2/2022,
mart-aprel
(№ 00058)

СОВЕРШЕНСТВОВАНИЕ ОБРАБОТКИ И АНАЛИЗА ДАННЫХ, ПОЛУЧЕННЫХ ПРИ ПЕРЕПИСИ НАСЕЛЕНИЯ

Тула Нодирбек Баҳодир ўғли
Институт повышения квалификации кадров и статистических исследований при Государственном комитете Республики Узбекистан.
nodir_bek1990@mail.ru
Article DOI: 10.55439/EIT/vol10_iss2/a29

Аннотация
В статье описаны новые методы обработки данных, полученных при переписи населения, способы их применения в условиях Узбекистана. Проводится корреляционно-регрессионный анализ среднегодовой численности населения Республики Узбекистан, родившимися, умершими, числом браков и числом разводов. Приведена ранговая и рейтинговая оценка регионов Узбекистана по численности населения, родившимся и умершим за 2020 год.
Ключевые слова: перепись населения, обработка данных, корреляционно-регрессионный анализ, сканирование, анализ данных.

Annotatsiya
Maqolada aholini ro'yxatga olish natijalariga ko'ra olingan ma'lumotlarni qayta ishlashning yangi usullari, ularni O'zbekiston sharoitida qo'llash yo'llari yoritilgan. O‘zbekiston Respublikasi aholisining o‘rtacha yillik soni, tug‘ilishlar, o‘limlar, nikohlar soni va ajralishlar sonining korrelyatsion-regression tahlili o‘tkaziladi. O‘zbekiston viloyatlarining 2020-yil uchun aholi soni, tug‘ilish va o‘limlar soni bo‘yicha reyting va reyting bahosi berilgan.
Kalit so'zlar: aholini ro'yxatga olish, ma'lumotlarni qayta ishlash, korrelyatsiya-regressiya tahlili, skanerlash, ma'lumotlarni tahlil qilish.

Abstract
The article describes new methods of processing data obtained from the population census, ways of their application in the conditions of Uzbekistan. A correlation-regression analysis of the average annual population of the Republic of Uzbekistan, births, deaths, the number of marriages and the number of divorces is carried out. The ranking and rating assessment of the regions of Uzbekistan in terms of population, births and deaths for 2020 is given.
Key words: population census, data processing, correlation-regression analysis, scanning, data analysis.

Введение

Ввод данных и другие виды процессов в области обработки данных являются сферами, где информационные технологии играют важнейшую роль в рамках всей операции переписи. Многие страны, перешли от процессов ручного ввода данных на автоматизированные системы, на основе сканирования, оптического распознавания символов и оптического распознавания меток.

При переписи населения традиционно используются различные технологии ввода данных, такие, как ввод с клавиатуры и оптическое распознавание меток. Для ввода с клавиатуры требуется простое программное обеспечение и простые модели аппаратного обеспечения. Однако этот подход требует намного больше персонала по сравнению с автоматизированными методами ввода данных и, как следствие, больше времени. Эффективность затрат при использовании этого метода зависит от соотношения расходов на персонал и затрат на оборудование/разработку систем, необходимых при использовании других методов.

Анализ литературы по теме

Исследованиями в области обработки и анализа данных, полученных от переписи населения, занимались разные ученые. В частности, Бажанова О.В. предлагает оптимизацию первичной обработки данных переписи на основе экономико-математических моделей [1].

Манжула О.В. считает, что качественное проведение переписи населения в современных условиях возможно на основе использования современных ИКТ и методики выбора рациональных методов сбора и обработки первичной информации в условиях региональных различий [2].

Методологические исследования по разработке методов сбора статистической информации о населении и автоматизации процессов обработки данных переписи отражены в научных трудах таких ученых, как В.И. Борткевич, А.Г. Ковалевский, Н.С. Четвериков, М.А. Королев, В.В. Шуракова, В.П. Божко, Я.Л. Циписа и др. [3].

Методология исследования

В исследовании использовался корреляционно-регрессионный метод использовался для определения зависимости между родившимися и среднегодовой численностью населения, метод ранговой и рейтинговой применялся для оценки регионов Узбекистана по численности населения, родившимся и умершим.

Анализ и результаты

Оптическое распознавание меток может являться эффективным вариантом в тех случаях, когда переписной лист содержит только вопросы с помечаемыми вариантами ответов. Для обработки рукописных ответов требуются дополнительные средства ввода данных, автоматизированного кодирования. Однако на смену оптическому распознаванию меток пришли технологии интеллектуального распознавания символов.

Для большинства стран наиболее эффективным вариантом, является сочетание оцифровки изображений, интеллектуального распознавания символов, корректировки данных и автоматического кодирования. Пример такого процесса кратко описан ниже.

Переписные формуляры обрабатываются с помощью сканеров для получения изображений. Программное обеспечение распознавания используется для выявления помеченных вариантов ответов и преобразования рукописных ответов в текстовые величины. Для определения того, какие ответы обладают приемлемым качеством, а какие ответы требуют дополнительной корректировки или проверки, устанавливаются доверительные уровни.

Автоматизированная корректировка призвана снизить необходимость вмешательства оператора и, как правило, предусматривает использование таблиц

словарного поиска и средств контекстуального редактирования. Словари разрабатываются с учетом подлежащих обработке вопросов переписи. Таким образом, словарь для страны рождения будет содержать только названия стран. В свою очередь, подготовительная работа по созданию словарей терминов естественного языка значительно повысит эффективность кодирования.

Ручная корректировка может использоваться в отношении изображений, не поддающихся распознаванию. Этот подход является эффективным только в отношении тех вопросов, в случае которых существует высокая вероятность автоматического кодирования исправленных данных.

Автоматическое кодирование опирается на компьютеризованные алгоритмы для сопоставления введенных ответов с соответствующими индексами. Ответы, оставшиеся без соответствия, затем подвергаются компьютеризированному кодированию. В целях ограничения затрат и повышения качества ответы, не поддающиеся кодированию, должны анализироваться для нахождения схожих ответов. Эти ответы могут быть либо добавлены в индексы кодирования, либо вновь переданы для проведения автоматического кодирования или же может быть проведено групповое кодирование другой формы.

Вышеописанное сочетание технологии интеллектуального распознавания символов/автоматического кодирования/формирования изображений, является наиболее эффективным решением для многих стран. Благодаря автоматическому кодированию и использованию этих систем сокращаются потребности в персонале. Использование изображений значительно снижает потребность в передаче бумажных переписных листов, и было продемонстрировано, что использование изображений для отслеживания хода кодирования ответов, которые не поддаются автоматическому вводу, является намного более эффективным, чем использование бумажных формуляров.

Важно отметить, что эта методология открывает возможности для повышения качества данных. Она может гарантировать непротиворечивую обработку идентичных ответов. Однако качество автоматизированного ввода и кодирования требует тщательного контроля в ходе обработки для обеспечения того, чтобы система функционировала в соответствии с установленными спецификациями. Коэффициенты замещения символов должны тщательно отслеживаться, а критические вопросы или части вопросов (такие, как год рождения в сопоставлении с днем рождения) могут требовать применения более жестких доверительных интервалов, что в свою очередь предполагает более высокий уровень контроля и обеспечения качества по сравнению с другими полями или значениями. Организация процесса корректировки и кодирования может значительно повысить эффективность и точность процесса за счет направления результатов ответов на конкретные вопросы специализированным операторам или кодирования по блокам вопросов.

Необходимо разработать процедуры непрерывного обеспечения качества выходных результатов системы, такие, как ручное перекодирование изображений выборки ответов и сопоставление их с автоматически введенными и закодированными ответами. Это должно помочь обеспечить надлежащий баланс между качеством и затратами, включая сокращение объема ручной корректировки, и

избежать расходования ресурсов на меры, дающие незначительное повышение качества.

По этой причине крайне необходимо, чтобы даже в том случае, когда эти системы отданы на «аутсорсинг», организаторы переписи четко понимали зависимость «качество/затраты», присущую «аутсорсингу», реализованным в программном обеспечении интеллектуального распознавания символов /корректировки данных, их влияние на коэффициенты замещения и конечное качество переписных данных. Договор о внешнем подряде должен позволять оперативную корректировку этих параметров для удовлетворения требований переписных органов, касающихся качества и операционных характеристик.

Переписные организации должны изучить вопрос о том, в каком формате данные будут проходить через процесс обработки. Традиционно обработка результатов переписи проводится с использованием двумерных файлов, которые постепенно обновляются, и более ранняя версия файла остается для резервирования и восстановления. Обычно использовалась пакетная обработка, при которой дискретная группа формуляров (обычно с одного переписного участка) обрабатывалась вместе. При этом данные с переписных листов вводятся, редактируются и кодируются в качестве группы. Это обеспечивает более высокий уровень контроля за рабочей нагрузкой. Базы данных позволяют ведение и обработку информации на уровне индивидуального поля. Это обеспечивает более высокий уровень гибкости, поскольку после электронного ввода переписных данных их можно легко организовать для максимального повышения эффективности и качества обработки благодаря тому, что схожие ответы могут быть сразу же сгруппированы и закодированы вместе. Однако введение данных переписи в формате базы данных требует более сложных систем для управления работами и представления результатов. Необходимо также учитывать механизмы резервного копирования и восстановления.

Анализ стандартных таблиц на ранних этапах позволит устранить внутренние противоречия и обеспечит сопоставимость с данными из других источников. Это позволит выявить ошибки кодирования и степень распознавания данных. С целью выявления ошибок кодирования и чрезмерного использования родовых кодов необходимо проводить экспертизу профилей кодирования по отдельным операторам. Такие системы, как правило, требуют намного больше затрат по разработке и тестированию по сравнению с традиционной системой обработки результатов переписи.

Рисунок 1. Процесс ввода данных с использованием интеллектуальной системы распознавания символов.

Существует ряд факторов, которые необходимо учитывать и интегрировать в ходе разработки систем, таких как организация работы остальных процессов документооборота.

Высокая пропускная способность сети также имеет чрезвычайно большое значение с учетом большого числа и размера файлов, связанных с изображениями. Такие методы, как «стирание» формуляра, когда с окончательного изображения удаляется фиксированная справочная информация, способны значительно уменьшить размер файлов.

Такая возможность должна быть предусмотрена при разработке вопросника и протестирована во время печатания вопросников с целью проверки соответствия оптической плотности изображения параметрам удаления текста [7].

Еще одним важным аспектом в обработке полученных данных в ходе переписи населения, является подготовка высококвалифицированных специалистов, которые будут обрабатывать данные переписи. Даже при использовании сложных программ интерпретации, необходимо некоторое количество операторов для распознавания изображений, но, к сожалению, они не все могут быть высококвалифицированными. Следовательно, на этом этапе имеется два важных требования: проверка должна быть простой и быстрой, и операторы не должны вносить дополнительных ошибок. Решение, состоит в так называемой «массовой проверки»: изображения всех интерпретируемых символов были представлены в группе в соответствии с их значением. Если символ появляется в неправильной группе, верификатор отбирает его, и он будет автоматически помещен в поле, где возник этот символ, для исправления. Сначала цифры подвергаются массовой верификации, затем буквы и, наконец, поля с отметками.

Важным моментом, сокращающим количество ошибок, является эффективный дизайн переписного листа. Правильный дизайн переписного листа выполняет две задачи: облегчает респондентам заполнение форм и ведет к минимизации ручного труда при считывании информации. Поэтому хорошо, когда в разработке дизайна форм участвуют специалисты по сканированию, которые понимают, какие технические аспекты следует принимать во внимание. Переписные листы имели четкую планировку, достаточное пространство для ответов, большой размер шрифта и соответствующие разрывы страниц. Вопросы-фильтры, предназначенные для разных подгрупп и для переходов, были выделены цветом.

Номер серии был напечатан на всех страницах переписного листа. Уникальный идентификатор переписных листов помогал при вводе данных, особенно в том случае, если лист был по ошибке отсканирован более одного раза.

Во время печатания переписных листов была введена процедура, в соответствии с которой случайная выборка ежедневного тиража проверялась системой сканирования. Несмотря на то, что каждый переписной лист проходил этот прагматический тест, в конце операций было обнаружено несколько неувязок в ограниченном количестве напечатанных вопросников. В системе сканирования был предусмотрен обходной путь для решения таких вопросов. Всего было напечатано около 1,200,000 буклетов (6 отдельных вопросников) [5].

Когда используется система работы с изображениями, необходимо обучать переписчиков как правильно заполнять формы, для того чтобы система ICR могла

правильно распознавать текст, написанный от руки. INSTAT проводил специальное занятие во время обучения переписчиков о том, как правильно записывать ответы в полях формы, какой ручкой пользоваться и пр. Было потрачено много времени и усилий, чтобы обеспечить как можно более аккуратное заполнение форм и их возврат в хорошем состоянии.

Успех операций по вводу данных до определенной степени зависит от квалификации персонала, участвующего в процессе. INSTAT организовал обучение для повышения осведомленности персонала об их обязанностях и важности четкого выполнения работы. Отдельные занятия были проведены для операторов сканеров по эксплуатации оборудования и использованию программного обеспечения, для операторов ручного ввода, обучая их процедуре массовой верификации и другим процедурам ПО, для верификаторов, знакомя их с процедурой выделения несогласованностей и общей стратегией валидации.

После получения данных переписи можно составить рейтинг регионов Узбекистана по рождаемости, смертности, численности, а также плотности населения. Данные, по рейтинговой оценке, могут быть использованы при распределении государственных программ финансирования.

Таблица 1.

Ранговая и рейтинговая оценка регионов Узбекистана по численности населения, родившимся и умершим за 2020 год.

	Регионы Узбекистана	Ранг по численность населения	Ранг по родившимся	Ранг по умершим
1	Республика Каракалпакстан	10	11	11
2	Андижанская	4	4	5
3	Бухарская	9	10	9
4	Джизакская	4	12	12
5	Кашкадарьинская	3	3	6
6	Навоийская	13	13	13
7	Наманганская	6	5	7
8	Самаркандская	1	1	3
9	Сурхандарьинская	7	6	8
10	Сырдарьинская	14	14	14
11	Ташкентская	5	7	4
12	Ферганская	2	2	2
13	Хорезмская	11	9	10
14	г. Ташкент	8	8	1

Как видно из таблицы 1, большинство рейтингов регионов по 3 показателям незначительно отличаются друг от друга. Логично предположить, что чем больше численность населения, тем больше умерших и родившихся. Вместе с тем г. Ташкент по численности населения и родившимся стоит на 8 месте, но по количеству умерших на первом. Это можно объяснить тем, что в г. Ташкент проживает больше пожилых людей, а также большим количеством медицинских учреждений, в которых случаются смертельные случаи.

Таблица 2.

Статистические показатели регионов Узбекистана по родившимся, умершим, количеством браков и разводам за 2020 год.

Паказатели	Родившиеся	Умершие	Браки	Разводы
Ср.знач	60129,57143	12545,5	21,2	2,016
Дисперсия	647590050,8	27980863,54	73,28	0,886
Коэф.вар.	10769,90964	2230,350607	3,45	0,439

Для выявления зависимости между численностью постоянного населения, рождаемости и смертности построим таблицу коэффициентов корреляции.

Таблица 3.

Коэффициенты корреляции между среднегодовой численностью постоянного населения Республики Узбекистан, родившимися, умершими, числом браков и числом разводов[1].[8]

	Коэффициенты корреляции между среднегодовой численностью постоянного населения Республики Узбекистан			
	Родившимися	Умершими	Числом браков	Числом разводов
Республика Узбекистан	0,52	0,69	0,75	0,27
Республика Каракалпакстан	0,06	-0,52	0,39	0,64
Андижанская	0,60	0,83	0,69	0,63
Бухарская	0,39	0,70	0,66	0,63
Джизакская	0,45	0,71	0,82	0,66
Кашкадарьинская	0,60	0,86	0,86	0,65
Навоийская	0,27	0,15	0,61	-0,31
Наманганская	0,52	0,77	0,79	0,76
Самаркандская	0,58	0,57	0,80	0,62
Сурхандарьинская	0,56	0,71	0,83	0,49
Сырдарьинская	0,28	0,32	0,70	0,60
Ташкентская	0,31	0,52	0,67	-0,21
Ферганская	0,43	0,76	0,67	0,19
Хорезмская	0,44	0,68	0,64	0,70
г. Ташкент	0,82	-0,33	0,56	-0,38

По данным таблицы 3. можно отметить, что лишь в городе Ташкенте зависимость между рождаемостью и среднегодовой численностью населения сильная и прямая коэффициент корреляции составляет 0,82. В большинстве остальных регионах коэффициент корреляции хоть и прямой, но не столь существенный. Наименьший коэффициент корреляции между рождаемостью и среднегодовой численностью населения в Республике Каракалпакстан, что обусловлено сильными миграционными процессами.

[1] Составлено автором на основании данных Госкомстата Республики Узбекистан за 1991-2020 гг.

Таблица 4.

Регрессионная статистика зависимости между родившимися и среднегодовой численностью населения Республики Узбекистан

Регрессионная статистика	
Множественный R	0,51969329
R-квадрат	0,27008111
Нормированный R-квадрат	0,24401258
Стандартная ошибка	3388,39781
Наблюдения	30

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	118950543,8	118950544	10,36	0,0032471
Остаток	28	321474712,1	11481239,7		
Итого	29	440425255,9			

	Коэффи-циенты	*Стандартная ошибка*	*t-статис-тика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	12650,369	4499,46293	2,81	0,01	3433,64	21867,1
Переменная X 1	22,3015074	6,928598499	3,22	0,00	8,11	36,49

Из таблицы 4. видно, что существует связь между родившимися и среднегодовой численностью населения Республики Узбекистан. При этом коэффициент детерминации (R квадрат) составляет 0,27, а стандартная ошибка 3388,39781. В то же время нормированный R-квадрат равен 0,244, это означает, что количество родившихся влияет на среднегодовую численность населения лишь на 24,4%. Значимость модели по критерию Стьюдента для среднегодовой численности населения Республики Узбекистан равна 2,81, а по родившимся 3,22. Значение плотности вероятности распределения Стьюдента равно 0,01, критерий Фишера равен 10,36, а его значение 0,0032471, это означает что модель значима. Нижние и верхние границы доверительных интервалов для среднегодовой численности населения Республики Узбекистан вычисленные с 95% доверительной вероятностью, составляют 3433,64 и 21867,1 соответственно, для родившихся 8,11 и 36,49 соответственно.

Для более наглядного представления зависимости между родившимися и среднегодовой численностью населения Республики Узбекистан построим рисунок 2.

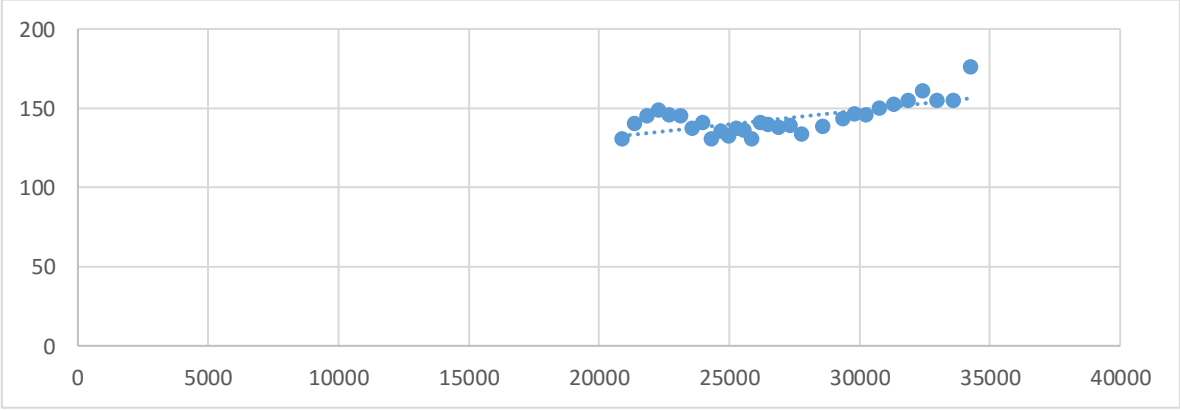

Рисунок 2. Зависимость между родившимися и среднегодовой численностью населения Республики Узбекистан

Из рисунка 2. видно, что зависимость между родившимися и среднегодовой численностью населения Республики Узбекистан, что является логичным, но в тоже время существуют другие факторы влияющие на численность населения в Узбекистане.

Таблица 5.

Регрессионная статистика зависимости между умершими и среднегодовой численностью населения Республики Узбекистан

Регрессионная статистика	
Множествен-ный R	0,686949
R-квадрат	0,471898
Нормированный R-квадрат	0,453038
Стандартная ошибка	7,429649
Наблюдения	30

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значи-мость F*
Регрессия	1	1381,10	1381,10	25,02	0,000028
Остаток	28	1545,59	55,20		
Итого	29	2926,69			

	Коэффи-циенты	*Стандартная ошибка*	*t-статис-тика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	95,53	9,65	9,90	0,00	75,76	115,30
Переменная X 1	0,00	0,00	5,00	0,00	0,00	0,00

Из таблицы 5 видно, что существует связь между умершими и среднегодовой численностью населения Республики Узбекистан. При этом коэффициент детерминации (R квадрат) составляет 0,47, а стандартная ошибка 7,429. В то же время нормированный R-квадрат равен 0,453, это означает, что количество родившихся влияет на среднегодовую численность населения лишь на 45,3%. Значимость модели по критерию Стьюдента для среднегодовой численности населения Республики Узбекистан равна 9,9, а по умершим 5,0. Значение плотности вероятности распределения Стьюдента равно 0,00, критерий Фишера равен 25,02, а его значение 0,000028, это означает что модель значима. Нижние и верхние границы доверительных интервалов для среднегодовой численности населения Республики Узбекистан вычисленные с 95% доверительной вероятностью, составляют 75,76 и 115,3 соответственно, для умерших 0,0 как для нижних, так и для верхних границ.

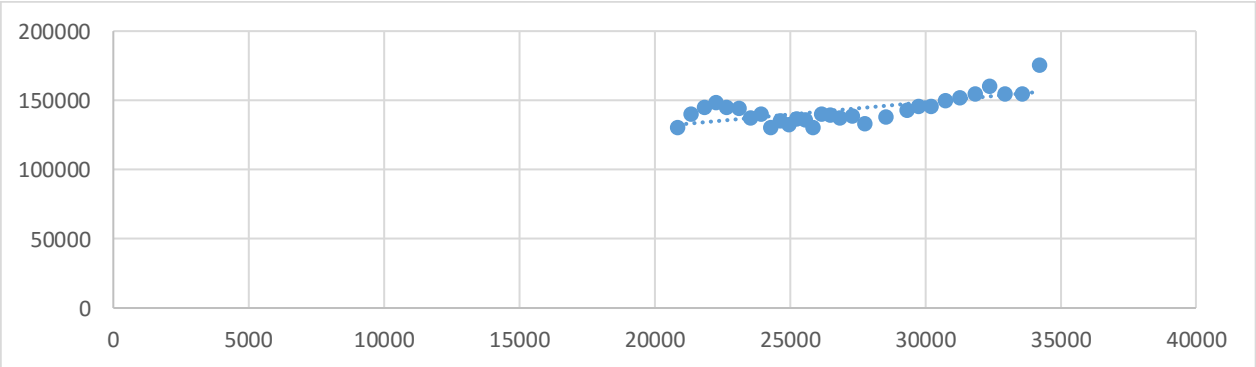

Рисунок 3. Зависимость между умершими и среднегодовой численностью населения Республики Узбекистан

Из рисунка 3. видно, что зависимость между умершими и среднегодовой численностью населения Республики Узбекистан прямая и это объясняется тем, что население Узбекистана постепенно входит в фазу «стационарного населения», выраженную постепенным повышением среднего возраста населения и доли населения старших возрастов. Проще говоря, чем больше становится численность населения, тем больше количество умерших.

Выводы и предложения

Таким образом, обработка данных, полученных при переписи населения, будет более эффективной с помощью автоматизированных систем, на основе сканирования, оптического распознавания символов и оптического распознавания меток в сочетании с технологией интеллектуального распознавания символов/автоматического кодирования/формирования изображений. В свою очередь анализ данных, полученных при переписи населения можно осуществить с помощью корреляционно-регрессионного метода. Можно сделать вывод, что сканирование является относительно дешевым способом существенного ускорения обработки данных. При численности населения Республики Узбекистан в 35 млн. человек, можно ожидать получение полного массива очищенных данных на несколько месяцев раньше, чем при ручном вводе, при привлечении небольшого количества сотрудников. Это приведет к гораздо более быстрому получению результатов переписи: ожидается, что окончательные таблицы будут сформированы через 9-10 месяцев после завершения работы переписчиков.

Список использованной литературы

1. Бажанова О.В., Моделирование процессов первичной обработки материалов сельскохозяйственной переписи. Дисс. к.э.н. М.: 2007.

2. Манжула О.В., Рационализация методов сбора и первичной обработки информации всероссийской переписи населения Дисс. к.э.н. М.: 2020.

3. Божко В.П., Лури А.В., Сычев Е.Б. Совершенствование процессов проведения статистических переписей и обследований. М.: МЭСИ, 2008.

4. Тула Н.Б. Совершенствование методологии переписи и учета занятости населения. Монография. Т.: 2021.

5. U.S. Census Bureau 2020 Census Operational Plan A New Design for the 21st Century / December 2018.

6. Bernard Baffour-Awah, Estimation of population totals from imperfect census, survey, and administrative records. University of Southampton, PhD Thesis. 2009.

7. Рекомендации конференции европейских статистиков по проведения переписей населения и жилищного фонда 2020 UCECE Нью-Йорк и Женева, 2015, с. 42-45.

8. www.stat.uz - официальный сайт Госкомстата РУз