Организация Объединенных Наций ECE/CES/GE.41/2019/15

Экономический и Социальный Совет

Distr.: General
10 July 2019
Russian
Original: English

Европейская экономическая комиссия

Конференция европейских статистиков

Группа экспертов по переписям населения и жилищного фонда

Двадцать первое совещание
Женева, 18–20 сентября 2019 года
Пункт 2 предварительной повестки дня
Результаты тестирования с точки зрения методологии, технологии, участия и других аспектов

Оценка по малым районам для исправления ошибок измерения в крупных регистрах населения с использованием переписи населения Израиля

Записка Центрального статистического бюро Израиля*

Резюме

Как и во многих других странах, в Израиле на национальном уровне имеется достаточно точный регистр населения, в котором зарегистрировано около 9 млн человек. Однако этот регистр гораздо менее точен в отношении малых территорий, причем средняя погрешность регистрации в случае таких территорий составляет около 13%. Основная причина неточностей на уровне территорий заключается в том, что люди, переезжающие в ту или иную территорию или покидающие ее, часто сообщают о смене адреса с опозданием. В настоящем документе рассматриваются способы корректировки оценок по малым территориям, полученных из регистра, с помощью обследования на основе регистра. В частности, в нем рассматриваются следующие вопросы:

a) как лучше всего формировать выборку?

b) как комбинировать информацию выборочного обследования с данными регистра?

c) как производить корректировку на непредставление ответов, которое не является случайным? и

d) каким образом оценивать среднеквадратичную погрешность результирующих оценок на основе данных переписи?

В нем представлены эмпирические иллюстрации, основанные на данных переписи 2008 года.

* Подготовлено Дэнни Пфефферманном, Дэном Бен-Хуром и Оливией Блум (см. также Pfeffermann et al, Statistics in Transition (SiT), Vol. 20, No. 1, Polish Statistical Association (PTS) and Statistics Poland (GUS)).

GE.19-11744 (R) 190719 250719

1911744

I. Введение

1. В настоящем документе предлагается новый метод проведения переписи, который сочетает использование обследования и административных данных. В нем рассматриваются альтернативные способы интеграции результатов обследования с административными данными для формирования единой переписной оценки в небольших географических районах с учетом погрешностей в обоих источниках данных и непредставления ответов, которое не является случайным (НОНС). Предлагаемый метод иллюстрируется с использованием данных переписи 2008 года в Израиле.

A. Описание последней переписи населения Израиля (2008 год)

2. Израиль располагает достаточно точным Центральным регистром населения (ЦРН); на страновом уровне он почти идеален. Однако в случае малых статистических районов ЦРН гораздо менее точен, и средняя погрешность регистрации составляет 13%, а 95-й процентиль – 40%. Израиль разделен примерно на 3 000 статистических районов, и по каждому из них в рамках переписи населения требуется такая информация, как численность населения и социально-экономическая информация. Основная причина неточностей в данных регистра на уровне района заключается в том, что люди, прибывающие в районы или покидающие них, часто сообщают о смене адреса с опозданием, а другие люди, заинтересованные в сохранении конкретного адреса (например, по причине налоговых льгот, школы, парковки и т. д.), не сообщают об изменении адреса, пока у них сохраняется такая заинтересованность. В 2008 году Центральное бюро переписи населения Израиля (ЦБПНИ) провело комплексную перепись, состоящую из регистра населения, данные которого были скорректированы с помощью оценок, полученных по двум выборкам по каждому району: полевой (районной) выборке лиц, проживающих в данном районе в день переписи, для оценки занижения в данных регистра (выборка «U») и выборке населения, зарегистрированного в этом же районе, для оценки завышения в данных регистра (выборка «O»). Выборка «U» использовалась также для сбора социально-экономической информации.

3. Окончательные оценки переписи были рассчитаны следующим образом: обозначим N_i истинное число лиц, проживающих в районе i в день переписи, и обозначим K_i число лиц, зарегистрированных в качестве проживающих в этом районе. Пусть $P_{i,L|R}$ означает отношение числа лиц, проживающих в районе i, к числу лиц, зарегистрированных в качестве проживающих в этом районе, а $P_{i,L|R}$ означает отношение числа лиц, зарегистрированных в районе i, к числу лиц, проживающих в данном районе. Тогда

$$N_i \times p_{i,R|L} = K_i \times p_{i,L|R} \Rightarrow \hat{N}_i = K_i \times \frac{\hat{p}_{i,L|R}}{\hat{p}_{i,R|L}} . \quad (1)$$

4. При использовании разложения Тейлора условная (выборочная) дисперсия $\hat{N}_i$ может быть аппроксимирована следующим образом:

$$\mathrm{Var}(\hat{N}_i \mid K_i) \cong K_i^2 \left[\frac{Var(\hat{p}_{i,L|K})}{[E(\hat{p}_{i,K|L})]^2} + \frac{[E(\hat{p}_{i,L|K})]^2}{[E(\hat{p}_{i,K|L})]^4} \times Var(\hat{p}_{i,K|L}) \right] . \quad (2)$$

5. За последнее десятилетие Израиль, как и многие другие страны, осуществил ускоренный процесс перехода к использованию административных данных для подготовки официальной статистики в целом и, в частности, расширил свои возможности по использованию административных данных для целей переписи населения. В результате методология переписи населения 2020 года в Израиле предусматривает использование новых источников данных, и впервые

геодемографический административный файл (ГДАФ) послужит основой выборки, которая будет использоваться для корректировки административных данных по малым статистическим районам.

6. Модификация планируемой методологии стала возможна благодаря двум ключевым фактам: a) надлежащая регистрация въезда в страну и выезда из нее; b) все люди в стране находятся на административном учете; граждане регистрируются в ЦРН, а иностранцы регистрируются в функциональных регистрах, таких как перечни разрешений на работу и виз. В будущем можно ожидать концептуального и практического шага в направлении проведения полностью административных переписей населения, хотя, к сожалению, это пока невозможно сделать в отношении следующей переписи населения Израиля в 2021 году, контрольная дата которой установлена на 31 декабря 2020 года.

B. Новый метод, который планируется использовать для проведения следующей переписи населения в Израиле

7. Для проведения переписи 2021 года планируется использовать другой метод, который, как ожидается, поможет ЦБПНИ в переходе в будущем к полностью административной переписи. Эта перепись будет объединять информацию, полученную из единой выборки, взятой из ГДАФ, с информацией, имеющейся в регистре и других административных файлах, главным образом для корректировки показателей, полученных из ГДАФ. По этой выборке будет собрана геодемографическая информация обо всех членах домохозяйств в день проведения переписи, а также социально-экономическая информация. Информацию планируется получать через Интернет, затем по телефону от тех лиц, которые не представили ответы через Интернет, а в случае отсутствия ответов с использованием любого из этих двух способов – путем личного опроса.

8. Прямые оценки, полученные по этой выборке, будут улучшены с помощью модели оценки Фэя-Херриота (F-H) с использованием соответствующей ковариационной информации, известной на уровне района, такой как число зданий и общий объем всех зданий в районе, причем объем определяется как площадь крыши здания, помноженная на его высоту. Другие ковариаты будут использоваться для оценки представляющих интерес социально-экономических показателей района.

9. Для оценки численности населения района модель F-H будет объединена с соответствующим подсчетом ГДАФ в целях получения окончательной, составной оценки переписи (см. ниже).

II. Предлагаемая трехступенчатая модель переписной оценки

A. Оценка методом прямого подсчета (этап 1)

10. Обозначим число жителей в стране на день проведения переписи N и N_i – число жителей в районе i таким образом, чтобы $N = \sum_i N_i$. Пусть $p_i = N_i / N$ означает истинную долю жителей в ГДАФ, проживающих в районе i, а $\hat{p}_i$ означает соответствующий оценочный показатель по прямой выборке, например долю выборки в случае простой случайной выборки. (В настоящее время изучаются более эффективные планы выборки и прямых оценок). Наконец, обозначим $K \cong N$ размер ГДАФ в день переписи. Тогда прямая оценка численности населения в районе i будет иметь формулу $\hat{N}_i = K \times \hat{P}_i$. Условная выборочная дисперсия $\hat{N}_i$ будет выглядеть следующим образом: $Var_D(\hat{N}_i \mid K) = K^2 Var_D(\hat{P}_i) = \sigma_{Di}^2$.

B. «Усовершенствованная» оценка Фэя-Херриота (этап 2)

11. Стандартная модель Фэя-Херриота (F-H, 1979) имеет следующий вид:

$$\hat{N}_i = \alpha + \mathrm{x}_i'\beta + u_i + e_i \,, \qquad (3)$$

где $\hat{N}_i$ означает оценку с помощью прямой выборки, x_i представляет собой ковариаты района (число жилых зданий в районе и общий объем всех жилых зданий в эмпирических иллюстрациях); в настоящее время разрабатываются более эффективные ковариаты), u_i означает случайный эффект, а e_i означает погрешность выборки прямой оценки.

12. Согласно модели (3), усовершенствованный эмпирический наилучший линейный несмещенный предиктор (ЭНЛНП) истинного показателя выглядит следующим образом:

$$\hat{N}_{i,IMP} = \hat{\gamma}_i \hat{N}_i + (1-\hat{\gamma}_i)\mathrm{x}_i'\hat{\beta}; \quad \hat{\gamma}_\mathrm{i} = \hat{\sigma}_u^2(\hat{\sigma}_u^2 + \hat{\sigma}_{D_i}^2)^{-1} \,, \qquad (4)$$

где $\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_{D_i}^2$ – соответствующие оценки по выборке.

C. Окончательные оценки численности населения переписи

13. Окончательная оценка числа жителей в районе i будет представлять собой средневзвешенный показатель усовершенствованной оценки F-H (4) и показателя ГДАФ. Для этого предполагается, что $K_i \sim Possion(N_i) \Rightarrow Var(K_i) = N_i$. Таким образом, окончательная составная оценка рассчитывается по следующей формуле:

$$\hat{N}_{i,COM} = \hat{\alpha}_i K_i + (1-\hat{\alpha}_i)\hat{N}_{i,IMP}; \;\; \hat{\alpha}_\mathrm{i} = \frac{\hat{\sigma}_{i,FH}^2}{\hat{\sigma}_{i,FH}^2 + \hat{Var}(K_i)} \,. \qquad (5)$$

III. Альтернативные методы оценки численности населения переписи

A. Включение показателей регистра в число ковариатов в качестве постоянных чисел

14. Вместо расчета составной оценки (5) предлагается включать показатель ГДАФ в модель F-H (3) в качестве дополнительного ковариата. Подгонка этой модели «как есть» подразумевает использование известного показателя регистра без учета его возможной погрешности. Окончательной оценкой переписи в данном случае является оценка F-H.

B. Учет погрешностей регистра

15. При использовании метода Ибарра и Лора (Ybarra and Lohr (2008)) показатель ГДАФ добавляется в набор ковариатов, но его возможная погрешность измерения учитывается следующим образом $K_i \sim N(N_i, Var(K_i))$. Обозначим $\tilde{\mathrm{x}}_i = (\mathrm{x}_i', K_i)$. Предполагая, что все остальные ковариаты измерены без погрешности.

$$C_i = Var(\tilde{x}_i) = \begin{bmatrix} 0 \ldots 0, & \ldots, & 0 \\ 0 \ldots 0, & \ldots, & 0 \\ \ldots & , & \ldots \\ \ldots & , & \ldots \\ \ldots & , & \ldots \\ 0 \ldots 0, & \ldots, & V(K_i) \end{bmatrix}, \text{ и}$$

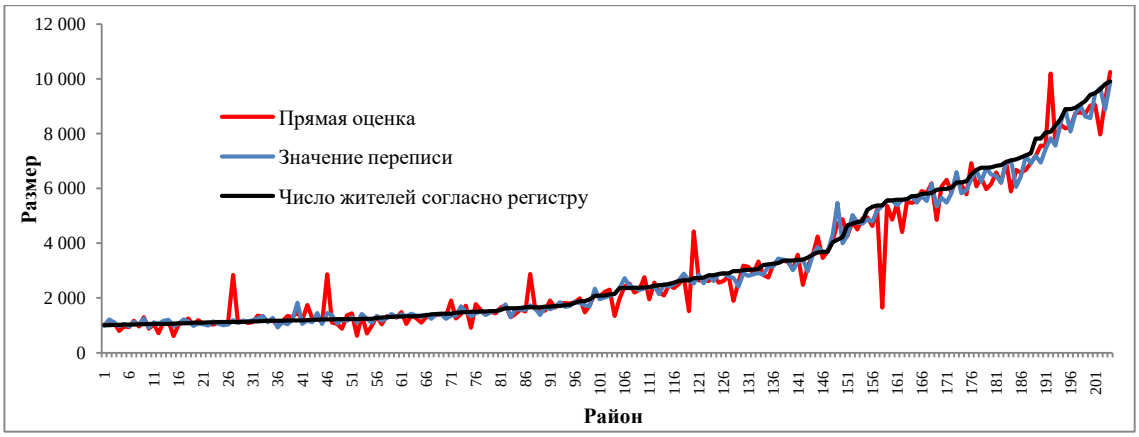
$$\hat{N}_{i,YL} = \hat{\delta}_i \hat{N}_i + (1-\hat{\delta}_i)\tilde{x}_i'\hat{\beta}; \quad \hat{\delta}_i = \frac{\hat{\sigma}_u^2 + \hat{\beta}' C_i \hat{\beta}}{\hat{\sigma}_u^2 + \hat{\beta}' C_i \hat{\beta} + \hat{\sigma}_{Di}^2}. \qquad (6)$$

IV. Эмпирические иллюстрации

16. Для иллюстрации этого метода и его различных вариантов используем избыточную выборку (O) из центрального регистра населения для переписи 2008 года. Общий размер выборки составляет приблизительно 600 000 человек. Объектом анализа являются 205 районов размером 1 000–10 000 жителей, согласно оценке переписи 2008 года, поскольку размеры этих районов соответствуют размерам представляющих интерес статистических районов следующей переписи. Выборочная совокупность формировалась методом стратифицированной простой случайной выборки. Ковариатами, используемыми в этих моделях, являются число жилых зданий в районе и общий объем всех жилых зданий. Параметры модели F-H были оценены с помощью оценки по методу максимального правдоподобия с использованием процедуры «PROC mixed» в программном приложении «SAS», которая предполагает нормальное распределение случайных эффектов и погрешностей выборки. Оценки переписи 2008 года (на основе выборок «O» и «U») взяты в качестве истинных значений (на диаграммах ниже они обозначены как «значения переписи»).

Диаграмма I

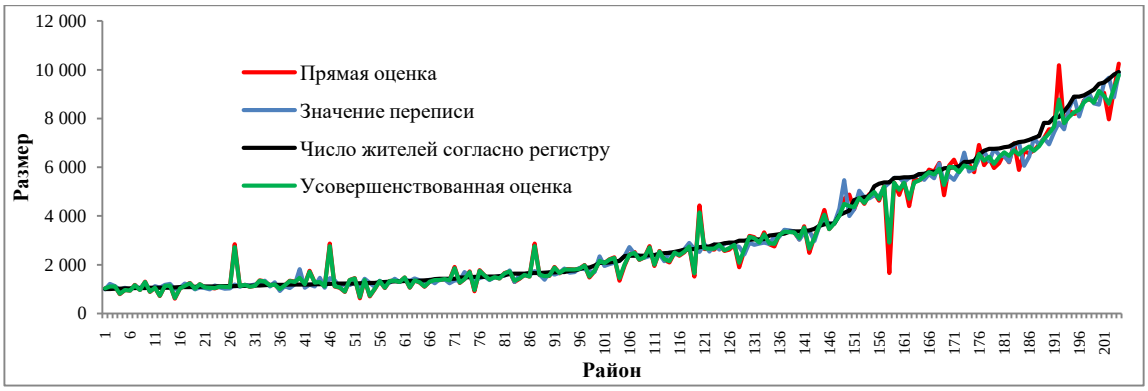
Прямая оценка, значение переписи и число жителей согласно регистру в 205 малых районах, отсортированных по их размеру в регистре



17. Как видно из диаграммы, прямая оценка является несмещенной, но характеризуется значительной дисперсией.

Диаграмма II

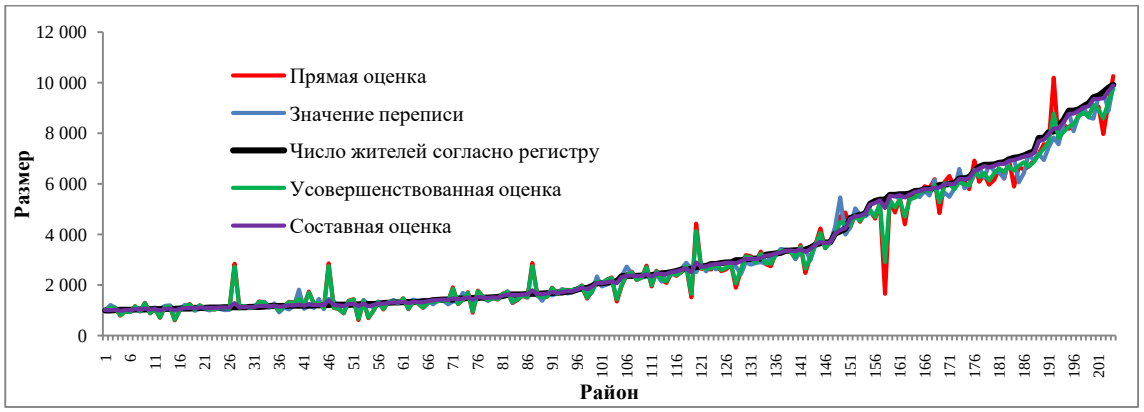
Прямая оценка, значение переписи, число жителей согласно регистру и усовершенствованная оценка (F-H)



18. Усовершенствованный метод оценки F-H позволяет лишь незначительно уменьшить дисперсию прямого метода оценки. В настоящее время ЦБПНИ ведет поиск более точных ковариатов. В частности, ожидается, что от электроэнергетической компании будет получен список всех жилых квартир и домов в Израиле, который должен значительно улучшить оценку F-H по сравнению с использованием только числа зданий.

Диаграмма III

Прямая оценка, значение переписи, число жителей согласно регистру, усовершенствованная оценка и составная оценка



19. Метод составной оценки, как представляется, позволяет получить более точную оценку истинных значений по сравнению с другими методами. В таблице 1 приводятся некоторые сводные статистические данные о результатах применения различных рассмотренных до настоящего времени методов оценки.

Таблица 1

Абсолютное относительное расстояние оценок от значений переписи I

Оценочный показатель	*Среднее значение*	*10-й процентиль*	*25-й процентиль*	*50-й процентиль*	*75-й процентиль*	*90-й процентиль*
Прямая оценка	0,1047	0,0101	0,0243	0,0556	0,1084	0,2202
Число жителей согласно регистру	0,0616	0,0010	0,0151	0,0507	0,0912	0,1344
Усовершенствованная оценка	0,0946	0,0112	0,0275	0,0573	0,0956	0,1959
Составная оценка	0,0598	0,0056	0,0189	0,0469	0,0834	0,1257

20. Наконец, на диаграмме IV и в таблице 2 представлены результаты, полученные при добавлении значений регистра в качестве дополнительного ковариата в модель F-H, с учетом погрешности измерения (FH_WME) и без ее учета (FH_NME). В последнем случае σ_u^2 и β оцениваются по методу модифицированных наименьших квадратов (Ybarra and Lohr, 2008).

Диаграмма IV

Оценочные значения при добавлении значений регистра в ковариаты модели Фэя-Херриота, с учетом и без учета погрешности измерения

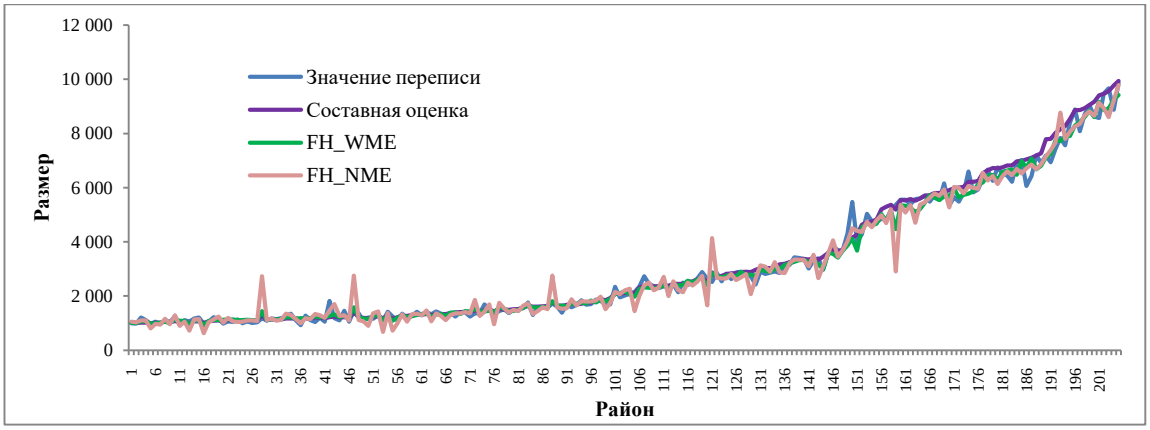


Таблица 2

Абсолютное относительное расстояние оценок от значений переписи II

Оценочный показатель	*Среднее значение*	*10-й процентиль*	*25-й процентиль*	*50-й процентиль*	*75-й процентиль*	*90-й процентиль*
Прямая оценка	0,1047	0,0101	0,0243	0,0556	0,1084	0,2202
Число жителей согласно регистру	0,0616	0,0010	0,0151	0,0507	0,0912	0,1344
Усовершенствованная оценка	0,0946	0,0112	0,0275	0,0573	0,0956	0,1959
FH_NME	0,0893	0,0100	0,0261	0,0540	0,0931	0,1877
Составная оценка	0,0598	0,0056	0,0189	0,0469	0,0834	0,1257
FH_WME	0,0603	0,0094	0,0227	0,0498	0,0793	0,1230

21. Как видно из приведенных примеров, если не учитывать погрешность измерения в регистре, то полученная оценка лишь незначительно лучше прямой оценки. Учет погрешности данных регистра весьма существенно улучшает результаты оценки модели F-H, но, как это ни удивительно, составной показатель дает несколько более точный результат, несмотря на использование ЭНЛНП в методе оценки Ибарра и Лора (Ybarra and Lohr (2008)). Хотя мы основываемся только на одном эмпирическом исследовании, возможное объяснение этого результата заключается в том, что в случае последнего метода данным регистра и другим (фиксированным) ковариатам присваивается одинаковый вес, тогда как составная оценка является более гибкой, позволяя использовать разные веса для данных регистра и других ковариатов. Для подтверждения этого результата необходимы дальнейшие теоретические исследования и эмпирические иллюстрации.

V. Учет непредставления ответов, которое не является случайным (НОНС)

22. Для оценки вероятности представления ответов в малых районах Сверчков и Пфефферманн (Sverchkov and Pfeffermann (2018)) предлагают метод, в котором используется принцип отсутствующей информации Орчарда и Вудбери (Orchard and Woodbury (1972)). Основная идея заключается в следующем: сначала необходимо определить вероятность, которая была бы получена, если бы отсутствующие конечные значения были бы также известны в отношении и тех респондентов, которые не представили ответы. Однако, поскольку отсутствующие значения на практике неизвестны, следует заменить вероятность на ожидаемое значение в отношении распределения отсутствующих результатов, учитывая все наблюдаемые данные. Последнее распределение получают из распределения наблюдаемых результатов в соответствии с наблюдаемыми значениями. См. работу Sverchkov and Pfeffermann (2018), в которой описывается взаимосвязь между распределениями наблюдаемых и отсутствующих результатов для заданных ковариатов и вероятностей представления ответов.

23. Было бы идеально показать, как этот метод работает при оценке истинного числа лиц, проживающих в каждом районе в день переписи, но эта информация практически неизвестна в отношении тестовых данных (используемая до сих пор выборка «О»). Поэтому ниже приводится иллюстрация эффективности этого метода для прогнозирования истинного числа разведенных лиц, зарегистрированных в каждом районе. Выборка «О» взята из центрального регистра населения, и поэтому истинное число разведенных лиц, зарегистрированных в каждом районе, известно.

24. Условимся, что итоговая переменная y_{ij} равна 1, если лицо j, зарегистрированное в районе i, разведено, и 0 в противном случае. Пусть показатель представления ответа R_{ij} имеет значение 1, если лицо j, зарегистрированное в районе i, представляет ответ, и 0 в противном случае. Анализ проводится только в отношении лиц в возрасте 20 лет и старше. Модель, использовавшаяся в отношении результатов представляющих ответы единиц, и модель, предложенная для вероятностей представления ответов, описаны уравнениями (7) и (8). Ковариаты, используемые для данной иллюстрации, приведены в таблице 3.

$$\Pr(y_{ij}=1\,|\,\mathrm{x}_{ij},u_i,R_{ij}=1)=\frac{\exp(\beta_0+\mathrm{x}'_{ij}\beta+u_i)}{1+\exp(\beta_0+\mathrm{x}'_{ij}\beta+u_i)};\quad u_i\sim N(0,\sigma_u^2) \qquad (7)$$

$$\Pr(R_{ij}=1\,|\,y_{ij},\mathrm{x}_{ij},\mathrm{u}_i;\gamma)=\frac{\exp(\gamma_0+\mathrm{x}'_{ij}\gamma+\gamma_y y_{ij})}{1+\exp(\gamma_0+\mathrm{x}'_{ij}\gamma+\gamma_y y_{ij})}. \qquad (8)$$

25. Очевидно, что для $\gamma_y\neq 0$ уравнение (8) дает информативный механизм представления ответов. Во-первых, вводится $\gamma_y=0$, т. е. предполагается, что статус «в разводе» не влияет на вероятность получения ответа, что соответствует предположению о случайном непредставлении ответа (СНО). Это осуществляется путем исключения семейного положения y_{ij} из модели представления ответов (8). Результаты приводятся в таблицах 3 и 4.

26. В таблице 3 приведены отношения шансов оцениваемой логистической модели вероятностей представления ответа для данного случая. Как и ожидалось, отношение шансов представления ответа увеличивается по мере увеличения числа телефонов, принадлежащих административной семье, и аналогично в отношении размера административной семьи. Возрастной группой с наименьшей вероятностью представления ответа является группа 30–39 лет (отношение шансов = 0,87), и в случае лиц, родившихся в Израиле, отношение шансов представления ответа намного выше, чем в случае лиц, родившихся за рубежом.

Таблица 3
Отношения шансов оценочной логистической модели вероятностей представления ответа при СНО

Переменная	*Отношение шансов в случае СНО*
Число телефонов на семью	1,70
Размер административной семьи	1,15
20–29 лет	0,98
30–39 лет	0,87
40+ лет	1,00
Евреи	1,04
Другие	1,00
Родился (родилась) в Израиле	1,27
Другие	1,00

27. В таблице 4 показано распределение оценочных вероятностей представления ответа согласно модели таблицы 3.

Таблица 4
Распределение оценочных вероятностей представления ответов согласно модели таблицы 3

Семейное положение	*Среднее значение*	*5-й процентиль*	*25-й процентиль*	*75-й процентиль*
Другие	0,815	0,489	0,822	0,885
В разводе	0,742	0,359	0,683	0,843
Итого	0,812	0,487	0,819	0,885

28. Из таблицы 4 видно, что предположение $\gamma_y = 0$ неверно. Вероятность представления ответа разведенными лицами значительно ниже, чем в случае других лиц. Следовательно, вероятности представления ответа были оценены путем включения бинарной переменной «в разводе» в качестве дополнительной объясняющей переменной.

Таблица 5
Отношения шансов оценочной логистической модели вероятностей представления ответов с учетом НОНС

Переменная	*Отношение шансов в случае СНО*	*Отношение шансов в случае НОНС*
Число телефонов на семью	1,70	1,83
Размер административной семьи	1,15	1,11
20–29 лет	0,98	0,95
30–39 лет	0,87	0,86
Другой возраст	1,00	1,00
Евреи	1,04	1,05
Другие	1,00	1,00
Родился (родилась) в Израиле	1,27	1,25
Другие	1,00	1,00
В разводе	–	0,531

29. Из таблицы 5 видно, что отношение шансов представления ответа разведенными лицами примерно вдвое меньше, чем в случае других лиц, что соответствует результатам таблицы 6. Интересно, что отношения шансов других ковариатов очень похожи на отношения шансов, полученные при допущении СНО.

30. Оценочные модели в таблицах 3 и 5 позволяют оценить вероятность представления ответа для каждого респондента в выборке. Используя представление ответа в качестве дополнительного этапа выборки, оценочные вероятности представления ответа можно применить для прогнозирования истинных средних значений целевой переменной по району (для данной иллюстрации – это доля разведенных лиц) с использованием стандартной теории выборки, например, с использованием приблизительно несмещенной по выборке оценки.

$$\hat{\bar{Y}}_i^{HB} = \sum_{j,(i,j)\in R} (y_{ij} / \tilde{\pi}_{j|i}) / \sum_{j,(i,j)\in R} (1/\tilde{\pi}_{j|i});\ \tilde{\pi}_{j|i} = \pi_{j|i}\,\hat{p}_r(y_{ij}, x_{ij}; \hat{\gamma}) \quad , \qquad (9)$$

где $\pi_{j|i}$ обозначает вероятность попадания в выборку. Сверчков и Пфефферманн в своей работе 2018 года также выводят эмпирический наилучший предиктор в рамках моделей (7) и (8), но в настоящем документе этот предиктор не рассматривается.

31. На диаграмме V и в таблицах 6 и 7 приводится сравнение эффективности следующих трех предикторов истинной доли разведенных лиц в различных районах: доля разведенных лиц в наблюдаемой выборке без учета непредставления ответов (далее по тексту – прямая оценка); оценка, полученная при допущении CHO, и оценка, полученная при допущении HOHC (уравнение 8).

Диаграмма V

Процентная доля разведенных лиц в районах: истинное значение, прямая оценка и оценки, полученные при допущении CHO и HOHC

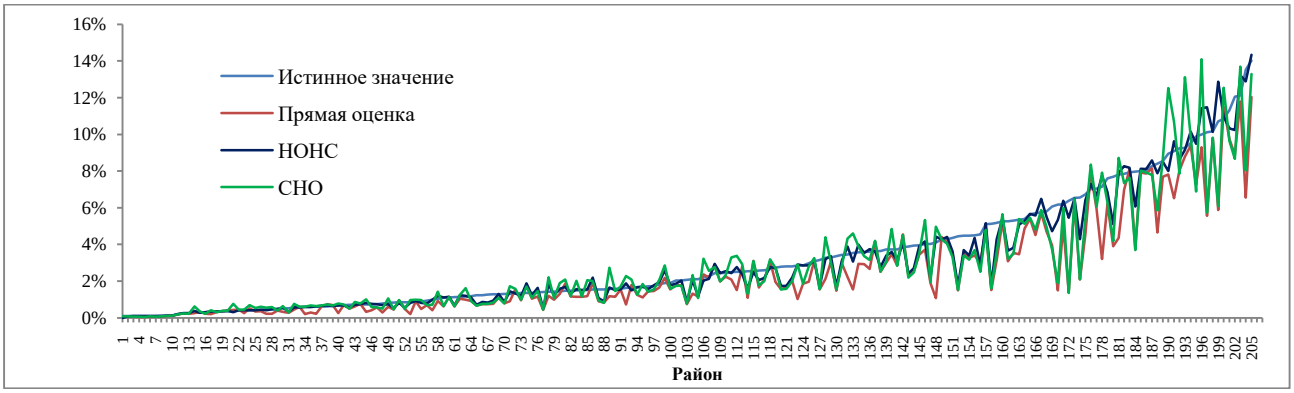


Таблица 6

Разница между истинными значениями и оценками по всем районам

Оценка	*Среднее значение*	*10-й процентиль*	*25-й процентиль*	*50-й процентиль*	*75-й процентиль*	*90-й процентиль*
Прямая оценка	0,0075	–0,0005	0,0006	0,0036	0,0099	0,0211
CHO	0,0033	–0,0077	–0,0018	0,0004	0,0057	0,0168
HOHC	0,0019	–0,0027	–0,0004	0,0001	0,0032	0,0094

Таблица 7

Абсолютное относительное расстояние оценок от истинных значений

Оценка	*Среднее значение*	*10-й процентиль*	*25-й процентиль*	*50-й процентиль*	*75-й процентиль*	*90-й процентиль*
Прямая оценка	0,270	0,042	0,121	0,233	0,406	0,551
CHO	0,256	0,032	0,113	0,216	0,379	0,472
HOHC	0,118	0,004	0,022	0,055	0,156	0,362

32. Как видно из диаграммы V и таблиц 6 и 7, оценки, полученные при учете НОНС, имеют наименьшее отклонение и наименьшее абсолютное относительное расстояние от истинных значений. Прямые оценки, которые не учитывают непредставление ответов, имеют большое отклонение и большое относительное расстояние от истинных значений.

VI. Оценка среднеквадратичной погрешности при НОНС

33. Как и в любой публикации официальной статистики, необходимо не только опубликовать оценки численности и другие социально-демографические параметры, но и оценивать их точность. Один из наборов оценочных методов (но определенно не единственный) – это оценка по каждому району среднеквадратичной погрешности (СКП) окончательного оценочного показателя. Это выглядит относительно просто, если в качестве окончательных оценок использовать прямые оценки по выборке и если ответы поступят от всех включенных в выборку единиц, но этого, очевидно, не произойдет. Оценка СКП более сложна при использовании оценочной модели Фэя-Херриота и учете НОНС и еще более сложна при использовании составной оценки, описанной в разделе II.

34. Сверчков и Пфефферманн в своей работе 2018 года предлагают для оценивания СКП оценок по малым районам с учетом НОНС метод «бутстрэп», который учитывает случайные процессы, предположительно формирующие значения численности населению, а также процессы формирования выборки и представления ответов. Это означает, что параметры целевого района (истинная доля лиц, проживающих в конкретном районе в день переписи, в числе всех лиц, зарегистрированных в ЦРН) рассматриваются как случайные, что отличается от классических выборочных обследований, в рамках которых конечные значения численности населения и, следовательно, целевые параметры рассматриваются как фиксированные значения. Пользователи оценочных данных выборочных обследований (официальной статистики) знакомы с такими методами измерения погрешностей, которые учитывают только вариативность, обусловленную случайностью формирования выборки (известной как рандомизированное распределение), а также непредставление ответов. Другими словами, пользователи привыкли к оценкам выборочной (рандомизированной) СКП (именуемой далее ВСКП) в рамках всех возможных вариантов выборки, когда значения численности населения переменных обследования (и, следовательно, значения целевых параметров) остаются неизменными. Оценка и публикация значений ВСКП (или ее квадратного корня) являются общепринятой практикой в национальных статистических управлениях во всем мире.

35. В одной из последних статей Пфефферманн и Бен-Гур (Pfeffermann and Ben-Hur (2019)) предлагают новую процедуру оценивания ВСКП основанных на модели предикторов по малым районам, которая хорошо зарекомендовала себя в обширном имитационном исследовании и превосходит другие методики оценки ВСКП, предложенные в научной литературе. В настоящее время применение этой процедуры расширяется на оценивание ВСКП предлагаемых составных оценочных показателей в целях учета, в частности, неизбежного НОНС и использования составного оценочного показателя, который сочетает оценку обследования с данными административного регистра населения.

VII. Заключительные замечания

36. В настоящем документе рассматривается новый метод проведения переписи, который сочетает использование оценок выборочного обследования и административных данных. Одним из основных преимуществ этого метода является то, что он не требует проведения личных опросов, за исключением тех случаев, когда респонденты не представляют ответа. Израиль по-прежнему не располагает достаточно надежным регистром жилых помещений, и использование полевого выборочного обследования требует предварительного получения перечня всех единиц жилья в выборке клеток по каждому статистическому району, что является довольно сложным с точки зрения логистики и очень дорогостоящим процессом. Также необходимо удостовериться, что в каждой из квартир кто-либо проживает.

37. В соответствии с новым методом из ГДАФ извлекается единая выборка лиц, которая считается в целом точной на национальном уровне, за исключением некоторых небольших «выпадающих» подгрупп населения, таких как незаконные иммигранты. Рассматриваются альтернативные пути интеграции информации, полученной в ходе обследования, с данными ГДАФ для составления единой окончательной оценки переписи, учитывающей погрешность выборки обследования и ошибки в адресах в ГДАФ. Предложена также простая описательная процедура проверки информативности отсутствующих выборочных данных и способ учета НОНС. Все вышеперечисленные темы иллюстрируются с использованием реальных эмпирических данных.

38. В настоящее время в двух статистических регионах Израиля планируется провести в следующем году пробную перепись населения, которая, как мы надеемся, предоставит ЦБПНИ еще одну возможность для проверки с помощью более актуальных данных рассмотренных в настоящей статье методик.

39. Более глубокое изучение административных источников информации открывает новые возможности с точки зрения процесса и результатов переписи. Использование данных о конкретной группе населения в рамках известной генеральной совокупности подразумевает существенное изменение концепции переписи. Границы района становятся не основным физическим аспектом переписи, а в некотором смысле виртуальным аспектом. Следует дополнительно изучить теоретические и социально-экономические последствия этих изменений, а также их влияние на разработку политики.

Справочная литература

Blum, O. (2018). From physical area to virtual lists: Toward an administrative census in Israel. UNECE group of experts on population and housing censuses. Geneva.

Fay, R. E. and Herriot, R. A. (1979). Estimation of income from small places: An application of James Stein procedures to census data. Journal of the American Statistical Association, 74, 269–277.

Orchard, T., and Woodbury, M.A. (1972). A missing information principle: theory and application. Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, 1, 697–715.

Pfeffermann, D. and Ben-Hur, D. (2018). Estimation of Randomization Mean Square Error in Small Area Estimation. International Statistical Review, 0, 0, 1–19 doi:10.1111/insr.12289.

Sverchkov, M., and Pfeffermann, D. (2018), "Small area estimation under informative sampling and not missing at random non-response". Journal of the Royal Statistical Society, Series A, 181, 981–1008.

Ybarra, L.M.R., and Lohr, S.L. (2008), "Small area estimation when auxiliary is measured with error", Biometrika, 95, 919–931.