана с каким-либо административным регистром, необходимо проверить отмечена ли запись каким-либо предостерегающим знаком⁶. Если запись отмечена знаком, то она будет иметь счетный фактор «X», где X имеет значение больше 0, необязательно является целым числом, которое должно рассчитываться с применением результатов обследования. В противном случае запись будет иметь счетный фактор, равный 1.

49. Агрегирование счетных факторов дает нам численность населения географической территории и результаты выборки будут откалиброваны к этой взвешенной основе.

Методология расчета X

50. Самый простой способ расчета «X» заключался бы в оценке записей с «предостерегающим знаком», используя данные выборки. Но это невозможно сделать, поскольку эти записи определяются административной обстановкой и эта информация, безусловно, не собирается в ходе переписи.

51. Стратегия будет заключаться в определении кластеров, исходя из социально-демографических переменных (например, граждане стран ЕС старше 65 лет, проживающие в средиземноморском регионе). В каждом кластере будут «предостерегающие» записи и записи со счетным фактором, равным 1. Каждая «предостерегающая» запись должна иметься только в одном кластере. Количество населения, включенного в кластер, не должно быть большим (около 5% всего населения), и должно быть столько кластеров, сколько нам необходимо, но в каждом из них должно быть небольшое количество населения.

⁶ Некоторое упрощенное приближение различных правил по присвоению записи предостережения:

- Если запись относится к гражданину Испании старше 13 лет, не имеющему соответствующего идентификационного номера
- Если запись имеет идентификационный номер, но отсутствует в базах данных полиции
- Если запись, которая относится к иностранцу, не меняется в течение длительного времени в PADRON.

52. Все «предостерегающие» записи, включенные в кластер, будут иметь один и тот же счетный фактор «X». «X» рассчитывается следующим образом:

Оцениваемый размер кластера (полученный из выборки) за минусом записей, включенных в кластер со счетным фактором, равным 1, деленный на количество «предостерегающих» записей, входящих в этот кластер.

V. Заключение

53. Методология, применяемая в Испании, обеспечивает возможность использования большей части информации, содержащейся в регистрах, и распространения детальной информации, основанной на относительно большой выборке.

54. Кроме того, она позволяет сократить бюджет и количество задействованного персонала и повышает качество проводимой работы на местах.

55. Помимо этого, такая альтернатива избавляет от различий между административными понятиями и статистическими переменными.

56. С другой стороны, она снижает детальность информации, которая может быть распространена, если сравнивать с предыдущими переписями. И, конечно, эта методология является абсолютно новой для Испании, поэтому приходится иметь дело с некоторыми незапланированными трудностями.
