УДК 311.2

ПЕРВИЧНАЯ ОБРАБОТКА ДАННЫХ ГЕОГРАФИЧЕСКИХ НАИМЕНОВАНИЙ В РАМКАХ ВЫБОРОЧНЫХ ОБСЛЕДОВАНИЙ МИГРАЦИОННЫХ ПЕРЕМЕЩЕНИЙ

Шарепина Е. А.

Аспирант, Преподаватель: Институт демографии имени А.Г. Вишневского

Кафедра демографии НИУ ВШЭ,

Ekaterinaandreeva96@yandex.ru

Родина О.А.,

студентка магистерской программы «Демография» НИУ ВШЭ,

oliarodina62@gmail.com

Новиков А. К.,

студент бакалаврской программы «Социология» НИУ ВШЭ,

schoolnovikov@yandex.ru

Аннотация: В статье рассматриваются методические подходы к первичной обработке данных направлений миграционных перемещений в рамках анкетного опроса. Проанализирована актуальность разработки методических рекомендации и специфика данных опроса на миграционную тематику. Показаны способы устранения ошибок и заполнения пропусков в переменных, содержащих географические наименования.
Ключевые слова: анкетный опрос, направления миграции, пропуски в данных географических наименований, очистка данных

THE SPECIFICS OF DATA CLEANING GEOGRAPHIC PLACE NAMES IN MIGRATION SURVEYS

Sharepina E. A.,

Rodina O.A.,

Novikov A.K.

Abstract: The methodological approaches for cleaning survey data of migration patterns were presented. This paper shows the importance of producing methodological recommendations and the specifics of data cleaning geographic place names in migration surveys. As a result, data cleaning problems and the approaches in handling errors and missing data in geographic place names were determined.
Keywords: surveys, migration patterns, missing data, spatial dimensions of migration, data cleaning

Одной из ключевых сложностей реализации исследований внутрироссийской межрегиональной и внутрирегиональной миграции

являются ограниченные возможности существующей статистики миграции [4, 5, 9]. Данные текущего учета и выборочные обследования Росстата не позволяют сколько-нибудь точно оценить миграционные биографии: направления и объемы, разные виды миграционных перемещений (маятниковую, вахтовую, временную миграции и др.), связь с социально-демографическими характеристиками мигрантов. В то же время понимание миграционных тенденций, миграционных траекторий представляется актуальным для понимания развития рынка труда, миграционных связей, для осуществления прогнозов социально-экономического положения региона [2, 7]. Все это обосновывает актуальность проведения выборочных обследований миграционных перемещений.

В российском исследовательском поле существует ряд работ, посвященных сбору и первичной обработке данных анкетных опросов в целом [1, 3, 6], однако они не учитывают специфику миграционных исследований. Нередки ошибки в методике концептуализации и сборе данных о миграции, в критериях учета мигрантов в разных источниках, смешение разных категорий миграционных перемещений, часто это связывается со скудностью учебно-методической литературы [8].

Фокус данных тезисов будет остановлен на проблемах и методологических приемах, позволяющих упростить первичную обработку данных направлений миграционных перемещений. Приведение к единому верному написанию топонимов (географических названий) важно для осуществления статистического анализа. Важнейшим принципом обработки данных является единообразность их написания.

Используемые данные были собраны методом уличного стихийного анкетного опроса в рамках экспедиционных выездов, посвященных анализу миграционных процессов в республике Хакасия (общее количество анкет - 1303), Республике Коми и Ямало-Ненецком АО (1012 анкет) в 2022 году. Выборка квотная, отражает половозрастной состав. Сбор данных осуществлялся с помощью программного обеспечения Arcgis Survey123.

Анкета включает 18 вопросов, среднее время заполнения анкеты 7 минут. Всего в базу были включены 34 переменных, содержащих географические названия.

Первый тип ошибок в названии населенных пунктов - это технические или грамматические ошибки (опечатки, лишние пробелы, непечатные символы, латынь, отсутствие или лишний дефис, кавычки, разный регистр букв и т.д.). Для данного типа ошибок необходимо специальным оператором в базе данных вывести все введенные топонимы и исправить в них опечатки. Одним из способов быстрого поиска опечаток является создание сводной таблицы в Excel - в которую включаются в алфавитном порядке все возможные варианты топонимов, или использование функции Excel «=УНИК», также выводящей все возможные варианта заполнения ячейки. Таким образом, все возможные варианты названия с ошибками наглядно предстают перед оператором базы данных. Далее ошибки могут быть исправлены с помощью функции «замена» или других функций Excel (например, «=СЖПРОБЕЛ»).

Второй распространенной проблемой является - неполный ответ респондента. Вариант ответа, когда респондент отказывается называть конкретный населенный пункт, а называет регион или район (не всегда административно-территориальную единицу) - например «Поволжье», «на Север», «средняя полоса», «деревня в Коми», «тундра». В этом случае мы восстанавливаем цепочку административно-территориальной принадлежности до максимально возможного (в случае Поволжья мы можем заполнить только страну - Россия), а исходный ответ переносим в категорию «другое».

Интересным и бросающим вызов оператору базы данных является тип ошибок, когда название населенного пункта услышано и записано интервьюером неверно и на первый взгляд не восстановимо. Здесь ключом к восстановлению топонима может быть насмотренность на часто встречающиеся населенные пункты - так и или иначе, ответы в анкете

помогают подсвечивать основные миграционные каналы и узнавать, например, село Шурышкары из всех возможных сочетаний букв, которыми оно могло быть занесено. Помимо насмотренности можно также опираться на другую информацию из анкеты и исследовательский запал - может быть, известен район и тогда можно проверить населенные пункты района или этот топоним в правильном написании встречается в другом поле анкеты.

В случае затруднений в определении административно-территориальной принадлежности населенного пункта (например, населенный пункт с таким названием существует в двух соседних районах или часто встречающихся регионах) рекомендуется заполнять поля как пользовательское пропущенное значение.

Для более быстрого восстановления административно-территориальной принадлежности можно создать общую базу введенных населенных пунктов со связями «страна - регион - район». Такая база позволит использовать функцию «=ВПР» в Excel для автоматизации заполнения пропусков.

После этапа выявления проблем в данных, непосредственной очистки и составления правил для такой очистки мы можем представить следующие результаты (Рис. 1). На графике представлено число пропущенных в отдельных переменных до обработки опроса и после. В качестве примеров представлены переменные, содержащие названия стран, регионов, районов и населенных пунктов, необходимые для восстановления миграционной истории.

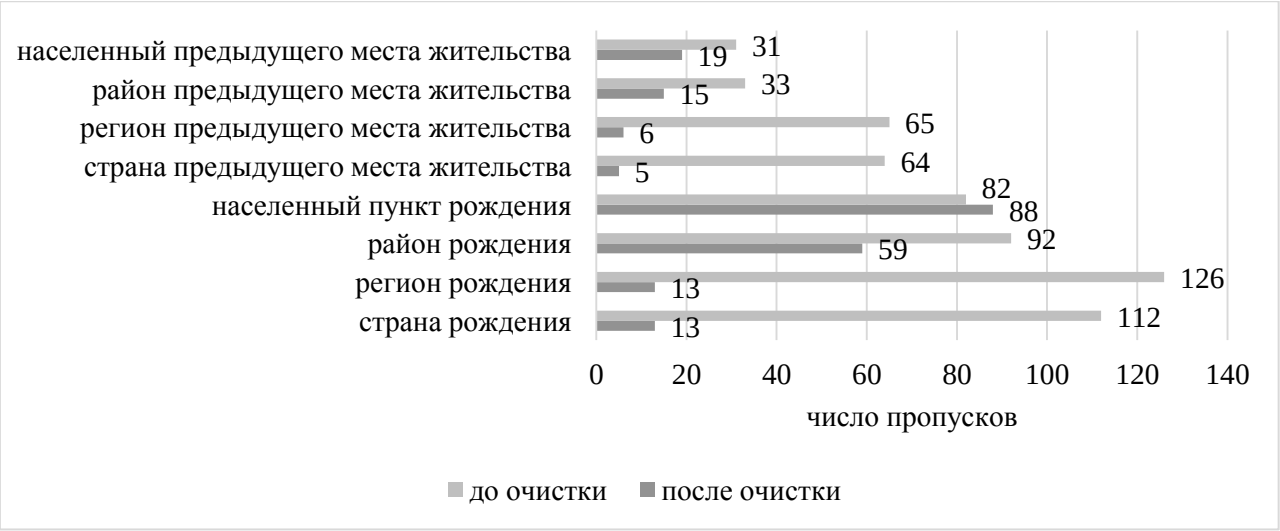

Рисунок 1. Результаты очистки базы данных по *Р*еспу*блике Коми и Ямало-Ненецкому АО,* числ*о* пропущенных по некоторым переменным
Источник: расчеты авторов.

Число пропусков значительно уменьшилось во всех представленных переменных, кроме переменной «населенный пункт рождения». Это произошло потому, что в ходе заполнения анкеты интервьюер часто использовал это поле в блоке вопросов места рождения для записи любого ответа, который давал ему респондент - зачастую, это были названия административных районов, регионов, которые, соответственно, были перенесены в соответствующую переменную - например, регион рождения, а в переменную «населенный пункт рождения» добавлялся код пользовательского пропущенного значения.

Библиографический список

1. Баранчиков А. И., Яковлев И. И., Клюева И. А. Использование методов очистки данных при реинжиниринге баз данных //Вестник Рязанского государственного радиотехнического университета. – 2021. – №. 76. – С. 35-41.
2. Брюханова Н.В., Сергиенко Л.С. Методология формирования системы мониторинга и прогнозирования баланса трудовых ресурсов как инструмента управления социально-экономическим развитием региона //Интеллектуальные ресурсы - региональному развитию. – 2014. – № 1. – С. 73-79. – EDN VRABJL.
3. Зангиева И.К. Проблема пропусков в социологических данных: смысл и подходы к решению //Социология: методология, методы, математическое моделирование (Социология:4М). – 2011. – Том. 0. – № 33. – С. 28-56.
4. Кашницкий И. С., Мкртчян Н. В., Лешуков О. В. Межрегиональная миграция молодежи в России: комплексный анализ демографической статистики //Вопросы образования. – 2016. – №. 3. – С. 169-203.
5. Мкртчян Н.В. Проблемы в статистике внутрироссийской миграции, порожденные изменением методики учета в 2011 г //Демографическое обозрение. – 2020. – Т. 7. – №. 1. – С. 83-99.

6. Татарова Г.Г., Бессокирная Г.П. Предметно-ориентированный подход к "борьбе" с пропущенными данными в типологическом анализе //Социологические исследования. – 2017. – №. 12. – С. 42-54.
7. Тимошенко Е.А. Роль и основные направления миграционной политики в Российской Федерации // Интеллектуальные ресурсы – региональному развитию. – 2015. – № 1. – С. 134-139. – EDN VNYTIJ.
8. Чудиновских О. С. О понимании статистики миграции //Вопросы статистики. – 2017. – №. 5. – С. 19-27.
9. Чудиновских О.С., Степанова А.В. О качестве федерального статистического наблюдения за миграционными процессами //Демографическое обозрение. – 2020. – Т. 7. – №. 1. – С. 54-82.