

**Проект Технических требований на создание типовой модели производства
статистических оценок численности населения на основе данных сотовых
операторов в рамках ТМПСИ**

Листов 135

2025

Аннотация

Настоящий документ является проектом Технических требований на создание типовой модели производства статистических оценок численности населения на основе данных сотовых операторов в рамках ТМПСИ.

В документе описаны основные наборы требований, правил, стандартов и методик, которые могут быть адаптированы и внедрены национальными статистическими службами государств-участников СНГ для оценки численности населения с использованием больших данных операторов сотовой связи.

Документ разработан в соответствии с требованиями ГОСТ 3.1001-2011 «Единая система технологической документации».

Содержание

Определения, обозначения, сокращения.....	7
1. Общие положения	9
1.1 Наименование работ и основания для их проведения.....	9
1.2 Перечень документов, на основании которых ведется разработка материалов	9
1.3 Организации, участвующие в разработке материалов	10
1.4 Сроки выполнения работ.....	10
1.5 Цели выполнения работ.....	11
2 Описание целей разработки требований к типовой модели производства статистических оценок численности населения на основе данных сотовых операторов в рамках ТМПСИ	13
3 Требования к определению потребностей Заказчика.....	14
3.1 Требования к определению потребностей в информации, к проведению консультаций и подтверждению потребностей	14
3.1.1 Понимание бизнес-процессов производства статистической информации о численности населения	14
3.1.2 Ограничения и недостатки существующих процессов	17
3.1.3 Вопросы, требующие решения (направления развития).....	18
3.1.4 Требования к описанию бизнес-процессов для реализации направлений развития	19
3.2 Требования к установлению целей формирования материалов.....	20
3.2.1 Варианты целей использования данных сотовых операторов в процессе производства официальной статистики о численности и составе населения.....	20
3.2.2 Ограничивающие принципы (недопустимые цели использования данных сотовых операторов в процессе производства официальной статистики о численности и составе населения).....	23
3.2.3 Критерии выбора и оценки альтернативных источников данных	23
3.2.4 Требования к постановке целей использования АИД	24
3.3 Требования к определению концепций.....	24
3.3.1 Правовые и этические требования.....	25
3.3.2 Технические требования к данным и их обработке.....	26
3.3.3 Методологические требования	27
3.3.4 Ресурсные и организационные требования	27
3.4 Требования к проверке наличия данных.....	28

3.4.1	Требования к предлагаемым перспективным решениям	28
3.4.2	Требования к соответствию альтернативного источника данных потребностям статистической организации	31
3.4.3	Требования к разделению обязанностей между поставщиком данных, агрегатором данных и статистической организацией	34
3.4.4	Требования по выявлению неполноты охвата поставщиком больших данных территориальных единиц страны или отдельных групп населения	36
3.5	Требования к подготовке и представлению типовой бизнес-модели	37
3.5.1	Требования к принципам построения бизнес-модели	37
3.5.2	Требования к структуре затрат	38
3.5.3	Требования к расчету бюджета	39
4	Требования к проектированию типового бизнес-процесса	50
4.1	Требования к проектированию типовых выходных материалов	50
4.1.1	Требования к формату выходных материалов	50
4.1.2	Типовая дорожная карта реализации бизнес-модели	52
4.2	Требования к проектированию типовых описаний переменных	54
4.3	Требования к проектированию сбора данных	54
4.3.1	Требования к описанию типовых принципов построения сотовых сетей и событий, фиксируемых сетями операторов сотовой связи	54
4.3.2	Требования к разработке принципов определения необходимого и достаточного количества сотовых операторов учетом обеспечения полноты покрытия территории страны и охвата населения	56
4.3.3	Требования к типовым принципам обработки первичных данных сотовых сетей для определения местонахождения абонентов в требуемый временной интервал	61
4.4	Требования к проектированию обработки и анализа	67
4.5	Требования к проектированию производственных систем и процесса	70
5	Требования к построению типового бизнес-процесса	72
5.1	Требования к построению механизмов сбора данных	72
5.1.1	Требования к технической готовности операторов сотовой связи к сбору данных	73
5.1.2	Требования к агрегатору данных и его опыту реализации задач, связанных с расчетом численности населения на базе больших данных	75
5.2	Требования к построению компонентов обработки и анализа	76
5.2.1	Требования к описанию терминов и определений с целью составления типового технического задания для двух (или более)	

операторов сотовой связи на подготовку предварительных агрегатов данных	76
5.2.2 Требования к составу технического задания для сотовых операторов	79
5.3 Требования к компоновке производственных процессов (Никита).....	80
5.3.1 Требования к описанию типовых форматов данных, получаемых в результате предварительной агрегации данных каждым из двух (или более) операторов сотовой связи	80
5.3.2 Требования к описанию принципов и правил предварительной агрегации исходных данных каждого из двух (или более) операторов сотовой связи	82
5.4 Требования к проверке систем производства.....	83
5.4.1 Требования к типовым критериям валидации и оценки качества предварительных агрегатов	83
5.4.2 Требования к типовым способам проверки корректности данных сотовых операторов.....	83
5.5 Требования к проверке статистического бизнес-процесса	87
5.5.1 Техническое задание на выгрузку данных сотовыми операторами составляется согласно пункту 5.2.2.	87
5.5.2 Сбор, обработка и верификация статистических отчетов, предоставляемых операторами, в целях формирования на их основе агрегатов.....	94
5.5.3. Критерии оценки качества агрегированных данных	103
5.5.4. Получение и верификация агрегированных статистических отчетов	106
6 Требования к сбору данных	111
6.1 Требования к формированию генеральной совокупности и выборки	111
6.2 Требования к организации сбора	111
6.2.1 Требования к технологиям сбора предварительных агрегатов, полученных от двух (или более) операторов сотовой связи.....	111
6.2.2 Требования к технологиям агрегации данных, полученных от двух (или более) операторов сотовой связи	113
6.3 Требования к проведению сбора данных.....	114
6.3.1 Требования к типовым методикам агрегации данных от нескольких операторов.....	114

6.3.2	Требования к типовым способам «досчета» итоговых данных для полного охвата населения исследуемой территории с учетом:	119
6.4	Требования к завершению сбора данных	120
Приложение 1	123

Определения, обозначения, сокращения

В настоящем документе применяют следующие сокращения с соответствующими обозначениями.

Сокращение	Расшифровка
BD	Big Data (Большие данные)
API	Application programming interface – Интерфейс программирования приложений
GSBPM	Generic Statistical Business Process Model (Типовая модель производства статистической информации)
IMSI	International Mobile Subscriber Identity — международный идентификатор мобильного абонента
IMEI	International Mobile Equipment Identity — международный идентификатор мобильного оборудования
СНГ	Содружество Независимых Государств
Статкомитет СНГ	Межгосударственный статистический комитет Содружества Независимых Государств
LAC	Location Area Code (Код зоны локации)
CID	Cell ID (Идентификатор соты)
FTP	File Transfer Protocol (Протокол передачи файлов)
SFTP	SSH File Transfer Protocol (Безопасный протокол передачи файлов)
АИД	Альтернативные источники данных
БД	База данных
ПО	Программное обеспечение
ТЗ	Техническое задание
ИАС	Информационная аналитическая система
ТМПСИ	Типовая модель производства статистической информации

В настоящем документе применяют следующие термины с соответствующими определениями.

Термин	Определение
Данные сотовых операторов (Mobile Network Data, MND)	Данные сотовых операторов (Mobile Network Data, MND). Обезличенные агрегированные данные, генерируемые оборудованием сетей мобильной связи в процессе их эксплуатации и предоставления услуг связи абонентам. Не являются персональными данными в контексте их использования для статистических оценок в соответствии с методологией, описанной в настоящем документе.
Агрегатор данных, Подрядчик	Юридическое лицо, выполняющее по договору со статистической службой и операторами связи функции сбора, обработки, агрегации и очистки данных от нескольких операторов связи, их обезличивания и преобразования в формат, пригодный для построения статистических оценок.
Заказчик	Статкомитет СНГ и/или национальные статистические службы
Поставщик	Оператор сотовой связи, выполняющий по договору с Агрегатором данных поставку предварительно агрегированных данных, собранных с оборудования сотовой сети
Предварительный агрегат	Промежуточный набор данных, сформированный оператором связи на уровне вышек/зон по заданной методологии, исключающий возможность идентификации отдельных абонентов и предназначенный для последующей обработки статистической службой.
Сетевое событие	Любое зафиксированное сетевое взаимодействие абонентского устройства с инфраструктурой оператора связи (например, выход в эфир, отправка SMS, начало сессии передачи данных, обновление локации, перемещение между сотами).

1. Общие положения

1.1 Наименование работ и основания для их проведения

Наименование работ: Оказание услуг по подготовке методических материалов, рекомендаций, стандартов, практических руководств, лучших практик по использованию больших данных и искусственного интеллекта в официальной статистике (этап 2025 года).

Основание: Работа выполняется в рамках проекта «Развитие статистики СНГ» в соответствии с Распоряжением Правительства РФ от 28.12.2023 №3996-рс, Календарным планом по реализации проекта «Развитие статистики СНГ» на 2025 г., одобренным координационной рабочей группой и Советом руководителей статистических служб государств-участников СНГ от 5 декабря 2024 г.

1.2 Перечень документов, на основании которых ведется разработка материалов

Основаниями для выполнения работ являются:

- Стандарт Единой системы технологической документации (ГОСТ 3.1001);
- ГОСТ Р 59853-2021 «Информационные технологии (ИТ). Комплекс стандартов на автоматизированные системы. Автоматизированные системы. Термины и определения»;
- ГОСТ Р 51275–2006 «Защита информации. Объект автоматизации. Факторы, воздействующие на информацию. Общие положения»;
- Организационно-распорядительные документы (далее – ОРД) Заказчика.
- GSBPM - Типовая модель производства статистической информации - Generic Statistical Business Process Model, была разработана ЕЭК ООН и Руководящей группой по статистическим метаданным Конференции европейских статистиков.
- Заключительный отчет по Контракту на оказание услуг по подготовке методических материалов и рекомендаций, обзора лучших практик по использованию больших данных и искусственного интеллекта в официальной статистике № CIS-SD/24-ICONS-2.8/1 от 15 августа 2024 г. (заключительный отчет, этап 2).

1.3 Организации, участвующие в разработке материалов

Заказчик: Межгосударственный статистический комитет Содружества Независимых Государств.

Краткое наименование: Заказчик, Статкомитет СНГ.

Юридический адрес: 107450, г. Москва, ул. Мясницкая, д.39, строение 1.

Исполнитель: Общество с ограниченной ответственностью «М 13».

Краткое наименование: Исполнитель, ООО «М13».

Юридический адрес: 117638, г. Москва, Кривокожская ул., д. 6А, стр. 2, пом. 366.

1.4 Сроки выполнения работ

Работы предусматривают определенные этапы/стадии и сроки их выполнения, представленные в Таблице 1.4.1.

Таблица 1.4.1 - Календарный план

№	Задача	Срок выполнения	Отчетные материалы
1	Разработка проекта Концепции формирования организационных и методологических подходов к использованию больших данных в статистических системах государств-участников СНГ	30.10.2025	Проект Концепции формирования организационных и методологических подходов к использованию больших данных в статистических системах государств-участников СНГ (п. 4.1 настоящего ТЗ).
2	Разработка требований к созданию типовой модели производства статистических оценок численности населения на основе данных сотовых операторов в рамках ТМПСИ	30.10.2025	Проект Технического задания на создание типовой модели производства статистических оценок численности населения на основе данных сотовых операторов в рамках ТМПСИ (п. 4.2 настоящего ТЗ).
3.	Разработка требований к созданию типовой модели	30.10.2025	Проект Технического задания на создание типовой

	использования спутниковых данных в процессе производства официальной статистической информации в области растениеводства на основе анализа возможностей новейших технологий спутникового мониторинга и традиционных методов сбора статистической информации		модели использования спутниковых данных в процессе производства официальной статистической информации в области растениеводства на основе анализа возможностей новейших технологий спутникового мониторинга и традиционных методов сбора статистической информации (п. 4.3 настоящего ТЗ).
4	Подготовка комплекта материалов для формирования исходной базы документов по использованию больших данных в официальной статистике на официальном Интернет-портале Статкомитета СНГ	10.11.2025	Комплект материалов по использованию больших данных в официальной статистике (п. 4.4 настоящего ТЗ).
5	Подготовка сводной документации	22.11.2025	Сводный отчет об оказании услуг.

1.5 Цели выполнения работ

Целью выполнения данной работы в соответствии с настоящим Техническим заданием является:

- разработка проекта Концепции формирования организационных и методологических подходов к использованию больших данных в статистических системах государств-участников СНГ. Концепция должна быть логическим продолжением, расширяющим основные положения Заключительного отчета по Контракту на оказание услуг по подготовке методических материалов и рекомендаций, обзора лучших практик по использованию больших данных и искусственного интеллекта в официальной статистике № CIS-SD/24-ICONS-2.8/1 от 15

- августа 2024 г. и включать в себя: подробный анализ современного международного опыта, перечень видов больших данных и анализ информационных источников, формулирование основных методологических подходов к формированию системы показателей официальной статистики с использованием больших данных, которые могут быть получены на их основе, методологические подходы, организационные механизмы, требования к информационно-технологической структуре, подготовку кадров – для организации использования больших данных в статистических системах стран СНГ;
- подготовка двух типовых ТЗ на разработку требований к типовой модели бизнес-процессов по использованию больших данных в официальной статистике (на основе данных сотовых операторов и на основе использования спутниковых данных): Оценка численности населения и Оценка сельскохозяйственных угодий с использованием больших данных в процессе производства официальной статистической информации в области растениеводства;
 - формирование на официальном Интернет-портале Статкомитета СНГ исходной базы документов по использованию больших данных в официальной статистике.

2 Описание целей разработки требований к типовой модели производства статистических оценок численности населения на основе данных сотовых операторов в рамках ТМПСИ

Целями настоящего документа являются:

1. Создание универсальной методологии: разработка согласованного набора требований, правил, стандартов и методик, которые могут быть адаптированы и внедрены национальными статистическими службами государств-участников СНГ для оценки численности населения с использованием больших данных операторов сотовой связи.
2. Обеспечение сопоставимости и качества данных: установление единых требований к исходным данным, алгоритмам обработки, процедурам агрегации и верификации для обеспечения высокого качества, надежности и сопоставимости получаемых статистических оценок как между различными регионами внутри одной страны, так и между государствами-участниками СНГ.
3. Формирование типового бизнес-процесса: определение четких требований к этапам работ, ролям и ответственности всех участников процесса (Статкомитет СНГ, национальные статистические службы, операторы сотовой связи, агрегаторы данных), а также к необходимым технологическим инфраструктурам и компетенциям.
4. Снижение рисков и обеспечение соответствия:
 - минимизация юридических, технических и методологических рисков за счет заблаговременного определения требований по обеспечению конфиденциальности и анонимности данных;
 - соблюдение соответствия национальным законодательствам в области связи и защиты персональных данных, а также международным принципам официальной статистики.
5. Оптимизация затрат и ресурсов: создание типовой модели позволяет избежать дублирования усилий и излишних затрат каждой страной в отдельности на разработку собственных методик, при этом сфокусировать ресурсы на непосредственной реализации и адаптации готовых проверенных решений.

3 Требования к определению потребностей Заказчика

3.1 Требования к определению потребностей в информации, к проведению консультаций и подтверждению потребностей

3.1.1 Понимание бизнес-процессов производства статистической информации о численности населения

Описание охватывает два основных метода производства статистических данных: перепись населения и текущий учет (оценка) численности населения.

3.1.1.1 Бизнес-процесс: Проведение переписи населения (ПН)

Назначение и цель: получение единовременных, детальных и точных данных о численности, составе и структуре населения по состоянию на определенный момент времени. Результаты переписи населения являются основой для текущих оценок в межпереписной период, таким образом, используются для формирования государственной политики (социальной, экономической, бюджетной). Эти данные служат базой для планирования развития регионов, строительства инфраструктуры, школ, больниц.

Входные данные: законодательство в отношении переписи населения, утвержденный график (как правило, раз в 10 лет).

Основные подпроцессы (этапы):

1. Подготовительный этап включает следующие обязательные шаги:
 - нормативно-правовое и методологическое обеспечение (разработка и принятие законодательных актов о переписи населения; утверждение государственным органом (например, Правительством) основных методологических принципов);
 - разработка и утверждение программно-методологических документов (бланков переписных листов, инструкций);
 - составление и актуализация списков адресов;
 - подготовка инфраструктуры (выбор информационной системы, организация логистики, разработка или приобретение специализированного программного обеспечения, создание центра обработки данных);
 - наем и обучение временного переписного персонала (проведение тренингов, разбор инструкций, пробные опросы);
 - информационно-разъяснительная работа с населением (запуск масштабной рекламной кампании в СМИ, интернете, на улицах;

объяснение целей переписи, важности участия, гарантий конфиденциальности; информирование о способах прохождения переписи).

2. Этап сбора данных длится несколько недель и имеет следующие особенности реализации:
 - Способ сбора: сплошное статистическое наблюдение.
 - Методы сбора:
 - обход жилых помещений переписчиками: переписчик с планшетом посещает каждое помещение на своем участке и заполняет переписные листы со слов респондентов;
 - самостоятельное заполнение населением электронных форм на государственном портале (если предусмотрено), причем данные автоматически поступают в базу;
 - сбор данных в стационарных переписных участках: граждане могут прийти и переписаться самостоятельно в специальных пунктах.
 - Ключевое действие: заполнение переписных листов на каждое лицо, постоянно или временно находящееся на изучаемой территории.
3. Этап обработки и контроля данных подразумевает следующие меры:
 - приемка и сканирование бумажных переписных листов (если применимо);
 - ввод и верификация данных;
 - автоматизированный и ручной логический и арифметический контроль полученной информации;
 - выборочные контрольные обходы для проверки качества и полноты учета.
4. Этап формирования результатов включает следующие шаги:
 - подсчет итогов: предварительные (численность), окончательные (полные данные по полу, возрасту, образованию и др.);
 - формирование баз данных, построение сложных таблиц, кросс-табуляция;
 - публикация итогов в виде статистических сборников, бюллетеней, распространение через официальный сайт статистической службы и данные открытого доступа.

Выходные данные: официальные публикации с итогами переписи, базы микроданных (анонимные), нормативно-справочная информация для текущих оценок.

3.1.1.2 Бизнес-процесс: Текущая оценка численности населения (межпереписной период)

Назначение и цель: ежегодная и ежеквартальная оценка численности и состава населения на основе данных последней переписи с учетом демографических событий (рождений, смертей, миграции).

Входные данные:

- Базовая численность населения (Отправная точка): итоговые данные последней проведенной переписи населения. Они являются «точкой отсчета» для всех последующих расчетов.
- Данные о естественном движении населения, а именно, ежемесячные данные из ЗАГС:
 - о рождениях (с указанием пола и даты рождения ребенка);
 - о смертях (с указанием пола, даты рождения и даты смерти умершего).
- Данные о миграции: ежемесячные данные от территориальных органов МВД (или миграционных служб):
 - о прибывших на постоянное место жительства (иммигранты);
 - о выбывших на постоянное место жительства (эмигранты);
 - также может учитываться внутренняя миграция (переезды внутри изучаемой территории).
- Нормативно-справочная информация: методологии расчета, коэффициенты, классификаторы (возрастные группы и т.д.).

Основные подпроцессы (этапы):

1. Сбор данных из административных источников:
 - получение данных от Минюста и МВД о регистрации рождений и смертей;
 - получение данных от МВД о регистрации и снятии с учета по месту жительства/пребывания (миграционные потоки);
 - проверка полноты, непротиворечивости и качества полученных данных.
2. Методологический расчет (балансовый метод):

– Формула:

Население на конец периода = Население на начало периода + Родившиеся за период - Умершие за период + Прибывшие иммигранты - Выбывшие эмигранты

- Расчет ведется в разрезе половозрастных групп, полученных по итогам последней переписи.
- Применение коэффициентов дожития (таблиц смертности) для «старения» населения.

3. Корректировка и верификация данных:

- внутренняя проверка, то есть сравнение полученных данных с предыдущими периодами для выявления аномальных скачков;
- сопоставление с данными из альтернативных источников (например, данные Пенсионного фонда о численности получателей пенсий);
- демографический анализ, а именно, проверка на соответствие демографическим закономерностям (например, соотношение полов при рождении, нормальные уровни смертности по возрастам);
- учет незарегистрированной миграции (экспертные оценки);
- корректировка результатов: внесение поправок в случае выявления существенных расхождений или ошибок в исходных данных.

4. Утверждение и публикация данных:

- подготовка оперативных и ежегодных данных о численности постоянного населения;
- согласование и официальное утверждение данных руководством статистической службы;
- публикация в демографических ежегодниках, статистических бюллетенях, на интерактивных порталах.

Выходные данные: официальные оценки численности населения на 1 января каждого года, ежеквартальные оперативные оценки, данные о компонентах изменения численности (естественный и миграционный прирост).

3.1.2 Ограничения и недостатки существующих процессов

Для переписи населения:

1. Высокая стоимость и ресурсоемкость. Процесс переписи требует привлечения огромного временного персонала и значительных финансовых затрат.
2. Низкая частота. Данные быстро устаревают, особенно в условиях динамичных миграционных процессов.
3. Риск недоучета определенных групп населения, таких как лица без определенного места жительства, временные трудовые мигранты, население удаленных территорий.
4. Риск сознательного сокрытия информации респондентами (например, из-за недоверия к организациям, занимающимся переписью).
5. Большой временной диапазон между моментом счета и публикацией окончательных итогов.

Для текущей оценки:

1. Зависимость от точности и полноты административных данных. Несовершенство учета миграции – ключевая проблема.
2. Накопление ошибки. Ошибки переписи и неточности учета событий накапливаются с каждым годом, к концу межпереписного периода точность оценки снижается.
3. Проблема учета временной и незаконной миграции. Данные МВД часто отражают лишь формальную регистрацию, а не реальное перемещение людей.
4. Методологические сложности учета долгосрочной миграции и определения «проживающего населения» в современных условиях высокой мобильности.
5. Неполное соответствие определений в разных ведомствах (например, разное определение «мигранта» в статистических организациях и МВД).

3.1.3 Вопросы, требующие решения (направления развития)

1. Интеграция данных и использование больших данных:
 - разработка методологии использования альтернативных источников данных (данные мобильных операторов, социальных сетей, иные большие данные) для верификации и дополнения официальных оценок, особенно по миграции;
 - создание единой межведомственной платформы обмена данными для минимизации расхождений в учете.
2. Совершенствование учета миграции:

- уточнение понятийного аппарата (ввод четких определений, таких как «долгосрочная/краткосрочная миграция»);
- разработка и внедрение надежных методов оценки незарегистрированной миграции (например, выборочные обследования, экспертные методы);
- налаживание более четкого учета эмиграции.

3. Модернизация переписи:

- постепенный переход на использование цифровых технологий с применением больших данных для сокращения издержек и сроков публикации;
- внедрение модели «регистрационной переписи» (на основе объединенных административных регистров) как основного или дополнительного метода;
- повышение доверия населения к переписи через прозрачность и разъяснение целей.

4. Методологическое развитие:

- разработка методов оперативного реагирования на кризисные ситуации (например, пандемия, массовые перемещения), влияющие на численность населения;
- совершенствование демографических прогнозов на основе новых методов анализа данных.

5. Повышение доступности и аналитичности данных:

- разработка интерактивных аналитических сервисов на основе публикуемых данных;
- предоставление данных в более детальном территориальном разрезе (муниципальном) и более частотном режиме.

3.1.4 Требования к описанию бизнес-процессов для реализации направлений развития

Требование 3.1.4.1. Заказчик (Статкомитет СНГ и национальные статистические службы) должен четко определить перечень статистических показателей, необходимых для формирования при помощи новых методов сбора статистических данных (например, среднесуточная численность населения, плотность населения, показатели маятниковой и трудовой миграции, оценка присутствующего населения в зонах туристической активности и т.д.). В текущей реализации таким показателем является численность населения на изучаемой территории.

Требование 3.1.4.2. Необходимо провести цикл консультаций с основными пользователями статистической информации (органы государственного управления, местного самоуправления, научное и бизнес-сообщество) для подтверждения актуальности и полноты запрашиваемых показателей.

Требование 3.1.4.3. Должна быть произведена оценка возможности дополнения традиционных методов сбора данных (переписи, выборочные обследования) предлагаемыми оценками на основе больших данных с точки зрения периодичности, детализации, оперативности и стоимости получения данных.

Требование 3.1.4.4. Должно быть предоставлено детальное описание шагов по каждому этапу сбора, обработки и анализа статистических данных, полученных от операторов сотовой связи как источника больших данных.

Требование 3.1.4.5. Должны быть представлены схемы взаимодействия участников процесса.

Требование 3.1.4.6. Требуется предоставить инструкции по работе со статистическими данными операторов сотовой связи, включая их проверку и оценку качества, а также методы агрегации с целью получения общей статистики по населению.

3.2 Требования к установлению целей формирования материалов

Настоящий раздел определяет возможные цели интеграции альтернативных источников данных (АИД) в процесс производства официальной статистики о численности и составе населения и требования к данным целям. Использование АИД не подразумевает замену традиционных методов (перепись, текущий учет), а направлено на их дополнение, верификацию и повышение оперативности, точности и детализации итоговых данных. Цели сгруппированы по приоритетности и функциональному назначению.

3.2.1 Варианты целей использования данных сотовых операторов в процессе производства официальной статистики о численности и составе населения

3.2.1.1 Цель №1: Верификация и повышение точности данных текущего учета и переписи

Требование 3.2.1.1.1. Снижение ошибки недо-/переучета населения.

Задача: выявлять и количественно оценивать расхождения между данными официального учета и сигналами из АИД (например, выявление

районов с аномально высокой/низкой численностью населения по сравнению с официальными значениями численности).

Критерий успеха: повышение репрезентативности данных, особенно для трудно учитываемых групп населения (временные мигранты, жители удаленных территорий, лица без определенного места жительства).

Требование 3.2.1.1.2. Контроль качества административных данных.

Задача: использовать АИД для проверки полноты и своевременности поступления данных из административных источников (например, данных МВД о регистрации).

Критерий успеха: своевременное выявление и корректировка пробелов в данных или неполного охвата в данных регистрации.

3.2.1.2 Цель №2: Оперативная оценка численности и перемещений населения в межпереписной период

Требование 3.2.1.2.1. Получение высокочастотных оценок.

Задача: обеспечивать оценку численности и плотности населения с высокой периодичностью (ежемесячно, ежеквартально), что невозможно при использовании только традиционных методов.

Критерий успеха: возможность публикации оперативных оценок, что повысит эффективность принятия управленческих решений на региональном и муниципальном уровне.

Требование 3.2.1.2.2. Мониторинг краткосрочных и сезонных миграционных потоков.

Задача: учесть при расчете численности населения интенсивность и сезонность внутренних и внешних миграционных потоков на основе данных о реальной мобильности (например, данные мобильных операторов).

Критерий успеха: формирование актуальной картины численности населения с учетом миграционной активности, не отражаемой в полной мере данными о регистрации.

3.2.1.3 Цель №3: Повышение детализации и аналитической ценности данных

Требование 3.2.1.3.1. Получение данных в разрезе малых территорий.

Задача: обеспечивать оценку численности населения не только на уровне региона или города, но и на уровне отдельных районов, микрорайонов и даже кварталов.

Критерий успеха: возможность предоставления данных для задач городского планирования, размещения объектов инфраструктуры и социальной сферы.

Требование 3.2.1.3.2 Анализ пространственного поведения и активности населения в качестве потенциального направления развития в сфере использования данных сотовых операторов в официальной статистике.

Задача (потенциальная): изучать не только место регистрации, но и реальные маршруты перемещений, места работы, учебы и досуга (в агрегированном и обезличенном виде).

Критерий успеха: формирование комплексных карт человеческой активности для социо-экономического анализа.

3.2.1.4 Цель №4: Подготовка и оптимизация проведения переписей населения

Требование 3.2.1.4.1. Актуализация картографического материала и списков адресов в качестве потенциального направления развития в сфере использования больших данных в официальной статистике.

Задача (потенциальная): использовать пространственные данные и данные о реальной застройке для уточнения переписных участков и маршрутов переписчиков.

Критерий успеха: снижение количества пропущенных домохозяйств и дублирования записей.

Требование 3.2.1.4.2. Валидация итогов переписи.

Задача: проводить оперативное сравнение предварительных итогов переписи с агрегированными данными АИД для выявления возможных аномалий и оперативной корректировки процесса подведения итогов.

Критерий успеха: повышение доверия к итоговым результатам переписи.

3.2.1.5 Цель №5: Исследование и апробация новых методов (Research & Development)

Требование 3.2.1.5.1. Разработка и калибровка статистических и математических моделей.

Задача: использовать АИД для создания и обучения моделей, которые в перспективе могут стать основой для оценок (например, модель оценки численности на основе ночной популяции и данных о мобильности).

Критерий успеха: создание прототипов методик, прошедших валидацию и готовых к внедрению в производственный процесс.

3.2.2 Ограничивающие принципы (недопустимые цели использования данных сотовых операторов в процессе производства официальной статистики о численности и составе населения)

Использование АИД должно исключать следующие цели:

- Идентификация личности. Целью не может быть получение или использование персональных данных конкретных физических лиц. Все данные должны использоваться исключительно в агрегированном и обезличенном виде.
- Надзор и контроль. Данные не могут использоваться для слежки, контроля перемещений или деятельности отдельных лиц или групп.
- Полная замена традиционных методов. АИД являются дополнением, а не заменой переписи и текущего учета, которые остаются легальной и методологической основой.
- Коммерциализация. Полученные данные не могут быть использованы в коммерческих целях или переданы третьим лицам для извлечения прибыли.

3.2.3 Критерии выбора и оценки альтернативных источников данных

Для достижения поставленных целей АИД должны оцениваться по следующим критериям:

- Репрезентативность: охват ключевых групп населения.
- Качество и точность: достоверность и уровень шума в данных.
- Периодичность и оперативность: частота обновления и скорость предоставления.
- Сопоставимость: возможность привязки к официальной сетке административно-территориального деления.

- Соответствие правовым нормам: соответствие законодательству о персональных данных и статистической деятельности.
- Стоимость и устойчивость: финансовая и операционная возможность долгосрочного использования источника.

Этот набор требований обеспечивает целенаправленное, этичное и методологически обоснованное включение новых источников данных в производство официальной статистики.

3.2.4 Требования к постановке целей использования АИД

Требование 3.2.4.1. Цели формирования материалов должны быть конкретными, измеримыми, достижимыми, релевантными и ограниченными по времени.

Требование 3.2.4.2. Необходимо учесть достижение следующих стратегических целей:

- Повышение оперативности получения статистической информации о населении.
- Увеличение пространственной и временной детализации статистических данных.
- Снижение нагрузки на респондентов и затрат на проведение традиционных выборочных обследований.
- Внедрение инновационных методов в производственный процесс официальной статистики.

Требование 3.2.4.3. Для каждой цели должен быть определен набор ключевых показателей эффективности (KPI), например:

- Срок от заказа операторам предварительных агрегатов до публикации итоговых показателей.
- Разрешающая способность (минимальная территориальная единица и временной интервал для расчета).

3.3 Требования к определению концепций

Данные сотовых операторов являются перспективным, но чувствительным источником информации для оценки численности и перемещения населения. Необходимо обеспечить требования к их использованию статистическими службами с целью обеспечения легитимности, этичности, безопасности и методологической корректности. Использование данных сотовых операторов рассматривается как дополнительный инструмент для верификации, дополнения и оперативного

анализа данных, полученных традиционными методами (перепись, текущий учет).

3.3.1 Правовые и этические требования

3.3.1.1 Законность и нормативное закрепление

Требование 3.3.1.1.1. Использование данных сотовых операторов должно быть разрешено законом. Закон должен четко определять:

- статус статистической организации как уполномоченного органа на запрос агрегированных и обезличенных данных сотовых операторов;
- цели использования (оценка численности, мониторинг миграции, верификация иных данных);
- принцип обязательности предоставления операторами данных в обезличенном и агрегированном виде по установленным статистической организацией форматам.

Требование 3.3.1.1.2. Соблюдение законодательства о персональных данных является абсолютным приоритетом. Обработка данных возможна только после процедуры окончательного и необратимого обезличивания и агрегации, проводимой оператором на стороне оператора. Передача персональных данных недопустима.

Требование 3.3.1.1.3. Обязательное заключение регламентированных соглашений между агрегатором данных и каждым сотовым оператором. Соглашение должно детально определять:

- объем, состав, формат и периодичность предоставляемых данных;
- юридически обязательные технические требования к обезличиванию и агрегации;
- порядок, сроки и защищенные каналы передачи;
- ответственных лиц с обеих сторон;
- ответственность за нарушение условий.

3.3.1.2 Этические принципы и прозрачность

Требование 3.3.1.2.1. Принцип «Privacy by Design» (конфиденциальность через проектирование): все процессы должны быть спроектированы так, чтобы конфиденциальность данных обеспечивалась по умолчанию.

Требование 3.3.1.2.2. Публичная декларация методологии: статистическая организация обязана опубликовать подробное методологическое пояснение о том, какие именно данные используются, как

они агрегируются, какие алгоритмы применяются для получения итоговых оценок и как обеспечивается анонимность.

3.3.2 Технические требования к данным и их обработке

3.3.2.1 Требования к составу и формату данных

Требование 3.3.2.1.1. Данные должны предоставляться операторами исключительно в агрегированном виде. Пример приемлемого формата агрегации:

- Агрегация в пространстве: данные агрегируются по выбранному разбиению исследуемой территории (например, административно-территориальное деление).
- Агрегация во времени: данные предоставляются за определенные интервалы (например, ежемесячно).
- Содержание агрегата: для каждой пространственно-временной ячейки передается только количество уникальных абонентов. Передача IMSI, IMEI, номеров телефонов, детализированной геолокации отдельных абонентов запрещена.

Требование 3.3.2.1.2. Для анализа численности необходимо выделять постоянно проживающее на изучаемой территории население, то есть тех людей, которые преимущественно пребывали в исследуемой территориальной единице в ночное время (с 23:00 до 06:00 местного времени) на протяжении последнего месяца.

3.3.2.2 Требования к инфраструктуре и безопасности

Требование 3.3.2.2.1. Создание статистической организацией защищенной платформы для приема и обработки данных от агрегатора. Платформа должна иметь:

- системы мониторинга;
- контроль физического и логического доступа с обязательной аутентификацией;
- процедуры регулярного аудита безопасности.

Требование 3.3.2.2.2. На стороне агрегатора также должна быть обеспечена безопасность данных:

- загрузка файлов от операторов должна производиться через FTP-сервер;
- аналитические операции должны проводиться внутри защищенной среды.

3.3.3 Методологические требования

3.3.3.1 Методологическая база

Требование 3.3.3.1.1. Концепция использования данных сотовых операторов должна быть основана на принципах Типовой модели производства статистической информации (GSBPM) и охватывать все ее этапы: от определения потребностей до распространения данных.

3.3.3.2 Калибровка и верификация

Требование 3.3.3.2.1. Обязательная калибровка алгоритмов агрегации на данных переписи населения. Данные, полученные на базе агрегатов сотовых операторов (например, «количество проживающих людей на территории») должны быть соотнесены с данными последней переписи для расчета корректирующих коэффициентов.

Требование 3.3.3.2.2. Валидация результатов. Получаемые оценки должны постоянно сверяться с данными из других источников: административные данные (Бюро технической инвентаризации, реестры избирателей и т.п.), данные иных операторов связи для перекрестной проверки.

3.3.3.3 Учет ограничений и погрешностей

Требование 3.3.3.3.1. Методология должна явно учитывать и количественно оценивать системные погрешности:

- неполный охват населения (дети, пожилые люди без телефонов);
- завышение количества людей из-за наличия у одного человека нескольких SIM-карт;
- население, временно находящееся на изучаемой территории, но не проживающее там постоянно.

3.3.4 Ресурсные и организационные требования

Требование 3.3.4.1. Подготовка квалифицированных кадров статистической организации: аналитиков данных, статистиков, демографов и специалистов по ИБ, способных работать с агрегированными данными сотовых операторов.

Требование 3.3.4.2. Выделение целевого финансирования на создание защищенной ИТ-инфраструктуры, закупку вычислительных мощностей и проведение работ по оценке численности населения с использованием данных сотовых операторов.

Требование 3.3.4.3. Назначение ответственных лиц в статистической организации, на стороне агрегатора и у каждого оператора связи за организацию процесса и соблюдение регламентов.

3.4 Требования к проверке наличия данных

3.4.1 Требования к предлагаемым перспективным решениям

3.4.1.1 Требования к перспективным решениям

Требования к перспективным решениям как часть проверки наличия и соответствия данных для производства статистических оценок численности населения на основе данных сотовых операторов в рамках ТМПСИ расширяют и повторяют уже описанные выше требования к определению концепций (п. 3.3), поэтому ниже требования перечислены кратко без детализации по каждому пункту.

Требование 3.4.1.1.1. Принцип анонимности и агрегирования

- Все данные от операторов должны предоставляться в агрегированном виде без персональной информации (без IMSI, IMEI, номеров телефонов).
- Данные должны агрегироваться по временным интервалам (например, 1 месяц) и географическим зонам (например, по административно-территориальному делению или по муниципальным образованиям).
- Для анализа должны использоваться только количество активных абонентов в зоне, а не данные о перемещениях конкретных людей.

Требование 3.4.1.1.2. Масштабируемость и стандартизация

- Решение должно быть масштабируемым для работы с несколькими операторами связи.
- Необходима разработка единых стандартов обмена данными (форматы данных, протоколы передачи, частота обновления).

Требование 3.4.1.1.3. Сверка с другими источниками данных

Решение должно позволять сверять данные сотовых операторов с другими источниками (например, данные переписи, административные данные, информация от органов власти) для повышения точности и верификации.

Требование 3.4.1.1.4. Автоматизация и оперативность

- Процесс сбора, обработки и анализа данных должен быть максимально автоматизирован.
- Решение должно обеспечивать получение данных с минимальной задержкой (например, ежемесячно).

Требование 3.4.1.1.5. Безопасность данных

- Все данные должны передаваться по защищенным каналам связи.
- Доступ к данным должен быть строго регламентирован.

3.4.1.2 Процесс сбора, верификации, агрегации данных и формирования показателей

Детально процессы описаны в Разделах 4 и 5. В данном пункте кратко перечислены требования, которые подробно расшифрованы в соответствующих параграфах.

3.4.1.2.1 Процесс сбора данных

Требование 3.4.1.2.1.1. Должно быть обеспечено заключение соглашений с операторами связи

- Определение правовых основ передачи данных.
- Разработка технических регламентов передачи данных.

Требование 3.4.1.2.1.2. Должна быть настроена автоматизированная передача данных

- Операторы настраивают автоматическую выгрузку агрегированных данных в соответствии с установленными стандартами.
- Данные передаются на защищенные серверы агрегатора на регулярной основе (например, ежемесячно).

Требование 3.4.1.2.1.3. Должен быть определен состав передаваемых данных

Для каждой географической зоны (привязка к административно-территориальному делению) и временного интервала (например, 1 месяц) передается:

- Численность населения.
- Временная метка.
- Идентификатор географической зоны.

3.4.1.2.2 Процесс верификации данных

Требование 3.4.1.2.2.1. Должен быть обеспечен контроль качества данных

- Проверка полноты данных.
- Проверка формата данных и соответствия стандартам.
- Выявление аномалий (например, резкие скачки количества абонентов).

Требование 3.4.1.2.2.2. Должна происходить кросс-валидация с другими источниками

- Сравнение данных от разных операторов между собой.

- Сравнение с данными переписи населения, административными данными (например, данные о численности населения из органов власти).

Требование 3.4.1.2.2.3. Необходимо предусмотреть калибровку алгоритмов расчета

- Настройка коэффициентов для учета доли охвата населения оператором связи.
- Учет доли населения, не охваченного сотовой связью (дети, пожилые люди).

3.4.1.2.3 Процесс агрегации данных

Требование 3.4.1.2.3.1. Пространственная агрегация должна быть построена следующим образом: данные от всех операторов агрегируются по единым географическим зонам (например, по административно-территориальному делению).

Требование 3.4.1.2.3.2. Временная агрегация должна быть построена следующим образом: данные агрегируются по необходимым временным интервалам (например, ежедневно, еженедельно, ежемесячно).

Требование 3.4.1.2.3.3. Агрегация по операторам должна происходить с учетом их доли на рынке и исключать двойной учет абонентов с несколькими SIM-картами.

3.4.1.2.4 Формирование бизнес-показателей для конечного заказчика

Требование 3.4.1.2.4.1. Должны быть определены потребности заказчика

- Заказчиками могут быть: органы государственной власти, местного самоуправления, исследовательские организации.
- Примеры потенциальных потребностей, которые в перспективе могут быть обеспечены при помощи данных операторов сотовой связи:
 - Оценка численности постоянного населения.
 - Анализ маятниковой миграции.
 - Мониторинг туристических потоков.
 - Оценка нагрузки на инфраструктуру.

Требование 3.4.1.2.4.2. Необходимо обеспечить расчет показателей, согласованных с заказчиком, исходя их потребностей.

- Численность постоянно проживающего населения на основе данных о ночном населении с учетом коэффициентов доли охвата населения оператором сотовой связи.

- Численность временно находящегося на территории населения.

Требование 3.4.1.2.4.3. Интерпретация результатов должна быть реализована следующим образом:

- Проведение анализа полученных данных;
- Подготовка аналитических отчетов и рекомендаций для заказчика.

Таким образом, предложенное решение позволяет создавать современную систему мониторинга численности и перемещения населения с высокой точностью и оперативностью, обеспечивая при этом полную конфиденциальность и безопасность данных.

3.4.2 Требования к соответствию альтернативного источника данных потребностям статистической организации

Использование данных сотовых операторов в качестве альтернативного источника информации для официальной статистики населения должно строго соответствовать целям, методологическим стандартам и правовым нормам статистической организации. Настоящий раздел определяет критерии, при которых данные сотовых операторов могут считаться релевантным, надежным и допустимым источником для формирования официальной статистики.

3.4.2.1 Требования к методологическому соответствию

3.4.2.1.1 Репрезентативность

Требование 3.4.2.1.1.1 Исходные данные должны обеспечивать достаточный охват генеральной совокупности. Агрегатор должен оценить:

- Охват населения сотовой связью: доля населения, не охваченного сотовой связью (дети, пожилые, отдельные социальные группы), должна быть статистически незначимой или поддающейся корректировке с помощью модельных коэффициентов.
- Географическое покрытие: данные должны охватывать всю территорию страны, включая сельскую местность и удаленные районы, с приемлемым уровнем сигнала.
- Доля рынка операторов: совокупность операторов, предоставляющих данные, должна покрывать не менее 50% рынка для минимизации систематических смещений.

Требование 3.4.2.1.1.2. Исходные данные должны позволять определять локацию абонента с точностью не менее чем на уровень административного района или крупного города.

Требование 3.4.2.1.1.3. Исходные данные должны иметь временную метку, позволяющую привязывать активность абонента к конкретному времени суток и определять продолжительность его нахождения в данной локации.

3.4.2.1.2 Точность и достоверность

Требование 3.4.2.1.2.1. Данные должны допускать верификацию и калибровку по отношению к эталонным источникам:

- Калибровка по данным переписи: оценки численности, полученные из данных сотовых операторов, должны быть скорректированы на основе точных данных последней переписи населения на уровне административных районов.
- Кросс-валидация с административными данными (Бюро технической инвентаризации, реестры избирателей и т.п.).
- Учет «шумов»: Методология должна явно учитывать и нивелировать такие факторы, как «двусимочность», население, временно находящееся на исследуемой территории.

3.4.2.1.3 Сопоставимость и согласованность

Требование 3.4.2.1.3.1. Данные должны быть приведены к единым статистическим понятиям:

- Определение «постоянного населения»: резидентное население должно определяться через критерий ночного присутствия (например, наиболее частая локация в ночные часы) в соответствии с методологией переписи.
- Привязка к административно-территориальному делению: данные отчетов должны содержать информацию о населении в границах муниципальных образований, территориальных субъектов и иных официальных территориальных единиц учета.

3.4.2.2 Требования к операционному соответствию

3.4.2.2.1 Периодичность и оперативность

Требование 3.4.2.2.1.1. Данные должны предоставляться с частотой, превышающей традиционные методы, например, ежемесячное агрегирование для формирования официальных оценок и отчетности.

Требование 3.4.2.2.1.2. Должны быть сформулированы требования о скорости и качестве предоставления данных.

3.4.2.2.2 Доступность и формат данных

Требование 3.4.2.2.2.1. Данные должны предоставляться в стандартизированном машиночитаемом формате (например, CSV), согласованном со статистической организацией.

3.4.2.2.3 Масштабируемость и автоматизация

Требование 3.4.2.2.3.1. Процесс сбора итоговых данных должен быть автоматизирован и интегрирован в ИТ-инфраструктуру статистической организации для минимизации ручного труда.

3.4.2.3 Требования к правовому и этическому соответствию

Требования к правовому и этическому соответствию данных аналогичны требованиям к концепциям, описанным в п. 3.3.1.

3.4.2.3.1 Законность

Требование 3.4.2.3.1.1. Использование данных сотовых операторов должно быть разрешено законодательством с четким определением:

- Статуса статистической организации как получателя данных.
- Цели использования – исключительно для формирования официальной статистики.
- Требований к обезличиванию и агрегации данных на стороне оператора.

Требование 3.4.2.3.1.2. Процесс должен полностью соответствовать законодательству о персональных данных. Передача персональных данных недопустима. Все данные должны быть агрегированы и обезличены до уровня, исключающего возможность идентификации личности.

3.4.2.3.2 Этичность и прозрачность

Требование 3.4.2.3.2.1. Общественная прозрачность: статистическая организация обязана публиковать подробное методологическое описание использования данных сотовых операторов, включая принципы агрегации, калибровки и меры по защите конфиденциальности.

Вывод: данные сотовых операторов могут считаться соответствующими потребностям статистической организации только при комплексном выполнении всех вышеперечисленных требований. Это превращает их из сырого источника больших данных в надежный, верифицируемый и этичный статистический инструмент, дополняющий традиционную систему учета.

3.4.3 Требования к разделению обязанностей между поставщиком данных, агрегатором данных и статистической организацией

Эффективное и безопасное использование данных сотовых операторов для официальной статистики требует четкого нормативного и технологического разграничения зон ответственности между участниками процесса. Настоящий раздел определяет требования к разделению обязанностей между сотовым оператором (Поставщик данных), Агрегатором данных (Подрядчик) и Статистической организацией (Заказчик).

Подробно обязанности и роль каждого из участников описаны в разделах 4 и 5 настоящего документа.

3.4.3.1 Общие принципы

1. Принцип минимальной достаточности: каждый участник получает доступ только к тем данным и функциям, которые абсолютно необходимы для выполнения его задач.
2. Принцип сквозного аудита: все действия участников должны протоколироваться и быть доступными для независимого аудита.
3. Принцип прозрачности методологии: алгоритмы агрегации и обработки должны быть документально зафиксированы и доступны для контроля со стороны статистической организации.

3.4.3.2 Обязанности и требования к сотовому оператору (Поставщик данных)

Роль: Источник первичных сырых данных о событиях в сети связи.

Требование 3.4.3.2.1. Оператор сотовой связи обязан обеспечить соблюдение законодательства в отношении использования сотовых данных с целью формирования статистики:

- Обеспечение правового основания для обработки и передачи агрегированных данных в соответствии с законодательством о связи и персональных данных.
- Заключение договора с Агрегатором, определяющего цели, объемы и условия передачи данных.

Требование 3.4.3.2.2. Оператор сотовой связи обязан провести техническую подготовку данных:

- Первичная агрегация и обезличивание: преобразование сырых данных сетевых событий в агрегированные временно-пространственные срезы.

- Формат данных: передача данных в строго согласованном формате (например, численность населения в административном районе на определенную дату).

Требование 3.4.3.2.3. Оператор сотовой связи обязан обеспечить безопасность данных:

- Передача данных по защищенным каналам связи.

Требование 3.4.3.2.4. Оператор сотовой связи обязан обеспечить качество и доступность данных:

- Обеспечение полноты, непрерывности и своевременности передачи данных в соответствии с регламентом.

3.4.3.3 Обязанности и требования к Агрегатору данных

Роль: Технический посредник, обеспечивающий консолидацию, дополнительную обработку и очистку данных множества операторов, их верификацию и итоговую агрегацию.

Требование 3.4.3.3.1. Агрегатор обязан провести консолидацию и стандартизацию данных:

- Прием данных от всех операторов по единому протоколу.

Требование 3.4.3.3.2. Агрегатор обязан обеспечить верификацию и калибровку данных:

- Сопоставление данных от разных операторов.
- Калибровка данных на основе эталонных источников (перепись населения, административные данные).
- Расчет и применение корректирующих коэффициентов (доля населения, пользующегося услугами связи оператора, «двусимочность»).

Требование 3.4.3.3.3. Агрегатор обязан обеспечить безопасность и контроль доступа:

- Функционирование на защищенной платформе.

3.4.3.4 Обязанности и требования к Статистической организации

Роль: Конечный заказчик и пользователь данных. Ответственность за методологию и публикацию официальной статистики.

Требование 3.4.3.4.1. Статистическая организация должна обеспечить конфиденциальность и безопасность:

- Создание и содержание защищенной платформы для приема и анализа данных.
- Соблюдение режима конфиденциальности на всех этапах работы.

- Проведение регулярного аудита процессов для обеспечения соответствия законодательству.

Требование 3.4.3.4.2. Статистическая организация должна обеспечить публичность результатов процесса:

- Публикация методик и объяснение того, как используются данные сотовых операторов для формирования статистики.
- Обеспечение прозрачности и подотчетности процесса.
- Подготовка публикаций и отчетных материалов.

Такое разделение обязанностей позволяет создать безопасную, масштабируемую и методологически корректную систему использования больших данных для общественного блага, минимизируя риски для приватности граждан.

3.4.4 Требования по выявлению неполноты охвата поставщиком больших данных территориальных единиц страны или отдельных групп населения

Использование данных сотовых операторов для статистических оценок населения критически зависит от полноты охвата сети. Неполнота охвата может приводить к систематическим ошибкам в оценках численности населения, особенно в удалённых и сельских районах, а также среди отдельных демографических групп. Настоящий раздел определяет требования к методологии выявления и оценки неполноты охвата.

3.4.4.1 Требования к мониторингу территориального охвата

Требование 3.4.4.1.1. Должна проводиться оценка уровня охвата населения сотовой связью:

- Сопоставление данных переписи населения и данных сотовых операторов на уровне малых территориальных единиц для выявления расхождений, указывающих на неполноту охвата.

3.4.4.2 Требования к выявлению неполноты охвата среди групп населения

Требование 3.4.4.2.1 Необходимо выявление групп с низким уровнем охвата сотовой связью:

- Проведение анализа демографической структуры абонентской базы операторов (в агрегированном виде) для выявления групп с низкой представленностью:
 - Дети младшего возраста.
 - Пожилые люди старше 75 лет.

- Сопоставление с внешними данными (например, с результатами переписи населения) для оценки представленности различных возрастных и социальных групп в данных операторов.

3.4.4.3 Требования к методологии оценки влияния неполноты охвата на статистические оценки

Требование 3.4.4.3.1. Должны быть разработаны корректирующие коэффициенты:

- Для территорий с неполным охватом должны быть разработаны корректирующие коэффициенты, учитывающие:
 - Уровень охвата сотовой связью
 - Долю населения, не охваченного связью.
- Коэффициенты должны регулярно пересматриваться с учётом изменения ситуации с покрытием и проникновением связи.

Требование 3.4.4.3.2. Должны применяться альтернативные методы оценки.

Для территорий с полным отсутствием покрытия должны применяться альтернативные методы оценки численности населения:

- Использование данных административных источников.

3.5 Требования к подготовке и представлению типовой бизнес-модели

Реализация проекта использования данных сотовых операторов для официальной статистики требует тщательной оценки и обоснования затрат. Настоящий раздел определяет требования к структуре, методологии и обоснованию затрат на внедрение и эксплуатацию предлагаемой бизнес-модели.

3.5.1 Требования к принципам построения бизнес-модели

Требование 3.5.1.1. Бизнес-модель должна описывать потоки создания стоимости и источники финансирования проекта (бюджетное финансирование, государственно-частное партнерство).

Требование 3.5.1.2. Модель должна быть устойчивой и предусматривать долгосрочное взаимодействие между всеми участниками процесса. Для реализации данного требования между сторонами процесса должны быть заключены соответствующие договоры, детально описывающие обязанности каждой из сторон и условия сотрудничества.

Требование 3.5.1.3. Модель должна включать в себя правовые аспекты, в том числе типовые формы соглашений о передаче данных между всеми

сторонами. Таким образом с правовой точки зрения будет обеспечена безопасность передаваемых партнерам данных.

3.5.2 Требования к структуре затрат

Затраты должны быть структурированы по следующим категориям: капитальные и операционные затраты.

Требование 3.5.2.1. Капитальные затраты (CAPEX) должны учитывать следующие:

- Разработка и внедрение программно-аппаратного комплекса:
 - затраты на разработку или приобретение программного обеспечения для сбора, обработки и анализа данных;
 - затраты на закупку серверного и сетевого оборудования;
 - затраты на создание защищенной инфраструктуры.
- Первоначальные подготовительные и методические работы:
 - затраты на разработку и тестирование методик обработки предварительно агрегированных данных операторов связи;
 - затраты на пилотные проекты с операторами связи.

Требование 3.5.2.2. Операционные затраты (ОРЕХ) должны учитывать следующие:

- Эксплуатационные расходы:
 - затраты на аренду каналов связи и облачных услуг;
 - затраты на техническое обслуживание и обновление оборудования и ПО;
 - затраты на электроэнергию.
- Затраты на закупку данных:
 - расходы на оплату услуг операторов связи за предоставление агрегированных данных;
 - расходы на оплату услуг Агрегатора за сбор, проверку, верификацию, агрегацию данных.
- Затраты на персонал:
 - фонд оплаты труда сотрудников (аналитики данных, статистики, IT-специалисты, юристы);
 - затраты на обучение и повышение квалификации персонала.
- Административные и прочие расходы:
 - затраты на юридическое и аудиторское сопровождение;
 - затраты на получение необходимых лицензий и сертификатов.

3.5.3 Требования к расчету бюджета

Требование 3.5.3.1. При детализации затрат на реализацию бизнес-модели применения данных сотовых операторов в производстве статистических оценок численности населения в расчете бюджета проекта требуется учесть работы и услуги, показанные в таблице 3.5.1.3.1 ниже.

В таблице представлен наиболее полный перечень работ и услуг, подразумевающий создание новой системы, в которую будут загружены статистические данные на основе данных операторов сотовой связи, а также визуализацию этих данных на интернет-портале. Однако не все проекты включают полный спектр работ и услуг. Так, например, настоящий документ отражает минимальную необходимую реализацию применения больших данных сотовых операторов в производстве статистических оценок численности населения, и такой вариант включает формирование статистической отчетности без необходимости разработки новых информационно-аналитических систем.

Таблица 3.5.3.1 – Перечень работ и услуг

№	Наименование
I	Работы:
1	Разработка ТЗ, Технорабочего проекта, включая сбор и подготовку требований
2	Разработка архитектуры системы ¹
3	Разработка (доработка) порталов, сайтов, клиентской части ИТ системы ²
4	Разработка баз данных, API, интеграционных решений

¹ В случае, если проект включает не только формирование статистических отчетов, но и создание информационной аналитической системы (ИАС). Минимально необходимая реализация, описанная в настоящем документе, не подразумевает создания отдельной ИАС.

² В случае, если проект включает не только формирование статистических отчетов, но и создание информационной аналитической системы (ИАС). Минимально необходимая реализация, описанная в настоящем документе, не подразумевает создания отдельной ИАС.

№	Наименование
5	Разработка интерфейсов в готовых системах, картографических системах ³
6	Разработка отчетов, аналитических бюллетеней, другой отчетной документации
II	Лицензии:
7	Приобретение прав на использование программ для ЭВМ, баз данных (далее по тексту совместно – ПО) по лицензионным договорам с правообладателем
III	Услуги:
8	Облачные сервисы (SaaS-сервисы/ПО, как услуга)
9	Техническая поддержка/сопровождение ПО
10	Действия по приобретению массивов данных
11	Сбор, обработка, верификация, агрегация и аналитика данных
12	Управление проектом (руководство проектом, управление командами)
13	Консультирование, сопровождение проекта

3.5.3.1 Разработка ТЗ, Технорабочего проекта, включая сбор и подготовку требований

Данные работы должны содержать:

- работы по разработке технического задания (далее – ТЗ) – основного исходного документа, определяющего порядок и условия выполнения работ/оказания услуг, направленных на реализацию бизнес-модели применения данных сотовых операторов в производстве статистических оценок численности населения (далее – Работ/Услуг соответственно), содержащего цель, задачи, принципы оказания, количественные и/или

³ Минимально необходимая реализация, описанная в настоящем документе, не подразумевает создания отдельной ИАС.

качественные и иные значимые характеристики Работ/Услуг, ожидаемые результаты и сроки оказания Работ/Услуг.

Кроме того, ТЗ должно содержать в том числе, но не ограничиваясь основные требования, предъявляемые к Работам/Услугам, порядок оказания и приемки Работ/Услуг, стадии выполнения Работ/оказания Услуг, состав отчетной документации, а также особые требования, обусловленные спецификой Работ/Услуг.

Разработка ТЗ включает выполнение следующих работ:

- сбор требований к Работам/Услугам,
 - проведение предварительных исследований,
 - анализ результатов предварительных исследований, расчётов.
 - анализ результатов детализированной информации, расчётов планируемых количественных показателей выполнения Работ/оказания Услуг.
- работы по разработке технорабочего проекта, являющегося одним из основных документов, которым руководствуются при создании (разработке) и внедрении ИАС в действие⁴.

На стадии технорабочего проектирования на основе утвержденного Технического задания разрабатываются основные положения проектируемой системы, принципы её функционирования и взаимодействия с другими системами; определяется структура Системы; разрабатываются проектные решения по обеспечивающим частям Системы.

Технорабочее проектирование охватывает большой круг задач:

- построение функциональной модели и модели данных,
- разработка протоколов взаимодействия с другими информационными системами и ПО,
- проектирование пользовательских интерфейсов,
- разработка технологических инструкций пользователям,
- разработка технологических инструкций администраторам системы,
- создание программы и методики приемочных испытаний.

Команда проекта, необходимая для выполнения данных работ:

- Аналитик ведущий

⁴ Технорабочий проект создается в случае проектирования ИАС. Минимально необходимая реализация, описанная в настоящем документе, не подразумевает создания отдельной ИАС

- Аналитик старший
- Аналитик
- Дополнительно сотрудники в случае разработки ИАС:
 - Специалист по тестированию ведущий
 - Специалист по тестированию младший
 - Frontend-разработчик
 - Backend-разработчик

3.5.3.2 Разработка архитектуры системы ⁵

Данная работа отражает фундаментальную организацию системы, реализованную в ее компонентах, связях этих компонентов друг с другом и внешней средой и принципах, определяющих структуру и развитие системы.

Разработка архитектуры системы включает в себя разработку следующих компонентов:

- спецификации требований к системе,
- конструирование концептуальной модели предметной области,
- спецификации обработки данных в системе,
- спецификации пользовательского интерфейса системы,
- спецификации деятельности в предметной области с учетом внедрения системы.

Команда проекта, необходимая для выполнения данных работ:

- Архитектор
- Инженер (системный администратор)

3.5.3.3 Разработка (доработка) порталов, сайтов, клиентской части отдельных систем⁶

Данный блок включает в себя работы по следующим основным направлениям:

- определение требуемых разделов портала, сайта, клиентской части отдельных систем,
- формулирование организационно-технических требований к административной системе управления порталом, сайтом, клиентской части отдельных систем,

⁵ Минимально необходимая реализация, описанная в настоящем документе, не подразумевает создания отдельной ИАС

⁶ Необязательный блок работ. В минимальной реализации возможно формирование отдельных отчетов без визуализации их на сайтах и порталах.

Проект: «Развитие статистики Содружества Независимых Государств»

- определение структуры портала, сайта, клиентской части отдельных систем с учетом поставленных задач, требуемых бизнес-процессов,
- формулирование требований к дизайну,
- разработка (доработка) портала, сайта, клиентской части отдельных систем,
- разработка(доработка) ядра системы (портала, сайта) и серверной логики,
- разработка (доработка) интеграционных сервисов для реализации функций взаимодействия портала, сайта с внешними системами, источниками данных.

Команда проекта, необходимая для выполнения данных работ:

- Frontend-разработчик ведущий
- Frontend-разработчик старший
- Frontend-разработчик
- Backend-разработчик ведущий
- Backend-разработчик старший
- Архитектор
- Аналитик ведущий
- Аналитик
- Специалист по тестированию ведущий
- Специалист по тестированию
- Инженер (системный администратор)

3.5.3.4 Разработка баз данных, API, интеграционных решений

Данный блок включает в себя работы по следующим основным направлениям:

- Разработка базы данных (далее – БД) содержит:
 - концептуальное проектирование – сбор, анализ и редактирование требований к данным,
 - логическое проектирование – преобразование требований к данным в структуры данных,
 - физическое проектирование – определение особенностей хранения данных, методов доступа и т.д.,
 - наполнение БД,
 - разработка витрин данных,
 - документирование.
- В разработку API входит:

Проект: «Развитие статистики Содружества Независимых Государств»

- создание прототипа API, используя шаблонный код,
 - тестирование API для предотвращения ошибок и дефектов,
 - документирование API с целью повышения удобства использования.
- Разработка интеграционных решений, ориентированных на обеспечение корректной работы информационных систем и на организацию «бесшовного» взаимодействия между разными сервисами и интерфейсами, предусматривает спектр работ по созданию инфраструктуры для интеграции программного обеспечения:
- сбор требований,
 - построение моделей процессов и бизнес-логики будущего решения,
 - выбор наиболее подходящей платформы или промышленного интеграционного продукта (разработка собственного интеграционного решения),
 - разработка целевой архитектуры,
 - тестирование, внедрение и сопровождение интеграционного решения.

Команда проекта, необходимая для выполнения данных работ:

- Backend-разработчик ведущий
- Backend-разработчик старший
- Backend-разработчик
- Frontend – разработчик старший
- Frontend – разработчик
- Архитектор
- Аналитик ведущий
- Специалист по тестированию ведущий
- Специалист по тестированию
- Инженер (системный администратор)

3.5.3.5 Разработка интерфейсов в готовых системах, картографических системах, отчетов, аналитических бюллетеней ⁷

- Проектирование пользовательского интерфейса,
- Разработка комфортного и понятного интерфейса в готовой ИАС или готовых картографических системах, а именно:

⁷ Минимально необходимая реализация, описанная в настоящем документе, не подразумевает создания интерфейсов.

- визуализация информации в соответствии с информационными потребностями пользователей (создание листов, необходимых фильтров, ключевых объемных показателей, создание объектов, написание формул, настройка работы фильтров),
 - создание структуры базы данных, обеспечивающей визуализацию информации с учетом требований, определяемых в ТЗ и/или спецификации готовой ИАС или картографической системы,
 - разработка функций загрузки данных в готовую ИАС или картографическую систему в объеме, необходимом и достаточном для формирования требуемых визуализаций.
- Создание отчетов разных видов:
- интерактивные отчеты, включающие результаты анализа данных, представленные в виде сложных таблиц, карт и/или диаграмм,
 - аналитические бюллетени, в которых излагаются сведения по определенным вопросам, собранным из множества разных источников.

Команда проекта, необходимая для выполнения данных работ:

- Frontend-разработчик ведущий
- Frontend-разработчик старший (2 человека)
- Frontend-разработчик
- Аналитик ведущий
- Аналитик
- Специалист по тестированию ведущий
- Специалист по тестированию
- Инженер (системный администратор)

3.5.3.6 Приобретение прав на использование программ для ЭВМ, баз данных по лицензионным договорам с правообладателем

По лицензионному договору правообладатель передает право на использование программ для ЭВМ, баз данных (программного продукта, программного обеспечения) в объеме, предусмотренном договором (на определенный срок, на определенной территории) другому лицу (лицензиату). Последний принимает на себя обязанность вносить лицензиару предусмотренные договором платежи (в случае если лицензионный договор

является возмездным), осуществлять иные действия, предусмотренные договором.

3.5.3.7 Облачные сервисы (SaaS-сервисы/ПО, как услуга).

SaaS (программное обеспечение как услуга, от английского Software as a Service) — это вариант оказания услуг, при котором экземпляр программы для ЭВМ или база данных пользователю во владение не передается, а ему предоставляется доступ к программному обеспечению через сеть Интернет на определенный (учетный) период.

При этом сам программный продукт (экземпляр ПО) расположен на серверах/оборудовании правообладателя и не требует локальной установки у пользователя.

Стоимость единицы Услуги (стоимость Услуги в месяц либо иной отчетный период) должна учитывать срок использования сервиса и количество пользователей.

Результатом оказания Исполнителем указанного вида Услуг будет являться доступ к ПО для определенного количества пользователей на условиях, предусмотренных соответствующим Заказом.

3.5.3.8 Техническая поддержка/сопровождение ПО

Исполнитель осуществляет техническую поддержку/сопровождение ПО в соответствии с условиями, представленными в договоре.

3.5.3.9 Действия по приобретению массивов данных

Данная услуга подразумевает приобретение массивов деперсонализованных данных (у сотовых операторов связи) для дальнейшего формирования агрегированных массивов данных о населении с единовременным охватом исследуемой территории и высоким уровнем пространственно-временной детализации (с помощью комплекса методов, алгоритмов и технологий) с последующим исследованием и статистическим анализом этих данных, исключая использование и обработку персональных данных людей на всех этапах сбора и обработки информации.

Для формирования статистических показателей о населении (геоаналитических данных) задействуются только технические данные сотовых сетей (радиочастотные события базовых станций), что определяет получение деперсонализированных данных от контрагентов и формирования агрегированных наборов количественных данных, содержащих информацию, необходимую для последующего получения на ее основе геоаналитических показателей.

3.5.3.10 Сбор, обработка, верификация и аналитика данных

Данный блок включает комплекс услуг:

- Сбор данных, включая создание баз данных и структур хранения данных для обеспечения проверок,
- Обработка исходных данных с использованием специальных методик и математических алгоритмов для дальнейшего формирования агрегированных наборов данных.
- Методики верификации наборов данных являются важной частью процесса формирования качественных валидных данных и включают в себя многоступенчатую проверку качества данных по критериям:
 - проверка непротиворечивости значений,
 - перекрестная проверка значений,
 - соотнесение с внешними открытыми источниками информации.
- Аналитика данных позволяет преобразовать массив данных в выводы, на основе которых будут приниматься решения и строиться действия. Она включает в себя ряд инструментов, технологий и процессов, используемых для поиска, обнаружения и интерпретации тенденций/закономерностей, а также для решения проблем с помощью данных. Аналитика данных помогает формировать отчеты, строить прогнозы и повышать эффективность принятия решений.

Команда проекта, необходимая для оказания данных услуг:

- Аналитик ведущий
- Аналитик старший
- Аналитик
- Специалист по тестированию ведущий
- Специалист по тестированию старший
- Backend-разработчик
- Инженер (системный администратор)

3.5.3.11 Управление проектом

Под управлением проектом (руководством проектом, управлением командами) понимаются услуги по решению задач и достижению поставленных целей проекта Заказчика в рамках утверждённого времени и бюджета.

Управление проектом делит рабочий процесс на части и контролирует бюджет, дедлайны и прогресс на каждом этапе. После завершения услуг можно оценить результаты по каждому процессу отдельно и в ракурсе конкретного проекта.

Управление проектом содержит в себе широкий спектр задач, включая, но не ограничиваясь:

- Постановка целей и задач;
- Проведение рабочих встреч (планерок, совещаний, собраний и т. д.);
- Построение системы эффективных коммуникаций в команде.

Управление командой включает в себя, но не ограничивается:

- Детализацию задач до уровня исполнения отдельным членом команды;
- Контроль технического стека, определенный при разработке архитектуры системы;
- Контроль процесса разработки и тестирования.

Команда проекта, необходимая для оказания данных услуг:

- Менеджер проектов
- Тимлид (Team Lead)

3.5.3.12 Консультирование, сопровождение проекта

В рамках процесса консультирования, сопровождения проекта Заказчика выделяют:

- Предоставление необходимых консультаций по формированию отчетов по численности населения на базе данных сотовых операторов, а также по эксплуатации системы в случае ее создания;
- Консультирование по выработке требований к формированию статистических отчетов на основе больших данных операторов сотовой связи,
- Участие в переговорах.

Команда проекта, необходимая для оказания данных услуг:

- Аналитик ведущий
- Инженер (Системный администратор) (в случае создания системы)

Таким образом, при формировании бизнес-модели процесса и расчете требуемого бюджета на его реализацию необходимо учитывать те работы и услуги, которые актуальны при реализации конкретного проекта.

В настоящем плане реализации не предусмотрено создание новой информационно-аналитической системы, поскольку речь идет о формировании готовых отчетов для их применения в производстве статистических оценок численности населения, поэтому из описанных выше работ и услуг при расчете бюджета мы учитываем следующие:

1. Разработка ТЗ, Технорабочего проекта, включая сбор и подготовку требований.
2. Разработка баз данных.
3. Разработка отчетов, аналитических бюллетеней, другой отчетной документации.
4. Облачные сервисы (SaaS-сервисы/ПО, как услуга).
5. Действия по приобретению массивов данных.
6. Сбор, обработка, верификация, агрегация и аналитика данных.
7. Управление проектом (руководство проектом, управление командами).
8. Консультирование, сопровождение проекта.

4 Требования к проектированию типового бизнес-процесса

4.1 Требования к проектированию типовых выходных материалов

4.1.1 Требования к формату выходных материалов

Требование 4.1.1.1. Для производства статистической информации о численности населения на основании данных операторов сотовой связи статистическое ведомство, выступающее в роли Заказчика, устанавливает:

- Период учета населения (как правило, календарный год). Оценка численности населения выполняется в установленный период ежемесячно, по состоянию на 23 часа 59 минут последнего числа каждого календарного месяца из установленного периода учета населения. Необходимость установки условного момента времени для ежемесячного учета связана с непрерывным изменением населения (рождения, смерти, переезды людей из одного места жительства в другое).
- Территорию, по которой необходимо выполнить оценку численности населения (исследуемая территория). Как правило, оценка численности населения выполняется для всей территории государства.
- Перечень выделяемых на исследуемой территории зон территориального деления, для каждой из которых необходимо выполнить оценку численности населения. Как правило, выделяемые для производства статистической информации зоны территориального деления соответствуют административно-территориальному устройству государства и устанавливаются Заказчиком на основании существующих общегосударственных классификаторов.
- Перечень возрастно-половых групп, по которым необходимо выполнить оценку численности населения. Как правило, при выполнении оценки численности населения по данным операторов сотовой связи, каждая возрастно-половая группа включает в себя возрастной диапазон 5 лет и более.

Требование 4.1.1.2. Результатом выполнения работ по производству статистической информации о численности населения на основе данных сотовых операторов являются следующие материалы:

- статистические данные о численности проживающего на исследуемой территории населения;
- справочник территориального деления;
- справочник возрастно-половых групп;

- отчет об оценке качества статистических данных о численности населения;
- данные операторов сотовой связи о численности населения (предварительные агрегаты);
- отчет об оценке качества предварительных агрегатов.

Требование 4.1.1.3. Статистические данные о численности населения содержат информацию о количестве лиц, постоянно и временно проживающих на территории каждой зоны территориального деления, в каждом календарном месяце из установленного Заказчиком периода учета населения.

Статистические данные о численности населения представляют собой текстовый csv-файл на электронном носителе информации, содержащий следующие поля:

- дата и время момента учета населения (календарный месяц);
- код зоны территориального деления;
- код возрастно-половой группы;
- количество лиц заданной возрастно-половой группы, постоянно проживающих на территории заданной зоны территориального деления;
- количество лиц заданной возрастно-половой группы, временно проживающих на территории заданной зоны территориального деления.

Требование 4.1.1.4. Справочник территориального деления представляет собой текстовый csv-файл на электронном носителе информации, как минимум содержащий следующие поля:

- код зоны территориального деления;
- название зоны территориального деления;
- код административно-территориальной единицы в соответствии с общегосударственным классификатором; либо описание границ зоны территориального деления в формате WKT с использованием системы координат WGS84, в том случае, если зона территориального деления не входит в общегосударственные классификаторы территорий.

Требование 4.1.1.5. Справочник возрастно-половых групп представляет собой текстовый csv-файл на электронном носителе информации, как минимум содержащий следующие поля:

- код возрастно-половой группы;
- возрастной диапазон;
- пол.

Требование 4.1.1.6. Отчет об оценке качества статистических данных о численности населения представляет собой текстовый документ на электронном носителе информации и содержит результаты оценки качества итоговых агрегатов, выполненные в соответствии с требованиями, которые изложены в разделе 6.4.

Требование 4.1.1.7. Данные операторов сотовой связи о численности населения – это полученные от операторов сотовой связи предварительные агрегаты. Требования к формату предварительных агрегатов приведены в разделе 5.3.1.

Требование 4.1.1.8. Отчет об оценке качества предварительных агрегатов представляет собой текстовый документ на электронном носителе информации и содержит результаты оценки качества предварительных агрегатов, выполненной в соответствии с требованиями, которые изложены в разделе 5.4.

4.1.2 Типовая дорожная карта реализации бизнес-модели

Требование 4.1.2.1. Процесс производства статистических оценок численности населения по данным операторов сотовой связи состоит из следующих этапов:

1. Статистическое ведомство, выступающее в роли Заказчика, на выполнение работ по производству статистических оценок численности населения по данным сотовых операторов, устанавливает требования к выходным материалам в соответствии с разделом 4.1.1, а именно: устанавливает период учета населения и территорию исследования; определяет перечень зон территориального деления и перечень возрастно-половых групп. В соответствии с установленными Заказчиком требованиями к выходным материалам составляется техническое задание на выполнение работ по производству статистических оценок численности населения.

2. Определяется подрядная организация (Подрядчик), удовлетворяющая требованиям, которые изложены в разделе 5.1.2, и готовая выполнить работы в соответствии с составленным Заказчиком на этапе 1 техническим заданием. Между Заказчиком и Подрядчиком заключается договор на выполнение работ по производству статистических оценок численности населения по данным операторов сотовой связи.

3. Подрядчик составляет техническое задание для сотовых операторов на поставку предварительных агрегатов в соответствии с требованиями, изложенными в разделе 5.2, и техническим заданием, составленным Заказчиком на этапе 1.

4. Подрядчик определяет необходимое и достаточное количество поставщиков статистических данных (операторов сотовой связи), которое обеспечит полноту покрытия исследуемой территории и охват населения, в соответствии с требованиями изложенными в разделе 4.3.2. Подрядчик выбирает известное количество поставщиков среди сотовых операторов, которые оказывают услуги связи на исследуемой территории, удовлетворяют требованиям, изложенным в разделе 5.1.1, и готовы выполнить работы в соответствии с составленным на этапе 3 техническим заданием. Подрядчик заключает договор на поставку предварительных агрегатов с каждым из выбранных поставщиков.

5. Поставщик обрабатывает первичную информацию о событиях сотовой сети для определения местоположения абонентов в соответствии с требованиями, изложенными в разделе 4.3.3. Поставщик формирует предварительные агрегаты в соответствии с требованиями, изложенными в разделе 5.3.2, и составленным на этапе 3 техническим заданием. Поставщик передает сформированные предварительные агрегаты Подрядчику.

6. Подрядчик оценивает качество полученных от операторов сотовой связи предварительных агрегатов в соответствии с требованиями, изложенными в разделе 5.4. Подрядчик формирует на основании полученных предварительных агрегатов в соответствии с требованиями, изложенными в разделе 6.3, итоговый агрегат – статистические данные о численности постоянно проживающего на исследуемой территории населения.

7. Подрядчик оценивает качество итогового агрегата в соответствии с требованиями, изложенными в разделе 6.4, и готовит отчет об оценке качества статистических данных о численности населения. Подрядчик передает Заказчику подготовленные выходные материалы: статистические данные о численности постоянно проживающего на исследуемой территории населения; справочник территориального деления; справочник возрастно-половых групп; отчет об оценке качества статистических данных о численности населения; предварительные агрегаты; отчет об оценке качества предварительных агрегатов.

Сроки каждого из перечисленных этапов должно согласовать Статистическое ведомство таким образом, чтобы срок каждого этапа соответствовал дорожной карте комплексного процесса, осуществляемого Статистическим ведомством. В среднем подготовка данных о численности населения на базе данных сотовых операторов занимает несколько месяцев.

4.2 Требования к проектированию типовых описаний переменных

Требование 4.2.1. В статистических данных о численности населения учитывается постоянное население, под которым понимаются лица, фактически проживающие на территории государства. К постоянному населению относятся: граждане государства, иностранные граждане и лица без гражданства, при условии проживания в стране в течение последнего месяца.

Требование 4.2.2. Отдельной категорией в статистических данных учёта населения считаются временно находящиеся на территории государства лица, постоянное место жительства которых расположено за границей. К ним относятся: граждане государства, иностранные граждане и лица без гражданства, находящиеся на территории государства на момент учета, но проживавшие в течение последнего месяца за границей.

Требование 4.2.3. При формировании статистических данных о численности населения не подлежат учету следующие категории лиц: граждане государства, иностранные граждане и лица без гражданства отсутствующие на территории государства на момент учета, при условии проживания за границей в течение последнего месяца.

Требование 4.2.4. Учет постоянно проживающего в стране населения осуществляется по месту фактического проживания, под которым понимается территориальная единица, где лицо преимущественно пребывало в ночное время (с 23:00 до 06:00 местного времени) на протяжении последнего месяца.

Для лиц, временно находящихся в стране (постоянно проживающих за рубежом), местом учета признается место их фактического проживания на момент проведения учета.

4.3 Требования к проектированию сбора данных

4.3.1 Требования к описанию типовых принципов построения сотовых сетей и событий, фиксируемых сетями операторов сотовой связи

Требование 4.3.1.1. Основной принцип устройства сотовой сети оператора связи – деления территории работы оператора на ячейки – соты, в центре каждой из которых находится базовая станция. Для каждой базовой станции известны точные координаты ее расположения. На базовой станции располагается антенна, для которой должно быть известно:

- азимут - направление главного лепестка антенны в горизонтальной плоскости;
- угол места – угол наклона антенны в вертикальной плоскости;
- мощность передатчика – т.е. мощность сигнала, которую излучает базовая станция;

- диаграмма направленности антенны – определение ширины главного лепестка и уровень боковых лепестков.

Все вышеприведенные атрибуты должны храниться у Оператора в виде справочника базовых станций и сот.

Оператор сотовой связи должен перед началом расчета очередного отчетного периода обновить справочник базовых станций и сот привязке к геометрии исследуемой территории.

Требование 4.3.1.2. Оператор сотовой связи должен перед началом расчета очередного отчетного периода отфильтровать и оставить для рассмотрения лишь те устройства, которые набрали за рассматриваемый период не менее 10 минут голосового трафика.

Требование 4.3.1.3. Оператор сотовой связи должен иметь возможность выделять следующие типы событий:

- Начало звонка;
- Окончание звонка;
- Пересечение границ сот во время звонка (хэндовер);
- Начало пакетной сессии;
- Окончание пакетной сессии;
- Пересечение границ сот во время активной пакетной сессии (хэндовер);
- Звонок;
- Смс;
- Пересечение границы LAC (локации внутри сот, требуется для уточнения местоположения абонента);
- Включение терминала;
- Выключение терминала;
- Регистрация терминала в сети после потери покрытия;

Требование 4.3.1.4. Оператор сотовой связи должен иметь возможность выделять следующие атрибуты для каждого рассматриваемого события:

- MSISDN - идентификатор абонента
- EventTime – дата и время обработки события (с точностью до секунды)
- EventType - тип события
- LAC, Cell ID - сота, в которой было обработано событие

4.3.2 Требования к разработке принципов определения необходимого и достаточного количества сотовых операторов учетом обеспечения полноты покрытия территории страны и охвата населения

Требование 4.3.2.1. При проведении статистического исследования на основе данных сотовых операторов необходимо выполнять комплексное обоснование выбора поставщиков данных, которое должно включать:

- Количественный критерий участников - минимальное количество операторов связи для обеспечения статистической значимости - не менее двух.
- Территориальный критерий покрытия: совокупное покрытие исследуемой территории от выбранных операторов должно составлять не менее 50% от общей площади исследуемой территории.
- Демографический критерий охвата: совокупная абонентская база выбранных операторов должна охватывать не менее 50% от общей численности на исследуемой территории.

Данные пороги (50% территории и 50% населения) являются минимальными для обеспечения базовой репрезентативности, но могут меняться в зависимости от целей исследований.

Требование 4.3.2.2. При работе с геоаналитическими данными требуется обосновать минимально возможную абонентскую базу. Необходимость обеспечения требуемой точности геоаналитических данных определяет требования к размеру активной на исследуемой территории абонентской базы одного или нескольких сотовых операторов.

Результаты статистических расчетов сотового оператора в заданный момент времени t определяются кумулятивной функцией $F(x, y; t)$ распределения вероятностей численности населения, где x, y – декартовы координаты, соответствующие текущей точке конкретной зоны разбиения, zid , территории исследуемого региона R .

Критерий Колмогорова-Смирнова об эмпирической функции распределения вероятностей $F^{(n)}(x, y; t)$, определенной по конечной выборке, позволяет определить объем выборки n , необходимый для оценки эмпирической функции распределения вероятностей, независимо от ее вида, с заданной надежностью. В двумерном случае, как показано в работах⁸, можно

⁸ Peacock J.A. Two-dimensional goodness testing in astronomy. Royal Astronomical Society. Provided by the NASA Astrophysics Data System, 1983; V.202; p. 615-627.

G. Fasano, A. Franceschini, "A multidimensional version of the Kolmogorov–Smirnov test", Monthly Notices of the Royal Astronomical Society, Volume 225, Issue 1, March 1987, Pages 155–170, <https://doi.org/10.1093/mnras/225.1.155>

оценить объем выборки n (численность активной абонентской базы оператора на территории исследуемого региона R), необходимый для достижения заданного уровня значимости α и заданной точности оценивания истинной функции распределения, например, 0,5%. Для величины максимального расхождения D_n между истинной и эмпирической функциями распределения:

$$D_n = \sup_{x,y} |F^{(n)}(x, y; t) - F(x, y; t)| \quad (2)$$

существует оценка [17], представляющая собой разложение в степенной ряд:

$$D_n = \sum_{i=0}^m \sum_{j=0}^{m-j} \sum_{k=0}^{m-i-j} a_{ijk} u^i v^j w^k, \quad (3)$$

где $u = \log\left(\frac{1,268-CC}{1,41}\right)$,

CC – параметр истинной функции распределения, характеризующей её симметрию,

$$v = \log(\alpha),$$

$$w = s \frac{4,989}{4,989-s},$$

$$s = \log(n) + 1.074.$$

В Таблице 4.3.2.2.1 приведены значения коэффициентов первых трех порядков.

Таблица 4.3.2.2.1 Коэффициенты разложения в ряд величины максимального расхождения истинной и эмпирической функцией распределения

i	j	k	aijk
0	0	0	0,710700
0	0	1	0,009863
0	0	2	0,008561
0	0	3	0,002800
0	1	0	-0,512600
0	1	1	-0,253900
0	1	2	0,037450
0	2	0	-0,329800
0	2	1	-0,014080
0	3	0	-0,076930
1	0	0	-0,061240
1	0	1	-0,191800
1	0	2	0,106400
1	1	0	-0,504200
1	1	1	0,036700
1	2	0	0,077500
2	0	0	0,777500

2	0	1	-0,232200
2	1	0	-0,127300
3	0	0	-0,991500

Доверительный интервал уровня $1-\alpha$ при любой истинной функции распределения $F(x,y;t)$ и для эмпирической функции распределения $F^{(n)}(x,y;t)$, полученной на основе выборки объема n , имеет следующий вид:

$$P\{F^{(n)}(x,y;t) - D_n < F(x,y;t) < F^{(n)}(x,y;t) + D_n: \forall x,y \in R\} = 1 - \alpha,$$
 (4)

где R – территория исследуемого региона.

Условие (4) позволяет для заданного доверительного уровня $1-\alpha$ определить необходимый объем выборки n (Таблица 4.3.2.2.2).

Таблица 4.3.2.2.2. Значения объема выборки n , необходимые для оценки кумулятивной функции распределения вероятностей с различными уровнями значимости α при точности оценивания функции распределения 0,005 (0,5%).

Уровень значимости α (вероятность ошибки I рода)	Точность оценки функции распределения вероятностей	Минимальный необходимый объем выборки
0,1	0,005	59915
0,05	0,005	73778
0,025	0,005	87641
0,01	0,005	105966
0,005	0,005	119829
0,001	0,005	152018
0,0005	0,005	165881
0,0001	0,005	198070
0,00005	0,005	211933
0,00001	0,005	244121
0,000005	0,005	257984

Для пространственной плотности населения критерий Колмогорова-Смирнова гарантирует, что её отклонения от истинной пространственной плотности не будут носить масштабный характер, чтобы существенно повлиять на кумулятивную функцию распределения $F^{(n)}(x,y;t)$. Выбор уровня значимости α может определяться бизнес требованиями.

При геоаналитических расчетах вычислять функцию распределения вероятностей нет необходимости. Однако оценки точности функции распределения представляют собой обоснование точности вышеуказанных расчетов по показателям операторов сотовой связи.

Как свидетельствуют данные, представленные в Таблице 4.3.2.2.2, для формирования статистически надёжных геоаналитических данных о численности населения на всей исследуемой территории в целом, достаточно данных одного крупного оператора. Однако, для получения геоаналитических данных с высокой детализацией по пространству и по времени, необходимо учитывать методические ограничения точности расчётов мобильного оператора, связанные с дискретизацией оператором траекторий абонентов, представлением данных по секторам 500x500 м и с другими аппроксимациями. Кроме того, в работе операторов сотовой связи могут иметь место сбои.

Требование 4.3.2.3. Методика должна включать в себя использование данных двух операторов, даже если абонентской базы одного оператора достаточно для необходимой точности геоаналитических данных в связи с потенциальными сбоями в работе операторов, как в следствии, и их данных. При использовании данных только одного оператора не исключена вероятность влияния случайных сбоев в системе сбора первичных данных на оборудовании оператора или особенностей сети этого оператора на результирующие расчеты геоаналитических данных. Для выявления ошибок в данных, связанных с такими событиями, используются методы перекрестных проверок данных двух и более операторов, аналогичных тем, которые использовались для выявления аномалий с помощью критерия χ^2 (1). В качестве теоретического распределения численности населения, E_j , используется распределение соответствующего показателя одного из операторов (например, второго), взвешенное относительно доли этого оператора в рассматриваемом регионе, а в качестве проверяемого, n_j – распределение для другого оператора (например, первого), также взвешенное относительно его доли в рассматриваемом регионе. Если в результате статистического теста нулевая гипотеза будет отвергнута, то проверка выявит неслучайные расхождения между данными двух операторов. Это расхождение будет свидетельствовать о вероятных ошибках в данных одного из операторов (по формальному смыслу критерия χ^2 – первого).

Если, благодаря другим тестам, будут выявлены подобные друг другу аномалии в данных обоих операторов, то будет обоснованно предположить, что наблюдается не случайная ошибка в данных, а особенность, имеющая

объяснимую причину, например, городское событие, которое нашло отражение в статистических данных.

Подобный вывод можно получить следующим образом. Любой оператор сотовой связи имеет возможности собирать и обрабатывать аппаратные события своей сотовой сети только на основе тех сигналов, которые формируются его абонентами, т.е. пользователями только его сети. При этом даже использование данных всех существующих сотовых операторов не дает 100% охват всего населения. Если k операторов охватывают всех N абонентов рассматриваемого региона с долями охвата λ_i , $i=1, \dots, k$, $\lambda_1 + \dots + \lambda_k = 1$, то надежность выводов на основе данных каждого i -го оператора в предположении о случайности ошибок в соответствии с основами теории надежности больших систем⁹ может быть охарактеризована вероятностью P_i :

$$P_i = 1 - \frac{\sqrt{-\frac{1}{2} \ln \frac{\alpha}{2}}}{\sqrt{N_i}} \quad (5)$$

$$N_i = N \lambda_i, \quad i = 1, \dots, k, \quad (6)$$

где α – вероятность ошибки первого рода. Эта оценка следует из теоремы Колмогорова о сходимости эмпирической функции распределения.

Применение данных нескольких операторов представляет позволяет повысить надежность посредством резервирования с избыточностью¹⁰. Поэтому суммарная надежность выводов по данным k операторов, а именно, вероятность правильного представления информации, равна P :

$$P = 1 - (1 - P_1)(1 - P_2) \dots (1 - P_k). \quad (7)$$

С учётом рисков ошибок, связанных не только со статистикой, но и с другими факторами, показатель надежности, достижимый с использованием нескольких операторов, существенно выше, чем надежность данных одного оператора со сколь угодно большим охватом.

Например, при численности населения региона $N = 10,000,000$ человек, доле охвата населения двумя операторами $\lambda_1 = 0.35$, $\lambda_2 = 0.30$, уровне значимости $\alpha = 0.001$, значения надежности выводов по данным одного оператора будут равны, соответственно, $P_1 = 1 - 1.04 \cdot 10^{-3}$ и $P_2 = 1 - 1.13 \cdot 10^{-3}$. Надежность при совместном использовании данных двух операторов равна $P_{12} = 1 - (1 - P_1)(1 - P_2) = 1 - 1.17 \cdot 10^{-6}$. Для рассматриваемого примера,

⁹ Гнеденко Б.В., Беляев Ю.К., Соловьёв А.Д. Математические методы в теории надёжности- М.: Наука, 1965. -524 с.

ГОСТ 27 002-2015. Надёжность в технике. – М.: Стандартиформ, 2016. - 24 с.

¹⁰ Гнеденко Б.В., Беляев Ю.К., Соловьёв А.Д. Математические методы в теории надёжности- М.: Наука, 1965. -524 с.

ГОСТ 27 002-2015. Надёжность в технике. – М.: Стандартиформ, 2016. - 24 с.

привлечение данных третьего оператора с рыночной долей $\lambda_3 = 0.25$ и надежностью $P_3 = 1 - 1.23 \cdot 10^{-3}$ позволяет повысить статистическую надёжность совместных данных до уровня $P_{123} = 1 - 1.45 \cdot 10^{-9}$. Требования к надёжности определяются бизнес-требованиями, поэтому для конкретных задач привлечение данных третьего оператора будет создавать ограниченную дополнительную ценность для статистической надёжности данных геоаналитики по отношению к обсуждаемым здесь рискам.

4.3.3 Требования к типовым принципам обработки первичных данных сотовых сетей для определения местонахождения абонентов в требуемый временной интервал

Требование 4.3.3.1. Первичными данными сетевой активности каждого абонента являются события, фиксируемые на сотовой сети. Для целей данного исследования выбираются события, описанные в требовании 4.3.1.3. Для каждого исследования должен быть определен временной период, в рамках которого рассматриваются события абонентов. Обработка первичных данных заключается в последовательной обработке событий каждого абонента в соответствии с нижеописанными требованиями.

В начало рассматриваемого периода времени добавляется вспомогательное событие, имеющие следующие свойства:

MSISDN	EventTime	EventType	LAC, Cell ID
Идентификатор рассматриваемого абонента	Начало рассматриваемого временного интервала Например, 2025.02.01. 00:00:00	Включение терминала	Сота, в которой было обработано первое событие рассматриваемого абонента в рассматриваемый временной интервал

В конец рассматриваемого периода времени добавляется вспомогательное событие, имеющие следующие свойства:

MSISDN	EventTime	EventType	LAC, Cell ID
Идентификатор рассматриваемого абонента	Конец рассматриваемого временного интервала	Выключение терминала	Сота, в которой было обработано последнее событие рассматриваемого абонента в

	Например, 2025.08.31. 23:59:59		рассматриваемый временной интервал
--	-----------------------------------	--	--

Требование 4.3.3.2. Основываясь на таблице событий абонентам необходимо сформировать таблицу, которая будет содержать интервалы пребывания абонента в каждой соте. Такая таблице называется таблицей временных интервалов.

Таблица временных интервалов имеет следующий набор столбцов:

- MSISDN – идентификатор абонента
- TimeInterval – интервал времени между моментами времени начала интервала (TimeStart) и окончания интервала (TimeEnd). При записи интервала квадратная (круглая) скобка означает, что конечный момент времени входит (не входит) в интервал.
- CellList - список сот, входящих в локацию пребывания абонента
- FirstEvent - время первого события в интервале
- LastEvent - время последнего события в интервале

Обработка таблицы событий идет построчно. Одновременно рассматривается две строки, т.е. два последовательных события:

Пусть для абонента зафиксировано 2 последовательных события, которые произошли в моменты времени T1 и T2 и были обработаны в сотах C1 и C2 соответственно.

Если выполнено хотя бы одно из следующих условий:

- А) между событиями прошло больше 24 часов
- Б) событие, произошедшее в момент времени T1, имеет тип «Выключение терминала»
- В) событие, произошедшее в момент времени T2, имеет тип «Включение терминала»

Тогда местоположение абонента в интервал времени (T1, T2) считается неизвестным.

В таблицу временных интервалов добавляются две строки:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	[T1, T1]	C1	T1	T1
Идентификатор абонента	(T1, T2)	Н/Д	Н/Д	Н/Д

Если не выполнено ни одно из условий из пункта А), а событие, произошедшее в момент времени T_2 , имеет тип “Пересечение границы LAC”, тогда считается, что абонент все время (T_1, T_2) находился в соте C_1 .

В таблицу временных интервалов добавляется строка:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	$[T_1, T_2)$	C_1	T_1	T_1

Если ни одно из условий пункта А) и Б) не выполнено, тогда считается, что в течение промежутка $(T_1, (T_1+T_2)/2]$ абонент находился в соте C_1 , а в течение промежутка $((T_1+T_2)/2, T_2)$ – в соте C_2 .

В таблицу временных интервалов добавляется две строки:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	$[T_1, (T_1+T_2)/2]$	C_1	T_1	T_1
Идентификатор абонента	$((T_1+T_2)/2, T_2)$	C_2	T_2	T_2

Таким образом, в сформированной таблице интервалов содержится информация о наборе сот, в которых находился абонент в любой момент времени.

Требование 4.3.3.3. Должны быть исключены непродолжительные периоды времени, во время которых местоположение абонента было не определено.

В случае, если временной интервал имеет $CellList = \text{Н/Д}$ и его длительность меньше 10 минут, то этот временной интервал прибавляется к предыдущему, т.е. две последовательные строчки из таблицы интервалов:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	$[T_{12}, T_{13}]$	C_8	T_{117}	T_{118}
Идентификатор абонента	(T_{13}, T_{14})	Н/Д	Н/Д	Н/Д

Заменяются одной:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	$[T_{12}, T_{14}]$	C_8	T_{117}	T_{118}

Требование 4.3.3.4. Полученная таблица временных интервалов должна быть уточнена путем объединения временных интервалов, составляющих непродолжительный цикл.

Назовем циклом последовательность, которая состоит из двух или большего количества временных интервалов, причем набор сот последнего временного интервала совпадает или полностью содержится в наборе сот первого временного интервала в последовательности:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	[T15, T16)	C9	T119	T120
Идентификатор абонента	[T17, T18)	C10	T127	T128
Идентификатор абонента	[T19, T20)	C11	T129	T130
Идентификатор абонента
Идентификатор абонента	[T21, T22)	C9' (причем C9' ⊆ C9)	T131	T132

При этом ни один из наборов сот в последовательности не должен быть равен Н/Д, а также ни одна из сот не должна относиться к подземным станциям метрополитена.

Длительность цикла равна времени, которое прошло между событиями, произошедшими в одном и том же наборе сот, т.е. для приведенного примера длительность цикла равна T131 – T120.

В случае, если длительность цикла меньше 5 минут, входящие в цикл интервалы объединяются в один интервал:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	[T15, T22)	C9, C10, C11, ...	T119	T132

После этого продолжается построчная обработка таблицы, начиная с добавленной строки.

Требование 4.3.3.5. Полученная таблица временных интервалов должна быть уточнена путем объединения последовательных интервалов времени, в которые абонент пребывал в соседних сотах. Соседние соты определяются исходя из геометрического пересечения сот, которые определены в справочнике базовых станций, описанным в требовании 4.3.1.1.

На данном этапе таблица интервалов времени уточняется при помощи справочника соседних базовых станций. Обработка таблицы идет построчно. Одновременно рассматривается две строки, т.е. два последовательных интервала времени, для которых местоположение абонента известно (т.е. значение столбца CellList не равно Н/Д) и кроме того, ни в одном из временных интервалов он не находился в подземной части метрополитена:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	[T6, T7)	C4 = {C4a, C4б, C4в, ...} – список из одной или нескольких сот	T113	T114
Идентификатор абонента	(T7, T8)	C5	T123	T124

Если среди сот из множества {C4a, C4б, C4в, ..., C5} есть хотя бы одна, для которой все соты из данного множества (кроме нее самой) являются соседними, то рассматриваемые две строки заменяются одной:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	[T6, T8)	C4a, C4б, C4в, ..., C5	T113	T124

После этого продолжается построчная обработка таблицы, начиная с добавленной строки.

Требование 4.3.3.6. В таблицу временных интервалов должны быть добавлены поля, показывающего, находился ли абонент в движении.

Для каждого временного интервала определяется, был ли абонент неподвижен или находился в движении. В таблицу, полученную на предыдущем этапе, добавляется столбец Status, значение которого для каждой строчки определяется следующим образом:

- Если местоположение абонента неизвестно (значение столбца CellList = Н/Д), то Status = Н/Д.
- Если местоположение абонента известно (CellList не равно Н/Д), тогда:
- Если длительность временного интервала (TimeEnd – TimeStart) не меньше 60 минут, то абонент считается неподвижным и в Status записывается Stay.

- Если длительность временного интервала меньше 60 минут, то абонент считается передвигающимся и в Status записывается Move.
- Если все соты из CellList относятся к подземным станциям метрополитена, то вне зависимости от длительности временного интервала в Status записывается Move.

Требование 4.3.3.7. Полученная таблица временных интервалов должна быть обработана путем добавления вспомогательных передвижений при двух последовательных неподвижных состояниях в разных сотах.

Будем считать, что абонент, который был неподвижен сначала в одном наборе сот, а потом в другом наборе сот, обязательно совершает между этими неподвижными состояниями передвижение.

Таким образом, между каждыми двумя последовательными временными интервалами, имеющими Status = Stay:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent	Status
Идентификатор абонента	[T9, T10]	{C6a,C6б,...}	T115	T116	Stay
Идентификатор абонента	(T10, T11)	{C7a,C7б,...}	T125	T126	Stay

Добавляется вспомогательная строчка, имеющая Status = Move:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent	Status
Идентификатор абонента	[T9, T10)	{C6a,C6б,...}	T115	T116	Stay
Идентификатор абонента	[T10, T10]	Н/Д	T10	T10	Move
Идентификатор абонента	(T10, T11)	{C7a,C7б,...}	T125	T126	Stay

Таким образом, таблица интервалов каждого абонента содержит информацию о наборе сот, в которых находится абонент в любой момент за рассматриваемый период времени. Кроме того, в ней есть информация, находился ли абонент в неподвижном состоянии (Status = Stay) или совершал поездку (Status = Move). В дальнейшем, под местоположением абонента в любой момент времени будет пониматься набор сот из колонки CellList того временного интервала, который содержит рассматриваемый момент времени.

4.4 Требования к проектированию обработки и анализа

Требование 4.4.1. Методика должна включать алгоритм определения временных интервалов нахождения абонента дома.

На основании таблицы временных интервалов определяется время, которое абонент провел в каждой соте ночью и днем. Для определения времени пребывания абонента в соте C1, учитываются все временные интервалы из таблицы, для которых в наборе сот (CellList) содержится C1.

Для определения времени ночного пребывания абонента в соте рассматривается время с 23:00 до 6:00 за все дни месяца (с учетом праздников и выходных). Из всех сот, в которых абонент был зафиксирован в ночное время, выбирается та, в которой он провел максимальное время ночью в рассматриваемом календарном месяце. Сот, удовлетворяющих этому условию максимальности, может быть несколько (в случае, если в нескольких удовлетворяющих указанному выше условию сотах абонент провел одинаковое время), тогда из них выбирается одна сота случайным образом. Время, которое абонент провел ночью в выбранной соте, будем называть ночным временем абонента, а саму выбранную соту – ночной сотой.

Если ночное время абонента больше определенного значения, например $(0.185 * \text{<количество дней в отчетном периоде> * 7)$ часов, то ночная сота будет называться основной домашней сотой абонента. В противном случае считается, что для данного абонента основная домашняя сота не определена. Например, для отчетов за март 2025 года основная домашняя сота будет определена для абонентов, ночное время которых превышает $0.185 * 31 * 7 = 40.145$ часов.

Стоит обратить внимание, что согласно введенной терминологии для каждого абонента определено ночное время (оно может быть равно, например, нулю) и ночная сота. При этом для абонента может не существовать основной домашней соты.

Для каждого абонента, для которого была определена основная домашняя сота, формируется набор домашних сот. Для этого перебираются все временные интервалы данного абонента за рассматриваемый календарный месяц. В случае, если для рассматриваемого временного интервала в поле CellList встречается основная домашняя сота, то все соты из данного списка CellList включаются в набор домашних сот.

Таким образом, для каждого абонента определяется ночная сота и ночное время. Кроме того, для некоторых абонентов определяется основная домашняя, основная дневная, а также набор домашних сот (либо их отсутствие).

В таблицу интервалов времени добавляется столбец POI_Home, значение которого будет показывать, находится ли абонент в данном интервале времени дома. В случае, если в рассматриваемый временной интервал хотя бы одна из сот в поле CellList входит в набор домашних сот, в столбец POI_Home записывается Home.

Требование 4.4.2. Методика должна включать алгоритм привязки временного интервала к зоне разбиения

Для каждого временного интервала известен набор сот (CellList), в которых находился абонент. Для всех временных интервалов каждой соте из набора сот CellList ставится в соответствие область на плоскости – область покрытия соты – в соответствии с данными радиопланирования. Назовем локацией, соответствующей набору сот, сумму областей покрытия сот, входящих в данный набор.

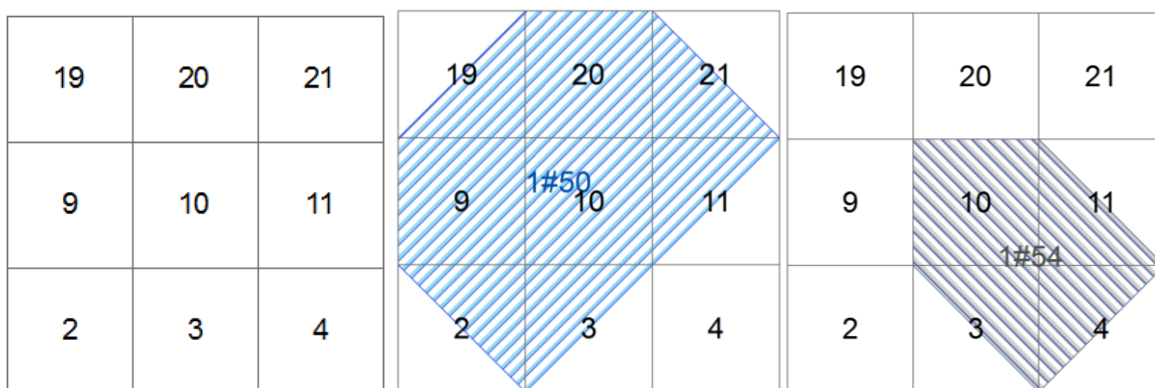
Если локация не имеет общих точек ни с одной зоной территориального деления (включая вспомогательные зоны), то зона нахождения абонента не определена. В противном случае, зона нахождения абонента для административного разбиения территории определяется следующим образом:

Определяются набор допустимых зон разбиения, в который входят все зоны, пересекающиеся с локацией.

Локация приписывается к зоне разбиения из набора допустимых зон, на территории которой (зоны) находится эта локация. Если разные части одной и той же локации находятся в разных зонах разбиения, то локация приписывается к зоне разбиения случайным образом, причем вероятность приписывания локации к зоне разбиения должна быть пропорциональная площади пересечения локации с зоной. При вычислении площади пересечения локации с зоной, следует вычислять сумму площадей пересечения каждой области покрытия соты, из которых состоит локация, с зоной разбиения. Таким образом, при вычислении площади, места наложения областей покрытия сот друг на друга будут учитываться несколько раз.

Приведем пример:

Пусть у нас есть разбиение на зоны, в котором каждая зона является квадратом со стороной 1 км, и локация, которая состоит из двух сот: 1#50 и 1#54. Зоны покрытия сот устроены так, как на рисунках ниже.



В таблице ниже приведены вычисления вероятности нахождения абонента в каждой из зон разбиения, соответствующие локации (1#50, 1#54).

Номер зоны	2	3	4	9	10	11	19	20	21	Сумма
Площадь пересечения зоны с сотой 1#50, кв. км	0.5	0.5	0	1	1	0.5	0.5	1	0.5	5.5
Площадь пересечения зоны с сотой 1#54, кв. км	0	0.5	0.5	0	1	0.5	0	0	0	2.5
Площадь пересечения зоны с набором сот, кв. км	0.5	1	0.5	1	2	1	0.5	1	0.5	8
Вероятность нахождения абонента в зоне	1/16	2/16	1/16	2/16	4/16	2/16	1/16	2/16	1/16	1
Нижняя граница P1	0	0.0625	0.1875	0.25	0.375	0.625	0.75	0.8125	0.9375	
Верхняя граница P2	0.0625	0.1875	0.25	0.375	0.625	0.75	0.8125	0.9375	1	

Для выбора одной зоны с нужной вероятностью осуществляется генерация случайного числа P в диапазоне от 0 до 1 ($0 \leq P < 1$). В зависимости

от значения P выбирается та зона разбиения, для которой выполнено двойное неравенство: $P_1 \leq P < P_2$.

После того, как определена зона нахождения абонента для рассматриваемого временного интервала, следует перейти к следующему временному интервалу и повторить алгоритм выбора зоны сначала (даже в том случае, если локация нахождения абонента в новом временном интервале совпадает с только что рассмотренным временным интервалом).

4.5 Требования к проектированию производственных систем и процесса

Требование 4.5.1. Порядок взаимодействия Подрядчика с Поставщиками данных:

Этап 1. Разработка технического задания.

Подрядчик осуществляет разработку технического задания на поставку предварительных агрегатов, в соответствии с изложенными в разделе 5.2.2 требованиями.

Разработанное техническое задание подлежит согласованию со всеми привлекаемыми Поставщиками в целях подтверждения его исполнимости и устранения неоднозначностей.

Этап 2. Отбор Поставщиков.

Подрядчик определяет необходимое и достаточное количество поставщиков статистических данных (операторов сотовой связи), которое обеспечит полноту покрытия территории и охват населения, в соответствии с требованиями изложенными в разделе 4.3.2.

Подрядчик проводит отбор определенного количества Поставщиков, удовлетворяющих требованиям к технической готовности операторов сотовой связи, изложенным в разделе 5.1.1, и готовых выполнить поставку данных в соответствии с утвержденным техническим заданием.

Этап 3. Обработка первичных данных и формирование предварительных агрегатов.

Каждый Поставщик выполняет на своей стороне внутренние процедуры предобработки первичных данных о сетевой активности абонентов, в соответствии с требованиями изложенными в разделе 4.3.3.

На основании обработанных первичных данных Поставщик формирует предварительные агрегаты, не содержащие персональных данных, в соответствии с требованиями технического задания.

Сформированный набор данных передается Подрядчику.

Этап 4. Оценка качества поставленных данных и приемка результатов.

Подрядчик выполняет комплексную проверку полученных от Поставщика данных, включающую:

- Валидацию формата: Соответствие структуры файлов, типов данных и кодировки требованиям технического задания.
- Проверку на внутреннюю непротиворечивость: Отсутствие логических противоречий, отрицательных значений, нарушений арифметических балансов.
- Проверку полноты: Наличие данных по всем запрошенным территориальным зонам и за указанный период.

В случае выявления несоответствий или невалидных данных Подрядчик формирует и направляет Поставщику замечания с детальным описанием выявленных дефектов.

Поставщик обязан в кратчайшие технически возможные сроки устранить указанные замечания и предоставить Подрядчику новую, исправленную версию набора данных.

Приемка результатов носит циклический (итерационный) характер до момента полного устранения всех замечаний и получения от Поставщика валидного набора данных, соответствующего критериям качества.

5 Требования к построению типового бизнес-процесса

5.1 Требования к построению механизмов сбора данных

Решение геоаналитики позволяет осуществлять сбор, обработку и аналитику данных о населении на исследуемой территории на основе получаемой информации с сетей сотовых операторов.

Ниже приводится описание основных элементов архитектуры решения геоаналитики. Выбор данной архитектуры основан на необходимости осуществлять следующие ключевые функции:

- Обеспечивать непрерывный (24x7) сбор данных с сетей сотовых операторов
- Обеспечивать удаление / хэширование конфиденциальных абонентских данных на площадке Поставщика данных
- Обеспечивать контроль за качеством собранных данных
- Иметь масштабируемость для подключения новых операторов в состав источников данных
- Обеспечивать сбор и агрегацию обработанных данных от Операторов
- Обеспечивать контроль качества поступающих агрегированных данных
- Осуществлять аналитику и визуализацию обработанных данных

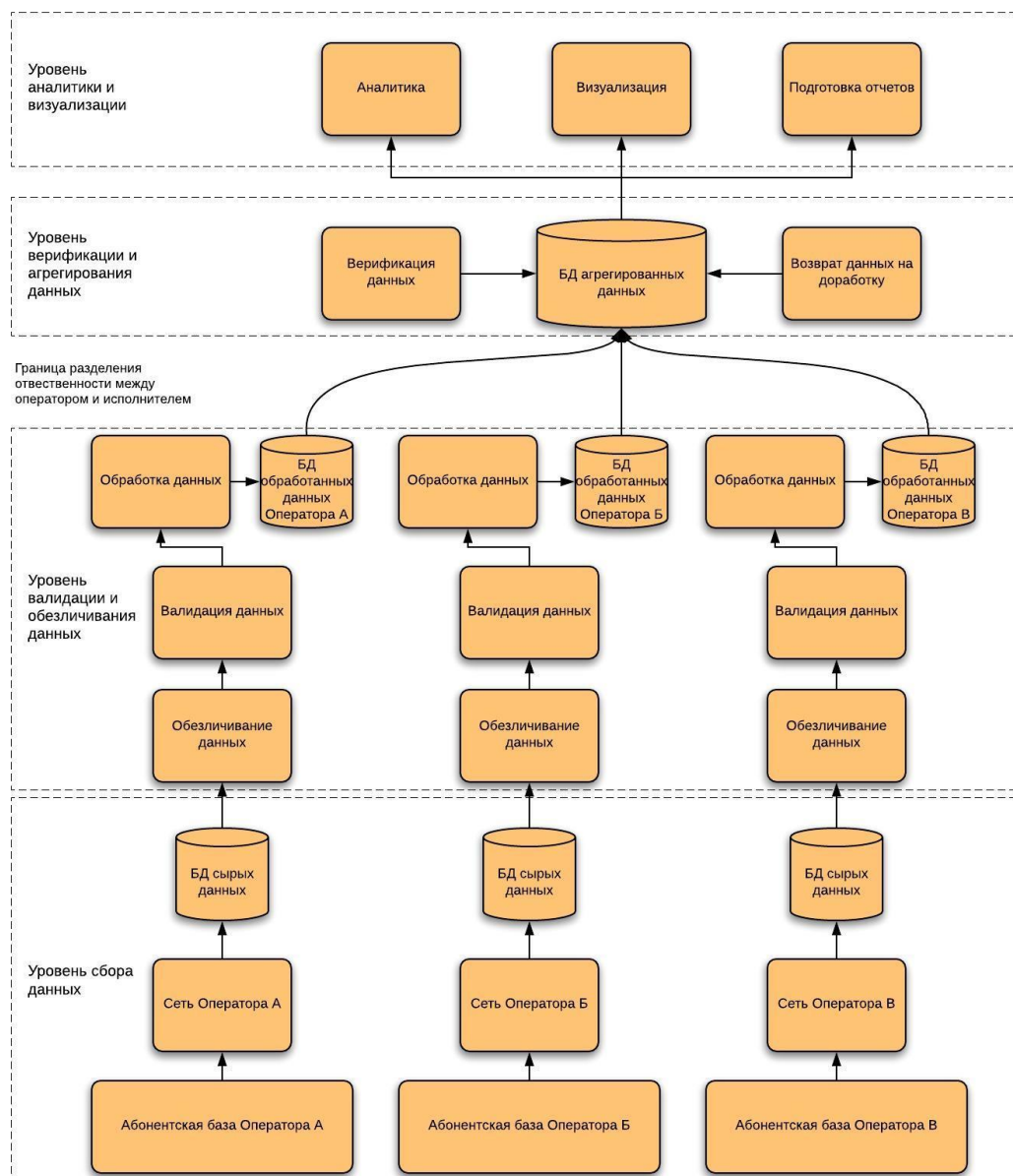


Рисунок 5.1.1. Механизм сбора данных

5.1.1 Требования к технической готовности операторов сотовой связи к сбору данных

Требование 5.1.1.1. Оператор связи должен обеспечить сбор, предобработку и хранение сырых данных.

Входные данные	Действие	Выход / Результат
–Информация об абонентской активности с привязкой к обслуживающим базовым станциям сотам	Фильтрация информации не связанных с геолокацией Контроль непрерывности сбора сырых данных	БД с сырыми геолокационными данными Информация об абонентской

–Сигнализация с различных интерфейсов сотовой сети	Сохранение предобработанных данных в БД	активности в сыром виде
--	---	-------------------------

Требование 5.1.1.2. Оператор должен проверить консистентность данных и деперсонализировать их.

Входные данные	Действие	Выход / Результат
Сырые геолокационные данные	Анализ консистентности сырых данных	БД деперсонализированных данных
Информация об абонентской активности в сыром виде	Удаление / хеширование конфиденциальной абонентской информации	Сырые данные готовы к анализу и последующей обработке
	Контроль отсутствия персональных данных	
	Передача данных в БД для последующей обработки и анализа	

Требование 5.1.1.3. Оператор должен обработать данные по алгоритмам, переданным агрегатором данных, а также провести расчет статистических метрик.

Входные данные	Действие	Выход / Результат
БД сырых деперсонализированных данных	Применение алгоритмов для уточнения геопозиции абонентов	БД агрегированных данных Оператора
Информация о конфигурации сотовой сети Оператора	Применение алгоритмов фильтрации абонентских терминалов	
Карта покрытия базовых станций Оператора	(минимизация эффекта “двойных симкарт”)	
Дополнительная статистика с сотовой сети	Выявление корреляций	
Разбиение территории	Исторический анализ	
	Выявление мест	

проекта на сегменты	жительство и работы Аналитика перемещений абонентов Сегментация абонентов по полу и возрасту на основе собственного алгоритма	
---------------------	---	--

5.1.2 Требования к агрегатору данных и его опыту реализации задач, связанных с расчетом численности населения на базе больших данных

Требование 5.1.2.1. Агрегатор должен иметь успешный опыт (не менее 2 реализованных проектов) обработки данных сотовых операторов для целей статистики или анализа пространственных перемещений.

Требование 5.1.2.2. Агрегатор должен предоставить операторам связи методику формирования таблицы временных интервалов алгоритмы расчеты статистических метрик, а также методы верификации данных.

Требование 5.1.2.3. Агрегатор должен подготовить shape-файл исследуемой территории, обеспечив необходимо территориальное деление.

Требование 5.1.2.4. Агрегатор должен обеспечить верификацию данных каждого оператора, а потом провести агрегацию данных.

Входные данные	Действие	Выход / Результат
БД агрегированных данных Оператора А БД агрегированных данных Оператора Б БД агрегированных данных Оператора В	Проверка консистентности входящих данных Применение специальных алгоритмов агрегации данных от нескольких операторов с учетом их области покрытия Корреляционный анализ Исторический анализ	БД объединенных данных от Операторов

Требование 5.1.2.5. Агрегатор должен обеспечить аналитику и визуализацию полученных данных.

Входные данные	Действие	Выход / Результат
БД Объединенных данных от Операторов Статистические данные по регионам Информация о покрытии сотовых операторов Информация о размещении БС сотовых операторов Исторические геоаналитические данные по регионам	Аналитика на основе полученных данных Визуализация и анализ полученных данных Корреляционный анализ Исторический анализ Выявление трендов и отклонений Устранение эффекта “двойных симок” Учет категорий жителей, которые не пользуются телефонами	Отчеты по геоаналитике Визуализация результатов для анализа

5.2 Требования к построению компонентов обработки и анализа

5.2.1 Требования к описанию терминов и определений с целью составления типового технического задания для двух (или более) операторов сотовой связи на подготовку предварительных агрегатов данных

Требование 5.2.1.1. В целях формирования статистических данных о численности населения, проживающего на определенной территории, осуществляется оценка численности абонентов мобильной связи, относящихся к соответствующим категориям. Категория абонента по гражданству определяется в следующем порядке.

1. Абонент.

Абонентом признается физическое лицо, идентифицируемое по международному идентификатору подвижного абонента (IMSI), зарегистрированное в сети оператора связи.

2. Домашняя страна.

Домашней страной признается страна, к которой приписан абонент. Метод определения Домашней страны зависит от статуса абонента:

- Для абонентов, обслуживаемых операторами связи государства, Домашняя страна определяется на основании сведений контрактной (абонентской) базы данных.
- Для абонентов, находящихся в режиме международного роуминга на территории государства, Домашняя страна определяется по коду Mobile Country Code (МСС) сети, к которой абонент приписан.

3. Категории абонентов по гражданству.

На основании установленной Домашней страны абоненты распределяются по следующим категориям:

- **Гражданин государства:** Абонент, Домашняя страна которого соответствует государству, на территории которого проводится оценка.
- **Иностраный гражданин или лицо без гражданства:** Абонент, Домашняя страна которого не соответствует государству, на территории которого проводится оценка.

Требование 5.2.1.2. Демографические параметры абонентов (пол и возрастная группа) определяются на основе вероятностных оценок, формируемых с применением статистических и поведенческих моделей, построенных по алгоритмам машинного обучения.

Принцип метода:

- Основанием для определения служат обезличенные и агрегированные данные о сетевой активности абонентов, включая, но не ограничиваясь: параметры вызовов, объем и структуру интернет-трафика, данные о перемещении.
- Прогнозная модель обучается на репрезентативной выборке абонентов, для которых демографические атрибуты (пол и возраст) известны из достоверных источников (например, данные договорной работы).

Ключевые характеристики метода:

- Обезличенность данных: Обработка осуществляется исключительно на основе обезличенных данных; метод не предполагает идентификации конкретного физического лица.
- Вероятностный характер оценки: Результат определения является статистическим прогнозом и не обладает абсолютной точностью. Достоверность оценки варьируется в зависимости от объема и характера доступных данных об активности абонента.

Требование 5.2.1.3. Локация и место ночевки абонента.

Под локацией абонента понимается наименьшая географическая территория, к которой может быть привязано местоположение абонента,

например, зона покрытия базовой станции. Определение границ локации осуществляется каждым оператором связи индивидуально с учетом топологии и структуры сети, включая конфигурацию и зоны покрытия базовых станций.

Местом ночевки на отчетную дату признается локация, в которой абонент провел наибольшее количество времени в период с 23:00 предыдущих календарных суток до 06:00 текущих календарных суток. Установленная локация подлежит привязке к единице территориального деления. В случае если зона покрытия локации пересекает несколько смежных единиц территориального деления, местом ночевки могут признаваться все соответствующие единицы.

В качестве особых случаев учитывается следующее: если локация находится за пределами территории государства, для которого выполняется оценка, фиксируется нахождение абонента за границей. Если данные о местоположении абонента в указанный период отсутствуют ввиду выключенного терминала или отсутствия сетевой активности, место ночевки считается неустановленным.

Требование 5.2.1.4. Определение категорий постоянного и временного населения.

К категории лиц, постоянно проживающих в государстве на отчетную дату, относятся:

- граждане государства, у которых место ночевки находилось на его территории не менее 16 дней за месяц, предшествующий отчетной дате;
- иностранные граждане, у которых место ночевки находилось на его территории не менее 16 дней за месяц, предшествующий отчетной дате.

К категории лиц, временно находящихся на территории государства, но постоянно проживающих за рубежом, относятся:

- граждане государства, которые находились на его территории в течение предыдущего месяца, но не удовлетворяют критериям постоянного проживания (то есть пробыли в стране менее 16 дней за последний месяц);
- иностранные граждане, которые находились на территории государства в течение предыдущего месяца, но не удовлетворяют критериям постоянного проживания (то есть пробыли в стране менее 16 дней за последний месяц).

Требование 5.2.1.5. Определение места проживания для постоянного и временного населения.

Место проживания для постоянного и временного населения определяется, как наиболее частое место ночевки абонента, установленное по данным за соответствующий календарный месяц.

5.2.2 Требования к составу технического задания для сотовых операторов

Требование 5.2.2.1. Техническое задание является единым для всех сотовых операторов, обеспечивая сопоставимость, однородность и возможность последующей агрегации предоставляемых данных.

Требование 5.2.2.2. Передаваемая сотовыми операторами информация должна быть строго обезличена и агрегирована. Запрещается передача любых персональных данных, позволяющих идентифицировать отдельного абонента.

Требование 5.2.2.3. Техническое задание должно содержать следующие обязательные разделы:

1. Методика определения места проживания абонента на учетную дату.

Данный раздел должен содержать детальное описание алгоритма, используемого оператором для установления локации проживания абонента. Методика должна включать:

- Определение учетной даты: Четкие временные рамки (дата и время, по состоянию на которые формируются данные).
- Критерии отнесения к месту проживания: Формализованные правила, основанные на анализе сетевых событий. Требования к правилам определения места проживания абонентов изложены в разделе 5.2.1.

2. Справочники территориальных зон.

Раздел должен включать приложения или ссылки на актуальные версии справочников, в разрезе которых требуется агрегация данных. По каждому справочнику необходимо указать:

- Наименование и версию справочника.
- Иерархию зон: Уровни детализации (например: субъект РФ -> муниципальный район -> городское/сельское поселение).
- Уникальные идентификаторы и наименования для каждой зоны каждого уровня.
- Географическое описание (например, координаты границ).

3. Формат и способ передачи данных.

Раздел регламентирует требования к структуре и каналам передачи итоговых файлов.

- Способ передачи: Указание на используемые защищенные каналы связи (например, SFTP-сервер, защищенная файлообменная платформа).
- Тип файла: Предпочтительный формат (например, CSV, XML, XLSX).
- Кодировка символов: Обязательная к применению кодировка (например, UTF-8).
- Состав и описание полей агрегированной таблицы: Детальное описание для каждого поля:
 - Наименование поля.
 - Формат данных (целое число, дата, текст и т.д.).
 - Описание данных. Допустимые значения или диапазоны.
- Структура данных: Указание на необходимость предоставления данных в разрезе каждого уровня территориальных зон из п. 2.
- Правила именования файлов: Шаблон имени файла, включающий идентификатор оператора, отчетный период и дату формирования (например, «OperatorXYZ_ZoneData_YYYY-MM-DD.csv»).

4. Регламент и сроки предоставления данных.

- Периодичность предоставления: Четкое указание (ежедневно, ежемесячно, ежеквартально).
- Крайняя дата и время предоставления данных за каждый отчетный период (например, не позднее 10:00 UTC+3 третьего рабочего дня месяца, следующего за отчетным).
- Процедура внесения изменений и предоставления уточненных данных.

5. Критерии оценки качества предоставленных данных.

Раздел определяет параметры для контроля целостности и достоверности полученной информации в соответствии с требованиями, изложенными в разделе 5.4.

5.3 Требования к компоновке производственных процессов (Никита)

5.3.1 Требования к описанию типовых форматов данных, получаемых в результате предварительной агрегации данных каждым из двух (или более) операторов сотовой связи

Требование 5.3.1.1. Данные от каждого оператора связи должны предоставляться в едином унифицированном формате, согласованном с Заказчиком, независимо от внутренней архитектуры сети оператора.

Требование 5.3.1.2. Формат файла предварительного агрегата – csv, кроме того:

- Кодировка – utf8
- Разделитель полей – “;”
- Разделитель строки - CHR(10)
- Первая строчка содержит названия полей.

Файл по каждому виду предварительного агрегата, а также каждая его версия предоставляется в заархивированном (gz) виде.

Требование 5.3.1.3. Структура файла с предварительными агрегатами должна включать следующие поля:

Номер	Наименование	Описание	Комментарии
1.	month	Месяц, за который производилось измерение	YYYY.MM
2.	zid	Номер территории соответствующего территориального деления, для которого производилось измерение	
3.	age	Возрастная категория, для которой производилось измерение	Категории возрастов: U (Undefined), 1,..., N
4.	sex	Пол	M/F/U (Undefined)
5.	cnt_home	Численность проживающего населения	Количество человек, для которых домашняя локация в рассматриваемом месяце находится на территории zid

Номер	Наименование	Описание	Комментарии
6.	cnt	Численность временного населения	Количество человек, для которых является ночным, но при этом не является домашним, районом в рассматриваемом месяце

Требование 5.3.1.4. Каждый файл должен сопровождаться с контрольной суммой, вычисленной по файлу с предварительным агрегатом.

5.3.2 Требования к описанию принципов и правил предварительной агрегации исходных данных каждого из двух (или более) операторов сотовой связи

Требование 5.3.2.1. Предварительная агрегация должна выполняться на стороне оператора связи до передачи данных Подрядчику для обеспечения конфиденциальности и снижения объема передаваемой информации.

Требование 5.3.2.2. Предварительная агрегация должна обеспечивать невозможность идентификации отдельных абонентов и восстановления их персональных данных.

Требование 5.3.2.3. Данные должны формироваться операторами на основании обработки технологических событий на базовых станциях оператора в соответствии с предложенной методикой. Формат выходных данных также должен соответствовать согласованному сторонами ТЗ.

Требование 5.3.2.4. Оператор должен отфильтровать предоставляемые данные от планшетов, модемов и устройств с несколькими SIM-картами.

Требование 5.3.2.5. Для всех подготавливаемых оператором данных должна соблюдаться перекрестная непротиворечивость данных – количественные показатели, определенные в одних таблицах, должны соответствовать количественным показателем других таблиц.

Требование 5.3.2.6. Оператор должен использовать данных обо всех абонентов, у которых была активность на сотовой сети оператора в рассматриваемый период на рассматриваемой территории, а также данных со

всех базовых станций, которые относятся к рассматриваемой территории, чтобы обеспечить полноту предоставляемых данных.

5.4 Требования к проверке систем производства

5.4.1 Требования к типовым критериям валидации и оценки качества предварительных агрегатов

Требования к типовым методикам критериям валидации предварительных агрегатов, полученных от каждого из двух (или более) операторов сотовой связи можно выделить в 4 основных группы:

Требование 5.4.1.1. Проверка целостности данных. Так как файлы с данными могут быть большого объема, то требуется проверка того, что файл не был изменен при передаче его оператором.

Требование 5.4.1.2. Соблюдение форматов данных. Среди этой группы проверок можно выделить такие проверки как: проверка наименований файлов, проверка количества полей и их названий, проверка форматов значений полей, проверка сортировки записей.

Требование 5.4.1.3. Проверка полноты данных. В отчете должны отсутствовать данные за некорректные временные периоды и по некорректным зонам территориального деления и присутствовать данные за требуемые временные периоды и по требуемым территориям.

Требование 5.4.1.4. Проверка качества данных. Набор проверок качества данных весьма широк и может включать в себя как проверки внутри одного отчета, так и перекрестные проверки между отчетами, а также сравнение отчетов за разные периоды. Кроме того, данные могут сравниваться с другими источниками, а также с данными другого оператора.

5.4.2 Требования к типовым способам проверки корректности данных сотовых операторов

Далее будут рассмотрены типовые способы проверки корректности данных сотовых операторов на предмет их качества.

Требование 5.4.2.1. Распределение показателей по зонам разбиения. Для примера рассмотрим отчет “Изменение численности населения”. На рисунке 5.4.2.1.1 представлена плотность проживающего населения для административных районов.

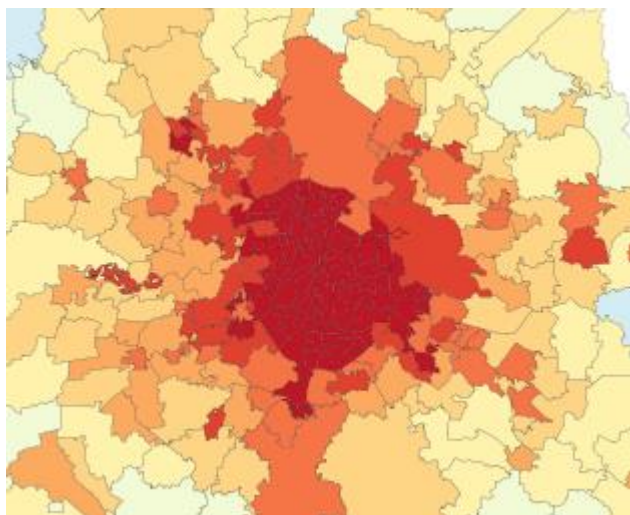


Рисунок 5.4.2.1.1. Плотность проживающего населения

Визуальная проверка может помочь выявить точки притяжения показателя, в данном случае, места проживания населения. В случае ячеек 500 на 500 метров можно визуально находить такие изъяны в данных, как наличие населения, проживающего на территории парков, лесов, водоемов и т.п. Геоаналитические данные о численности населения могут также отражать актуальную информацию о заселении новостроек или помогать выявлять места ночевки населения там, где, казалось бы, их не должно быть. Подобные выводы необходимо подкреплять либо данными из внешних источников, либо данными второго оператора, поскольку всплески или падения данных оператора могут быть связаны с проблемами на оборудовании или ошибками в расчетах.

Требование 5.4.2.2. Сравнение данных с внешними источниками. Статистический отчет оператора может быть соотнесен с другими источниками данных. Например, распределение численности населения в соответствии с административно-территориальным делением можно сравнить с данными Федеральной службы государственной статистики. Хотя, два сравниваемых источника оперируют различными методологиями, данные этих двух источников, должны коррелировать. Для сравнения двух выборок можно воспользоваться критерием χ^2 ¹¹. Этот критерий позволяет проверить гипотезу о том, что случайная величина подчиняется некому (теоретическому) закону распределения. В данном случае в качестве теоретического закона рассматриваются данные внешнего источника, в качестве проверяемой случайной величины – распределение данных сотового оператора, взвешенное

¹¹ Лагутин М. Б. Наглядная математическая статистика. (Том 2, стр. 174) — М.: П-центр, 2003.

относительно доли оператора в рассматриваемом регионе. В качестве нулевой гипотезы H_0 предполагается, что распределение случайной величины совпадает с теоретическим распределением. Критерий согласия имеет вид:

$$\chi^2 = \sum_{j=1}^k \frac{(n_j - E_j)^2}{E_j} \sim \chi_{k-1}^2, \quad (1)$$

где k – количество административно-территориальных районов;

n_j – количество абонентов в j -ом административно-территориальном районе, взвешенное с учётом рыночной доли оператора;

E_j – численность населения в j -ом административно-территориальном районе согласно теоретическому распределению.

Нулевая гипотеза H_0 отвергается при выбранном уровне значимости, если величина критерия превышает критическое значение распределения χ_{k-1}^2 . В этом случае возникает задача поиска аномальных административно-территориальных районов или зон разбиения территории. Для этого выполняется полный перебор сравнений групп из нескольких административных районов с помощью критерия χ^2 . Применение критерия χ^2 в этом случае аналогично описанному выше. Перебор вариантов может выполняться до тех пор, пока не будут выявлены все аномальные значения.

Требование 5.4.2.3. Сравнение данных одного отчета по разным разбиениям. На рисунках 5.4.2.3.1 и 5.4.2.3.2 представлено распределение численности населения по административно-территориальному делению и разбиению на зоны 500 на 500 метров. Как видно из рисунков, в зонах 500 на 500 метров данные распределены неравномерно. Подобную картину можно наблюдать при наличии на территории парковых зон, лесов, водоемов, озер и т.п.

Для сравнения данных двух разбиений, определим площадь пересечения ячеек и административных районов. Численность населения пересчитаем с ячеек на территорию административных районов, используя площади пересечений. Для сравнения исходного показателя для административных районов с рассчитанным показателем используется критерий χ^2 . В качестве теоретического закона распределения можно взять исходное распределение показателя по административным районам, а в качестве проверяемого закона – показатель, пересчитанный в административные районы из ячеек. Если нулевая гипотеза отвергается, согласно критерию, то это может свидетельствовать о неслучайных ошибках в данных сотового оператора.

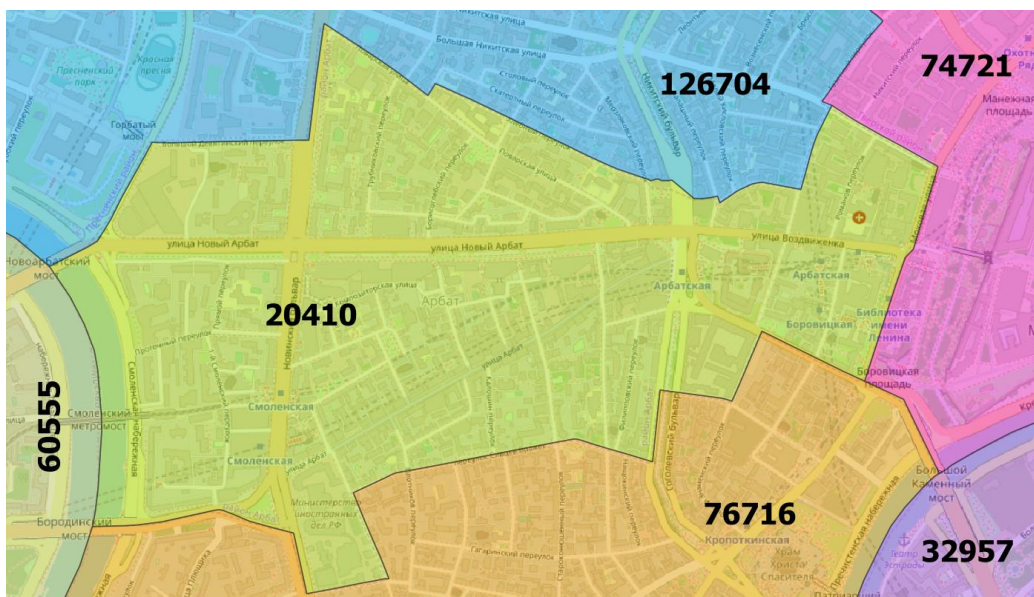


Рисунок 5.4.2.3.1. Плотность проживающего населения для административного-территориального деления территории

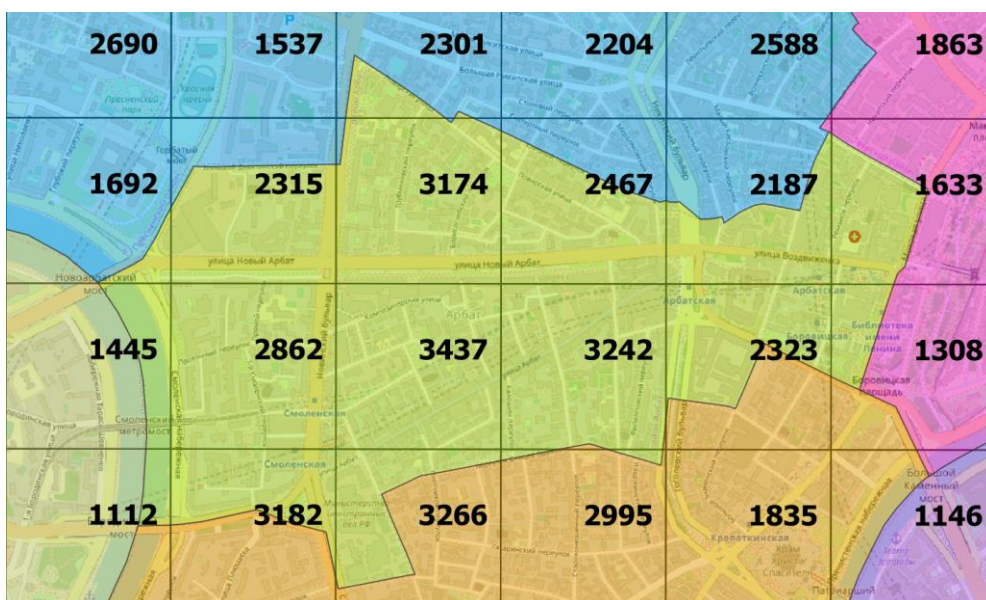


Рисунок 5.4.2.3.2. Численность проживающего населения для разбиения территории на зоны 500 на 500 метров

Требование 5.4.2.4. Сравнение данных с предыдущими периодами. В качестве примера можно рассмотреть график изменения численности проживающего населения за период с февраля 2023 года по январь 2025 года. На графике видны сезонные тренды, например, спад численности населения в летние месяцы и в канун новогодних праздников. Подобные зависимости могут быть проанализированы в разрезе отдельных ячеек или административных районов.

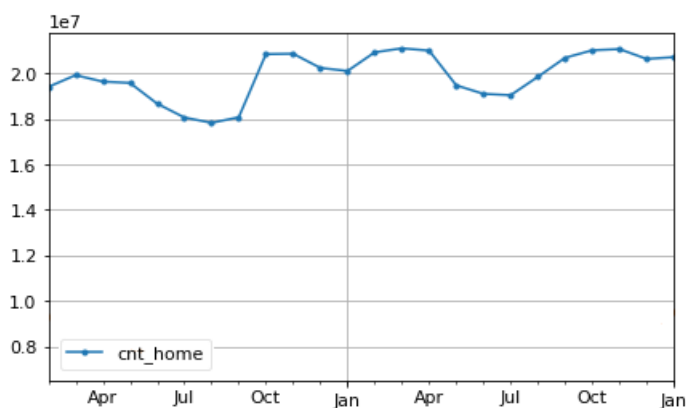


Рисунок 5.4.2.4.1. Численность проживающего и работающего населения с февраля 2023 года по январь 2025 года

Для сравнения изменений показателя во времени можно воспользоваться критерием χ^2 . В качестве теоретического закона можно взять значения показателя, усредненные, например, за предыдущие 3 месяца, а в качестве проверяемого – распределение показателя за текущий месяц. Если нулевая гипотеза отвергается, согласно критерию, то это может свидетельствовать об аномальных неслучайных тенденциях, однако это не всегда будет означать ошибку. Ряд ситуаций будет обусловлен ошибкой метода или ошибкой на оборудовании оператора, другие же ситуации могут обосновываться реальностью, поэтому такие аномалии могут верифицироваться лишь за счет внешних данных или данных второго оператора.

5.5 Требования к проверке статистического бизнес-процесса

5.5.1 Техническое задание на выгрузку данных сотовыми операторами составляется согласно пункту 5.2.2.

1. Методика определения места проживания и ночевки абонента.

Место ночевки - локация, административно-территориального деления, в которой физическое лицо в рассматриваемом месяце провело наибольшее количество времени суммарно за все дни в промежутке с 23.00 до 06.00

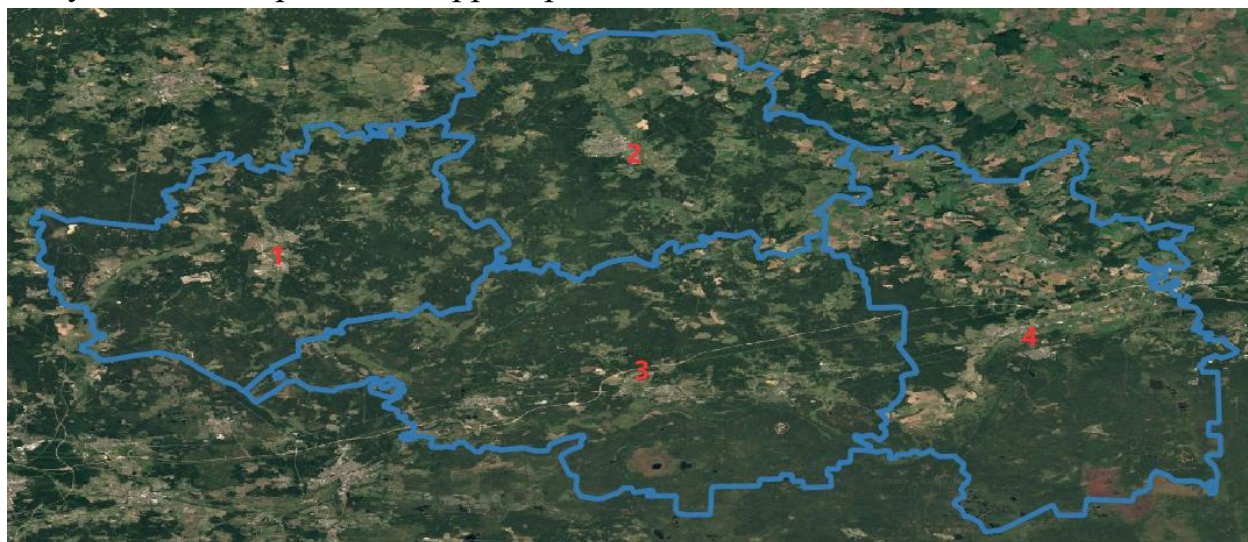
Место проживания – место ночевки в рассматриваемом месяце, если суммарное время за все дни в промежутке с 23.00 до 06.00, проведенное в месте ночёвки в рассматриваемом месяце более или равно 20% суммы всего ночного времени с 23.00 до 06.00

Данные выгружаются согласно методике, представленной в Приложении №1 к настоящему документу.

2. Справочник территориальных зон представлен shape-файлом со следующей структурой:

№	Наименование поля в файле	Описание поля	Формат
1	zid	Идентификатор территориальной зоны	Целое число
2	raion_id	Идентификатор района (идентификаторы второго уровня административно-территориального деления)	Целое число из списка
3	raion_name	Наименование района	Текстовое
4	subject_id	Идентификатор субъекта (идентификатор первого уровня административно-территориального деления)	Целое число из списка
5	subject_name	Наименование субъекта	Текстовое
6	geometry	Геометрия территории района	Полигон, Мультиполигон

Визуальное отображение территориальных зон с подложкой:



где цифры (1, 2, 3, 4) – идентификаторы соответствующих территориальных зон.

3. Формат и способ передачи данных:

- Способ передачи данных – SFTP-сервер;
- Тип передаваемого файла – CSV;
- Кодировка символов в передаваемом файле – UTF-8;

- Разделитель полей – “;”;
- Разделитель строки – CHR (10);
- Наименование отчета – изменение численности населения;
- Состав и описание полей агрегированной таблицы отчета «Изменение численности населения»:

№	Наименование поля в файле	Описание поля	Формат
1	dt	Месяц, за который производилось измерение	YYYY.MM
2	zid	Идентификационный номер территории согласно справочнику территориальных зон (в соответствии с требованиями, установленными в п 3 настоящего ТЗ).	Целое число из списка
3	gender	Пол: 1. М 2. Ж 3. Н/Д	Целое число из списка
4	age	Возраст: 1. До 30 лет 2. 30-50 лет 3. 50-65 года 4. Старше 65 лет 5. Н/Д	Целое число из списка
5	cnt_home	Количество человек, для которых zid является территорией проживания	Целое неотрицательное число
6	cnt_night	Количество человек, для которых zid является территорией самого популярного ночного времяпровождения	Целое неотрицательное число

- Наименование и очередность полей – первая строка в csv-файле.
- Правила именования файлов – csv-файлы с данными именуются согласно следующей маски «01_Location_n_xxx_YYYYMM.csv», где 01_Location – аббревиатура наименования отчета, n – номер выгрузки

(например, 1 (первая выгрузка)), xxx – аббревиатура оператора (например, Вее, Tele2, MTS), YYYYMM – месяц, за который производилось измерение (например, 202402 - февраль 2024 года).

- В дополнение к каждому файлу оператор должен предоставлять значение контрольной суммы md5. Контрольная сумма должна храниться в дополнительно предоставляемом файле <название набора статистических данных>.md5 для подтверждения объема, целостности и подлинности передаваемых данных.

1) Регламент и сроки предоставления данных.

Отчеты предоставляются ежемесячно в течении 15 рабочих дня месяца, следующего за отчетным, в период с 1 января 2024 г. по 31 декабрь 2024 г. Услуги по сбору и хранению статистических отчетов операторов сотовой связи осуществляются на программно-аппаратных средствах операторов.

2) Критерии оценки качества предоставленных данных

№ п/п	Наименование проверки	Краткое описание критериев проверки
1.1 Административно-территориальное разбиение		
Проверки соблюдения форматов данных		
1.1.1	Проверка имени файла с набором данных	Наименование файла с набором данных, содержащих информацию за 1 календарный месяц, должно соответствовать следующему наименованию: 01_Location_xxx_n_YYYYMM.csv(gz),
1.1.2	Проверка количества и полей и их названий в файле с набором данных	Количество и названия полей в наборе данных должны быть следующими: 1) dt 2) zid 3) gender 4) age 5) cnt_home 6) cnt_night
1.1.3	Проверка форматов значений в каждом поле файла	Формат значений для каждого поля в файле с набором данных должен строго соответствовать следующим форматам: 1. Для поля «dt» - YYYY.MM

№ п/п	Наименование проверки	Краткое описание критериев проверки
		2. Для поля «zid» - целое число из списка допустимых значений 3. Для поля «gender» - целое число из списка допустимых значений 4. Для поля «age» - целое число из списка допустимых значений 5. Для поля «cnt_home» - целое неотрицательное число; 6. Для поля «cnt_night» - целое неотрицательное число;
1.1.4	Проверка сортировки записей (строк)	Сортировка строк должна выполняться по следующим полям в порядке их перечисления: 1) dt 2) zid 3) gender 4) age
Проверки целостности данных		
1.1.5	Проверка контрольной суммы	Контрольная сумма (хеш-сумма) csv-файла с набором данных должна соответствовать контрольной сумме, указанной в контрольном файле с наименованием: 01_Location_xxx_n_YYYYMM.csv(gz).md5
Проверки полноты данных		
1.1.6	Проверка наличия данных за все временные периоды и по всем зонам разбиения	Данные, содержащиеся в csv-файле, должны быть выгружены за соответствующий временной период (календарный месяц), и по всем зонам территориального разбиения
1.1.7	Проверка отсутствия данных за некорректные временные периоды и/или по некорректным зонам разбиения	В наборе данных должны отсутствовать данные за некорректные временные периоды и по некорректным зонам территориального разбиения, а именно, не принадлежащим установленному множеству зон территориального разбиения.
Проверки качества данных		

№ п/п	Наименование проверки	Краткое описание критериев проверки
1.1.8	Проверка общей численности абонентов, проживающих на заданной территории	Общая численность абонентов, проживающих на заданной территории (для поля: cnt_home, сумма за месяц по всем зонам территориального разбиения), не должна существенно изменяться с течением времени. Этот показатель проверяется начиная с набора данных, предоставленного за второй и последующие временные диапазоны (месяцы).
1.1.9	Проверка общей численности абонентов, ночующих на заданной территории	Общая численность абонентов, ночующих на заданной территории (для поля: cnt_night, сумма за месяц по всем зонам территориального разбиения), не должна существенно изменяться с течением времени. Этот показатель проверяется начиная с набора данных, предоставленного за второй и последующие временные диапазоны (месяцы).
1.1.10	Проверка изменения численности проживающих абонентов по отдельным зонам территориального разбиения	Численность абонентов, проживающих на заданной территории для каждой зоны территориального разбиения, не должна существенно меняться с течением времени. Этот показатель проверяется начиная с набора данных, предоставленного за второй и последующие временные диапазоны (месяцы).
1.1.11	Проверка изменения численности ночующих абонентов по отдельным зонам территориального разбиения	Численность абонентов, ночующих на заданной территории для каждой зоны территориального разбиения, не должна существенно меняться с течением времени. Этот показатель проверяется начиная с набора данных, предоставленного за второй и последующие временные диапазоны (месяцы).
1.1.12	Перекрестная проверка между данными операторов изменения общей численности	Относительные изменения общей численности проживающих абонентов с течением времени не должны существенно отличаться по данным двух операторов. Этот показатель проверяется начиная с набора данных, предоставленного за

№ п/п	Наименование проверки	Краткое описание критериев проверки
	проживающих абонентов	второй и последующие временные диапазоны (месяцы).
1.1.13	Перекрестная проверка между данными операторов изменения общей численности ночующих абонентов	Относительные изменения общей численности ночующих абонентов с течением времени не должны существенно отличаться по данным двух операторов. Этот показатель проверяется начиная с набора данных, предоставленного за второй и последующие временные диапазоны (месяцы).
1.1.14	Перекрестная проверка между данными операторов изменения численности проживающих абонентов по отдельным зонам территориального разбиения	Относительные изменения численности проживающих абонентов для каждой зоны территориального разбиения с течением времени не должны существенно отличаться по данным двух операторов. Этот показатель проверяется начиная с набора данных, предоставленного за второй и последующие временные диапазоны (месяцы).
1.1.15	Перекрестная проверка между данными операторов изменения численности ночующих абонентов по отдельным зонам территориального разбиения	Относительные изменения численности ночующих абонентов для каждой зоны территориального разбиения с течением времени не должны существенно отличаться по данным двух операторов. Этот показатель проверяется начиная с набора данных, предоставленного за второй и последующие временные диапазоны (месяцы).
1.1.16	Распределения показателей по зонам территориального разбиения	Визуальные проверки (карты, графики).
1.1.17	Сравнение данных операторов с внешними данными (данные Федеральной	Набор данных оператора должен быть непротиворечив данным с внешних источников.

№ п/п	Наименование проверки	Краткое описание критериев проверки
	службы государственной статистики, данные о площади жилой застройки)	

5.5.2 Сбор, обработка и верификация статистических отчетов, предоставляемых операторами, в целях формирования на их основе агрегатов.

За Период с 1 января 2024 г. по 31 декабрь 2024 г. Операторы сформировали и предоставили Исполнителю статистические отчеты, приведенные в Таблице:

№	Наименование статистического отчета	Краткое описание статистического отчета	Целевое назначение	Кол-во отчетов
1	Изменение численности населения	Данные о распределение населения за заданный период времени на исследуемой территории в разрезе пола и возраста категорий населения	Оценка распределения проживающего, ночующего населения. Оценка половозрастного распределения проживающего, ночующего населения.	12

Рассмотрим сформированные отчеты операторов за один месяц (февраль 2024 года). Выгрузка из отчета первого оператора в таблице ниже:

dt	zid	gender	age	cnt_home	cnt_night
2024.02	1	1	1	950	755
2024.02	1	2	1	766	1064
2024.02	1	3	1	0	0
2024.02	1	1	2	1145	1124
2024.02	1	2	2	990	734
2024.02	1	3	2	0	0
2024.02	1	1	3	625	591

Проект: «Развитие статистики Содружества Независимых Государств»

dt	zid	gender	age	cnt_home	cnt_night
2024.02	1	2	3	967	713
2024.02	1	3	3	0	0
2024.02	1	1	4	382	467
2024.02	1	2	4	562	612
2024.02	1	3	4	0	0
2024.02	1	1	5	0	1
2024.02	1	2	5	0	1
2024.02	1	3	5	12	20
2024.02	2	1	1	1092	1082
2024.02	2	2	1	820	1433
2024.02	2	3	1	0	0
2024.02	2	1	2	1343	1057
2024.02	2	2	2	1174	1586
2024.02	2	3	2	0	0
2024.02	2	1	3	904	878
2024.02	2	2	3	904	655
2024.02	2	3	3	0	0
2024.02	2	1	4	330	410
2024.02	2	2	4	989	1329
2024.02	2	3	4	0	0
2024.02	2	1	5	0	1
2024.02	2	2	5	0	0
2024.02	2	3	5	4	5
2024.02	3	1	1	1244	1576

Проект: «Развитие статистики Содружества Независимых Государств»

dt	zid	gender	age	cnt_home	cnt_night
2024.02	3	2	1	1582	1181
2024.02	3	3	1	0	0
2024.02	3	1	2	1296	1950
2024.02	3	2	2	1520	2000
2024.02	3	3	2	0	0
2024.02	3	1	3	1050	799
2024.02	3	2	3	1192	1335
2024.02	3	3	3	0	0
2024.02	3	1	4	505	431
2024.02	3	2	4	942	1535
2024.02	3	3	4	0	0
2024.02	3	1	5	0	0
2024.02	3	2	5	0	1
2024.02	3	3	5	15	17
2024.02	4	1	1	1342	919
2024.02	4	2	1	831	1551
2024.02	4	3	1	0	0
2024.02	4	1	2	843	1410
2024.02	4	2	2	1018	1024
2024.02	4	3	2	0	0
2024.02	4	1	3	530	742
2024.02	4	2	3	1274	1033
2024.02	4	3	3	0	0
2024.02	4	1	4	411	482

dt	zid	gender	age	cnt_home	cnt_night
2024.02	4	2	4	669	735
2024.02	4	3	4	0	0
2024.02	4	1	5	1	3
2024.02	4	2	5	2	2
2024.02	4	3	5	8	11

Выгрузка из отчета второго оператора в таблице ниже:

dt	zid	gender	age	cnt_home	cnt_night
2024.02	1	1	1	2459	2654
2024.02	1	2	1	2512	2214
2024.02	1	3	1	0	0
2024.02	1	1	2	2049	2070
2024.02	1	2	2	2227	2483
2024.02	1	3	2	0	0
2024.02	1	1	3	1412	1446
2024.02	1	2	3	1662	1916
2024.02	1	3	3	0	0
2024.02	1	1	4	829	744
2024.02	1	2	4	2102	2052
2024.02	1	3	4	0	0
2024.02	1	1	5	2	3
2024.02	1	2	5	2	2
2024.02	1	3	5	24	36
2024.02	2	1	1	3074	3084
2024.02	2	2	1	2937	2324

Проект: «Развитие статистики Содружества Независимых Государств»

dt	zid	gender	age	cnt_home	cnt_night
2024.02	2	3	1	0	0
2024.02	2	1	2	2170	2456
2024.02	2	2	2	2861	2449
2024.02	2	3	2	0	0
2024.02	2	1	3	1576	1602
2024.02	2	2	3	2244	2493
2024.02	2	3	3	0	0
2024.02	2	1	4	1119	1039
2024.02	2	2	4	2368	2028
2024.02	2	3	4	0	0
2024.02	2	1	5	2	2
2024.02	2	2	5	2	2
2024.02	2	3	5	13	22
2024.02	3	1	1	4099	3767
2024.02	3	2	1	3480	3881
2024.02	3	3	1	0	0
2024.02	3	1	2	3770	3116
2024.02	3	2	2	3651	3171
2024.02	3	3	2	0	0
2024.02	3	1	3	2114	2365
2024.02	3	2	3	2842	2699
2024.02	3	3	3	0	0
2024.02	3	1	4	1378	1452
2024.02	3	2	4	3385	2792

dt	zid	gender	age	cnt_home	cnt_night
2024.02	3	3	4	0	0
2024.02	3	1	5	1	2
2024.02	3	2	5	1	3
2024.02	3	3	5	26	48
2024.02	4	1	1	3185	3608
2024.02	4	2	1	3260	2540
2024.02	4	3	1	0	0
2024.02	4	1	2	3307	2740
2024.02	4	2	2	3174	3168
2024.02	4	3	2	0	0
2024.02	4	1	3	1928	1716
2024.02	4	2	3	1921	2162
2024.02	4	3	3	0	0
2024.02	4	1	4	964	1033
2024.02	4	2	4	2493	2427
2024.02	4	3	4	0	0
2024.02	4	1	5	6	5
2024.02	4	2	5	3	4
2024.02	4	3	5	17	26

В целях контроля надлежащего выполнения Операторами Технического задания проводится проверка качества статистических отчетов, предоставленных каждым из Операторов за каждый месяц, в соответствии с критериями оценки качества предоставленных статистических данных операторов.

По результатам проверки готовятся протоколы проверки статистических отчетов отдельно по каждому отчетному месяцу и по каждому Оператору.

По результатам проверки статистических отчетов операторов формируется отчет. Отчет за февраль 2024 года первого оператора:

№ п/п	Наименование проверки	Результаты проверки набора данных
1.1 Административно-территориальное деление		
1.1.1	Проверка имени файла с набором данных	Данные удовлетворительны
1.1.2	Проверка количества и полей и их названий в файле с набором данных	Данные удовлетворительны
1.1.3	Проверка форматов значений в каждом поле файла	Данные удовлетворительны
1.1.4	Проверка сортировки записей (строк)	Данные удовлетворительны
1.1.5	Проверка контрольной суммы	Данные удовлетворительны
1.1.6	Проверка наличия данных за все временные периоды и по всем зонам разбиения	Данные удовлетворительны
1.1.7	Проверка отсутствия данных за некорректные временные периоды и/или по некорректным зонам разбиения	Данные удовлетворительны
1.1.8	Проверка общей численности абонентов, проживающих на заданной территории	Представленные данные лежат в допустимых пределах
1.1.9	Проверка общей численности абонентов, ночующих на заданной территории	Представленные данные лежат в допустимых пределах
1.1.10	Проверка изменения численности проживающих абонентов по отдельным зонам территориального разбиения	Представленные данные лежат в допустимых пределах
1.1.11	Проверка изменения численности ночующих абонентов по отдельным зонам территориального разбиения	Представленные данные лежат в допустимых пределах
1.1.12	Перекрестная проверка между данными операторов изменения	Представленные данные лежат в допустимых пределах

№ п/п	Наименование проверки	Результаты проверки набора данных
	общей численности проживающих абонентов	
1.1.13	Перекрестная проверка между данными операторов изменения общей численности ночующих абонентов	Представленные данные лежат в допустимых пределах
1.1.14	Перекрестная проверка между данными операторов изменения численности проживающих абонентов по отдельным зонам территориального разбиения	Представленные данные лежат в допустимых пределах
1.1.15	Перекрестная проверка между данными операторов изменения численности ночующих абонентов по отдельным зонам территориального разбиения	Представленные данные лежат в допустимых пределах
1.1.16	Распределения показателей по зонам территориального разбиения	Данные удовлетворительны
1.1.17	Сравнение данных операторов с внешними данными (данные Федеральной службы государственной статистики, данные о площади жилой застройки)	Противоречий не обнаружено

Отчет за февраль 2024 года второго оператора:

№ п/п	Наименование проверки	Результаты проверки набора данных
1.1 Административно-территориальное деление		
1.1.1	Проверка имени файла с набором данных	Данные удовлетворительны
1.1.2	Проверка количества и полей и их названий в файле с набором данных	Данные удовлетворительны
1.1.3	Проверка форматов значений в каждом поле файла	Данные удовлетворительны
1.1.4	Проверка сортировки записей (строк)	Данные удовлетворительны

№ п/п	Наименование проверки	Результаты проверки набора данных
1.1.5	Проверка контрольной суммы	Данные удовлетворительны
1.1.6	Проверка наличия данных за все временные периоды и по всем зонам разбиения	Данные удовлетворительны
1.1.7	Проверка отсутствия данных за некорректные временные периоды и/или по некорректным зонам разбиения	Данные удовлетворительны
1.1.8	Проверка общей численности абонентов, проживающих на заданной территории	Представленные данные лежат в допустимых пределах
1.1.9	Проверка общей численности абонентов, ночующих на заданной территории	Представленные данные лежат в допустимых пределах
1.1.10	Проверка изменения численности проживающих абонентов по отдельным зонам территориального разбиения	Представленные данные лежат в допустимых пределах
1.1.11	Проверка изменения численности ночующих абонентов по отдельным зонам территориального разбиения	Представленные данные лежат в допустимых пределах
1.1.12	Перекрестная проверка между данными операторов изменения общей численности проживающих абонентов	Представленные данные лежат в допустимых пределах
1.1.13	Перекрестная проверка между данными операторов изменения общей численности ночующих абонентов	Представленные данные лежат в допустимых пределах
1.1.14	Перекрестная проверка между данными операторов изменения численности проживающих абонентов по отдельным зонам территориального разбиения	Представленные данные лежат в допустимых пределах

№ п/п	Наименование проверки	Результаты проверки набора данных
1.1.15	Перекрестная проверка между данными операторов изменения численности ночующих абонентов по отдельным зонам территориального разбиения	Представленные данные лежат в допустимых пределах
1.1.16	Распределения показателей по зонам территориального разбиения	Данные удовлетворительны
1.1.17	Сравнение данных операторов с внешними данными (данные Федеральной службы государственной статистики, данные о площади жилой застройки)	Противоречий не обнаружено

5.5.3. Критерии оценки качества агрегированных данных

№ п/п	Наименование проверки	Краткое описание критериев проверки
1.1 Административно-территориальное разбиение		
Проверки соблюдения форматов данных		
1.1.1	Проверка имени файла с набором данных	Наименование файла с набором данных, содержащих информацию за 1 календарный месяц, должно соответствовать следующему наименованию: 01_Location_agg_n_YYYYMM.csv(gz),
1.1.2	Проверка количества и полей и их названий в файле с набором данных	Количество и названия полей в наборе данных должны быть следующими: 1) dt 2) zid 3) gender 4) age 5) cnt_home 6) cnt_night
1.1.3	Проверка форматов значений в каждом поле файла	Формат значений для каждого поля в файле с набором данных должен строго соответствовать следующим форматам:

№ п/п	Наименование проверки	Краткое описание критериев проверки
		7. Для поля «dt» - YYYY.MM 8. Для поля «zid» - целое число из списка допустимых значений 9. Для поля «gender» - целое число из списка допустимых значений 10. Для поля «age» - целое число из списка допустимых значений 11. Для поля «cnt_home» - целое неотрицательное число; 12. Для поля «cnt_night» - целое неотрицательное число;
1.1.4	Проверка сортировки записей (строк)	Сортировка строк должна выполняться по следующим полям в порядке их перечисления: 5) dt 6) zid 7) gender 8) age
Проверки целостности данных		
1.1.5	Проверка контрольной суммы	Контрольная сумма (хеш-сумма) csv-файла с набором данных должна соответствовать контрольной сумме, указанной в контрольном файле с наименованием: 01_Location_agg_n_YYYYMM.csv(gz).md5
Проверки полноты данных		
1.1.6	Проверка наличия данных за все временные периоды и по всем зонам разбиения	Данные, содержащиеся в csv-файле, должны быть выгружены за соответствующий временной период (календарный месяц), и по всем зонам территориального разбиения
1.1.7	Проверка отсутствия данных за некорректные временные периоды и/или по некорректным зонам разбиения	В наборе данных должны отсутствовать данные за некорректные временные периоды и по некорректным зонам территориального разбиения, а именно, не принадлежащим установленному множеству зон территориального разбиения.

№ п/п	Наименование проверки	Краткое описание критериев проверки
Проверки качества данных		
1.1.8	Проверка общей численности абонентов, проживающих на заданной территории	Общая численность абонентов, проживающих на заданной территории (для поля: cnt_home, сумма за месяц по всем зонам территориального разбиения), не должна существенно изменяться с течением времени. Этот показатель проверяется начиная с набора данных, предоставленного за второй и последующие временные диапазоны (месяцы).
1.1.9	Проверка общей численности абонентов, ночующих на заданной территории	Общая численность абонентов, ночующих на заданной территории (для поля: cnt_night, сумма за месяц по всем зонам территориального разбиения), не должна существенно изменяться с течением времени. Этот показатель проверяется начиная с набора данных, предоставленного за второй и последующие временные диапазоны (месяцы).
1.1.10	Проверка изменения численности проживающих абонентов по отдельным зонам территориального разбиения	Численность абонентов, проживающих на заданной территории для каждой зоны территориального разбиения, не должна существенно меняться с течением времени. Этот показатель проверяется начиная с набора данных, предоставленного за второй и последующие временные диапазоны (месяцы).
1.1.11	Проверка изменения численности ночующих абонентов по отдельным зонам территориального разбиения	Численность абонентов, ночующих на заданной территории для каждой зоны территориального разбиения, не должна существенно меняться с течением времени. Этот показатель проверяется начиная с набора данных, предоставленного за второй и последующие временные диапазоны (месяцы).
1.1.12	Распределения показателей по зонам территориального разбиения	Визуальные проверки (карты, графики).

№ п/п	Наименование проверки	Краткое описание критериев проверки
1.1.13	Сравнение данных агрегата с внешними данными (данные Федеральной службы государственной статистики, данные о площади жилой застройки)	Набор данных должен быть непротиворечив данным с внешних источников.

5.5.4. Получение и верификация агрегированных статистических отчетов

На основе предоставленных операторами статистических отчетов за двенадцать месяцев сформирован агрегированный статистический отчет (агрегат) в соответствии с пунктом 6.3. Выгрузка из полученного отчета за февраль 2024 представлена в таблице ниже:

dt	zid	gender	age	cnt_home	cnt_night
2024.02	1	1	1	5981	5981
2024.02	1	2	1	5751	5751
2024.02	1	3	1	0	0
2024.02	1	1	2	5604	5604
2024.02	1	2	2	5644	5644
2024.02	1	3	2	0	0
2024.02	1	1	3	3573	3573
2024.02	1	2	3	4612	4612
2024.02	1	3	3	0	0
2024.02	1	1	4	2124	2124
2024.02	1	2	4	4673	4673
2024.02	1	3	4	0	0
2024.02	1	1	5	4	7

dt	zid	gender	age	cnt_home	cnt_night
2024.02	1	2	5	3	6
2024.02	1	3	5	64	98
2024.02	2	1	1	7308	7308
2024.02	2	2	1	6591	6591
2024.02	2	3	1	0	0
2024.02	2	1	2	6164	6164
2024.02	2	2	2	7079	7079
2024.02	2	3	2	0	0
2024.02	2	1	3	4351	4351
2024.02	2	2	3	5523	5523
2024.02	2	3	3	0	0
2024.02	2	1	4	2542	2542
2024.02	2	2	4	5889	5889
2024.02	2	3	4	0	0
2024.02	2	1	5	4	5
2024.02	2	2	5	3	3
2024.02	2	3	5	30	47
2024.02	3	1	1	9374	9374
2024.02	3	2	1	8881	8881
2024.02	3	3	1	0	0
2024.02	3	1	2	8887	8887
2024.02	3	2	2	9072	9072
2024.02	3	3	2	0	0
2024.02	3	1	3	5551	5551

dt	zid	gender	age	cnt_home	cnt_night
2024.02	3	2	3	7077	7077
2024.02	3	3	3	0	0
2024.02	3	1	4	3303	3303
2024.02	3	2	4	7591	7591
2024.02	3	3	4	0	0
2024.02	3	1	5	2	3
2024.02	3	2	5	2	7
2024.02	3	3	5	72	114
2024.02	4	1	1	7942	7942
2024.02	4	2	1	7178	7178
2024.02	4	3	1	0	0
2024.02	4	1	2	7281	7281
2024.02	4	2	2	7354	7354
2024.02	4	3	2	0	0
2024.02	4	1	3	4312	4312
2024.02	4	2	3	5605	5605
2024.02	4	3	3	0	0
2024.02	4	1	4	2413	2658
2024.02	4	2	4	5547	5547
2024.02	4	3	4	0	0
2024.02	4	1	5	13	14
2024.02	4	2	5	9	11
2024.02	4	3	5	44	65

В соответствии с критериями верификации оценки качества агрегированных данных проводится проверка агрегированных геоаналитических отчетов. Отчет по результатам проверки:

№ п/п	Наименование проверки	Результаты проверки набора данных
1.1 Административно-территориальное деление		
1.1.1	Проверка имени файла с набором данных	Данные удовлетворительны
1.1.2	Проверка количества и полей и их названий в файле с набором данных	Данные удовлетворительны
1.1.3	Проверка форматов значений в каждом поле файла	Данные удовлетворительны
1.1.4	Проверка сортировки записей (строк)	Данные удовлетворительны
1.1.5	Проверка контрольной суммы	Данные удовлетворительны
1.1.6	Проверка наличия данных за все временные периоды и по всем зонам разбиения	Данные удовлетворительны
1.1.7	Проверка отсутствия данных за некорректные временные периоды и/или по некорректным зонам разбиения	Данные удовлетворительны
1.1.8	Проверка общей численности абонентов, проживающих на заданной территории	Представленные данные лежат в допустимых пределах
1.1.9	Проверка общей численности абонентов, ночующих на заданной территории	Представленные данные лежат в допустимых пределах
1.1.10	Проверка изменения численности проживающих абонентов по отдельным зонам территориального разбиения	Представленные данные лежат в допустимых пределах
1.1.11	Проверка изменения численности ночующих абонентов по отдельным зонам территориального разбиения	Представленные данные лежат в допустимых пределах
1.1.12	Распределения показателей по зонам территориального разбиения	Данные удовлетворительны

№ п/п	Наименование проверки	Результаты проверки набора данных
1.1.13	Сравнение данных агрегата с внешними данными (данные Федеральной службы государственной статистики, данные о площади жилой застройки)	Противоречий не обнаружено

6 Требования к сбору данных

6.1 Требования к формированию генеральной совокупности и выборки

Требование 6.1.1. Генеральная совокупность – это полная совокупность всех единиц наблюдения (отдельных лиц), относящихся к изучаемому явлению на определенной территории в заданный момент времени.

При проведении оценки численности населения в качестве генеральной совокупности рассматривается все население, постоянно проживающее и временно находящееся на территории страны, а также граждане, постоянно проживающие в стране, но временно находящиеся за ее пределами. Конкретные критерии отнесения отдельных лиц к постоянному или временному населению приведены в разделе 4.2.

Ключевые принципы формирования генеральной совокупности:

- Полнота охвата: Включение всех единиц наблюдения, подлежащих учету.
- Однозначность: Каждая единица совокупности должна быть учтена единственный раз и только в одном месте.
- Актуальность: Соответствие состава совокупности заданному моменту или периоду времени.

Требование 6.1.2. При выполнении оценки численности и пространственного распределения населения на основе обезличенных данных операторов сотовой связи выборочная совокупность формируется не на основе прямого отбора людей, а на основе обработки и моделирования полного массива сетевых событий (активности SIM-карт) с последующей экстраполяцией на генеральную совокупность.

В выборку должны быть включены агрегированные данные по всем абонентам тех сотовых операторов, которые участвуют в выполнении оценки. Конкретные требования к количеству операторов сотовой связи, необходимых и достаточных для выполнения оценки численности населения приведены в разделе 4.3.

6.2 Требования к организации сбора

6.2.1 Требования к технологиям сбора предварительных агрегатов, полученных от двух (или более) операторов сотовой связи

Требование 6.2.1.1. Система сбора должна быть построена по централизованной FTP-архитектуре с поэтапной обработкой данных от операторов связи до загрузки в аналитическую СУБД с поддержкой распределенных вычислений.

Требование 6.2.1.2. Процесс сбора должен включать следующие обязательные этапы:

- Загрузка файлов операторами на выделенный FTP-сервер.
- Проверка контрольной суммы файлов для верификации целостности.
- Перенос файлов в аналитическую СУБД с поддержкой распределенных вычислений во временную схему.
- Запуск скриптов валидации данных.
- Формирование отчета о результатах проверки.
- Миграция успешно проверенных данных в схему с чистовыми данными.

Требование 6.2.1.3. FTP-сервер должен обеспечивать круглосуточную доступность 24/7/365 с целевым показателем uptime не менее 99,9%.

Требование 6.2.1.4. Дисковое пространство FTP-сервера должно быть рассчитано на хранение данных от всех операторов за не менее 5 последовательных итераций формирования отчетов.

Требование 6.2.1.5. FTP-сервер должен поддерживать изолированные аккаунты для каждого оператора с индивидуальными квотами и правами доступа.

Требование 6.2.1.6. Для обеспечения безопасности должны использоваться FTPS (FTP over SSL/TLS) или SFTP (SSH File Transfer Protocol) с обязательным шифрованием передаваемых данных.

Требование 6.2.1.7. Каждый файл, загружаемый оператором, должен сопровождаться файлом контрольной суммы в формате SHA-256.

Требование 6.2.1.8. Процесс проверки должен выполняться при загрузке каждого файла с ведением журнала результатов проверки.

Требование 6.2.1.9. Кластер с аналитической СУБД с поддержкой распределенных вычислений должен иметь достаточный объем хранилища для размещения данных от всех операторов за последние 12 месяцев с учетом коэффициента резервирования 1.5.

Требование 6.2.1.10. Должен вестись централизованный журнал всех операций сбора и проверки данных с возможностью аудита и восстановления истории обработки.

6.2.2 Требования к технологиям агрегации данных, полученных от двух (или более) операторов сотовой связи

Требование 6.2.2.1. Система агрегации должна быть построена по распределенной масштабируемой архитектуре, обеспечивающей обработку данных от неограниченного количества операторов связи.

Требование 6.2.2.2. Архитектура системы агрегации данных от нескольких операторов должна поддерживаться модульный принцип построения с четким разделением функций:

1. Модуль приема и валидации предварительных агрегатов.
2. Модуль восстановления предварительных агрегатов.
3. Модуль вычисления суммарной доли используемых операторов.
4. Модуль суммирования данных используемых операторов и учет их суммарной доли.
5. Модуль досчета отсутствующих категорий.
6. Модуль проверки качества итоговых агрегатов.

Требование 6.2.2.3. Результат каждого этапа агрегации хранится в аналитической СУБД с поддержкой распределенных вычислений вместе с исходными данными.

Требование 6.2.2.4. Для работы процессов агрегации должна использоваться платформа управления обработкой данных с возможностями:

- Мониторинга выполнения задач в реальном времени.
- Перезапуска неудавшихся задач.
- Ведения детализированных журналов выполнения.

Требование 6.2.2.5. Восстановление данных должно включать:

- Временную интерполяцию пропущенных значений на основе сезонных паттернов
- Пространственную интерполяцию для территорий с недостаточным покрытием
- Коррекцию аномалий на основе данных других операторов

Требование 6.2.2.6. Вычисление суммарной доли используемых операторов должно учитывать демографические сегменты (пол, возраст) и территориальные сегменты с использованием данных переписи населения, данных о площадях жилых помещений и других альтернативных источников данных.

Требование 6.2.2.7. Досчет должен учитывать категории населения, слабо представленные в данных сотовых операторов:

- Дети и подростки (низкая проникновение мобильной связи)

- Пожилые люди (низкая цифровая активность)
- Жители удаленных территорий (плохое покрытие сети)

6.3 Требования к проведению сбора данных

6.3.1 Требования к типовым методикам агрегации данных от нескольких операторов

Требование 6.3.1.1. Методика агрегации должна включать в себя алгоритм поиска аномалий по полу и возрасту и их исправления с помощью данных этого же оператора за прошлые периоды:

Шаг 1. Подготовка исторического эталона

- Сформировать базу исторических данных за последние 12-24 месяца с ежемесячной детализацией
- Рассчитать средние значения и стандартные отклонения для каждой демографической группы (пол/возраст)
- Выявить сезонные паттерны и тренды для каждого демографического сегмента

Шаг 2. Анализ текущего распределения

- Сгруппировать пользователей текущего периода по полу и возрастным группам
- Рассчитать доли каждой группы в общей численности пользователей оператора
- Нормализовать данные для исключения влияния общего роста/снижения абонентской базы

Шаг 3. Выявление статистических аномалий

- Для каждой демографической группы рассчитать Z -score как отношение отклонения от исторического среднего к стандартному отклонению
- Определить аномалии: значения с Z -score > 2.5 считаются значительными отклонениями
- Проверить устойчивость аномалии по временным периодам (неделя/месяц)

Шаг 4. Коррекция аномальных значений

- Для аномальных групп рассчитать скорректированное значение на основе взвешенного скользящего среднего
- Учесть сезонный коэффициент для соответствующего времени года
- Применить сглаживание с учетом тренда последних 3-6 месяцев

Шаг 5. Валидация результатов

- Проверить, что после коррекции распределение соответствует историческим паттернам

Требование 6.3.1.2. Методика агрегации должна включать в себя алгоритм поиска аномалий по полу и возрасту и их исправления с помощью данных второго оператора за этот же период:

Шаг 1. Нормализация данных для сравнения

- Привести данные обоих операторов к единой возрастной группировке и периоду анализа
- Рассчитать относительные доли демографических групп для каждого оператора в отдельности
- Нормализовать данные по общей численности пользователей

Шаг 2. Сравнительный анализ распределений

- Построить матрицу расхождений между долями демографических групп у разных операторов
- Рассчитать статистическую значимость расхождений с использованием критерия χ^2
- Выявить группы с аномально высокими расхождениями (>15% относительного отклонения)

Шаг 3. Определение эталонного распределения

- Проанализировать репрезентативность каждого оператора в разных демографических группах
- Выбрать оператора с более стабильным историческим распределением как эталон
- Для смешанного эталона рассчитать средневзвешенное распределение с учетом качества данных

Шаг 4. Коррекция аномальных распределений

- Для оператора с аномалиями рассчитать корректирующие коэффициенты для каждой проблемной группы
- Применить поэтапную коррекцию, начиная с групп с наибольшими отклонениями
- Сохранить общую численность пользователей оператора после коррекции

Шаг 5. Верификация корректировки

- Провести повторное сравнение скорректированного распределения с эталонным

- Убедиться, что расхождения не превышают допустимую погрешность (5-7%)
- Проанализировать остаточные отклонения и при необходимости выполнить дополнительную тонкую настройку

Требование 6.3.1.3. Методика агрегации должна включать в себя алгоритм поиска аномалий в разрезе территорий и их исправления с помощью данных этого же оператора за прошлые периоды:

Шаг 1. Построение территориальных профилей

- Сгруппировать данные по административно-территориальным единицам (области, районы, города)
- Рассчитать исторические показатели плотности пользователей для каждой территории
- Выявить сезонные колебания и долгосрочные тренды по территориям

Шаг 2. Анализ текущих территориальных показателей

- Рассчитать текущие значения плотности пользователей по территориям
- Сравнить с историческими данными за аналогичные периоды
- Выявить территории с аномальными отклонениями (>2 стандартных отклонений)

Шаг 3. Классификация территориальных аномалий

- Разделить аномалии на временные (связанные с событиями) и системные (устойчивые изменения)
- Проанализировать возможные причины аномалий: изменения в сети, миграции, экономические факторы
- Оценить достоверность аномалии на основе данных смежных территорий

Шаг 4. Коррекция территориальных данных

- Для временных аномалий применить интерполяцию на основе данных соседних периодов
- Для системных аномалий скорректировать данные с учетом выявленного тренда

Шаг 5. Валидация территориальной коррекции

- Проверить пространственную согласованность скорректированных данных
- Убедиться в отсутствии резких границ между смежными территориями

- Проанализировать влияние коррекции на общие демографические показатели

Требование 6.3.1.4. Методика агрегации должна включать в себя алгоритм поиска аномалий в разрезе территорий и их исправления с помощью данных второго оператора за этот же период:

Шаг 1. Сопоставительный территориальный анализ

- Построить карты покрытия и плотности пользователей для каждого оператора
- Выявить территории со значительными расхождениями в показателях между операторами
- Рассчитать индекс территориальной согласованности для каждой административной единицы

Шаг 2. Анализ причин расхождений

- Идентифицировать технические факторы (разное покрытие сети, качество сигнала)
- Выявить коммерческие факторы (разная популярность оператора в регионе)
- Учесть демографические особенности территорий, влияющие на проникновение связи

Шаг 3. Построение эталонной территориальной модели

- Выбрать оператора с более полным территориальным покрытием как основной эталон
- Для территорий с равным покрытием рассчитать усредненный эталон
- Учесть региональные особенности при построении эталона

Шаг 4. Территориальная коррекция данных

- Для каждой территории с аномалией рассчитать корректирующий коэффициент
- Применить коррекцию с учетом относительного веса операторов на территории
- Использовать итеративный подход для территорий со сложной структурой покрытия

Шаг 5. Комплексная валидация

- Проверить согласованность скорректированных данных с внешними источниками
- Проанализировать изменение общих показателей после коррекции

- Оценить стабильность результатов при использовании разных методов коррекции

Требование 6.3.1.5. Методика агрегации должна включать в себя алгоритм расчета суммарной доли операторов по полу, возрасту и территории:

Шаг 1. Подготовка унифицированных данных

- Привести данные всех операторов к единой классификации по полу, возрасту и территориям

Шаг 2. Расчет взвешенных долей

- Назначить весовые коэффициенты операторам на основе качества и репрезентативности данных
- Рассчитать средневзвешенные доли для каждой демографическо-территориальной ячейки
- Учесть корреляции между различными сегментами при расчете долей

Шаг 3. Построение многомерных распределений

- Создать кросс-таблицы по полу, возрасту и территории
- Рассчитать маргинальные распределения для каждого измерения
- Проверить внутреннюю согласованность многомерного распределения

Шаг 4. Статистическое сглаживание

- Применить методы сглаживания для ячеек с малым количеством наблюдений
- Использовать байесовские методы для улучшения оценок в разреженных данных
- Провести итеративную коррекцию для обеспечения согласованности оценок

Требование 6.3.1.6. Методика агрегации должна включать в себя алгоритм расчета на основе этой суммарной доли полной совокупности людей:

Шаг 1. Оценка охвата населения

- Рассчитать долю населения, охваченную данными операторов связи
- Построить карты охвата для визуальной оценки репрезентативности

Шаг 2. Расчет коэффициентов экстраполяции

- Для каждой демографическо-территориальной ячейки рассчитать коэффициент пересчета

- Учесть систематические смещения, связанные с особенностями пользователей мобильной связи
- Разработать модель корректировки на основе данных переписи населения и данных о площадях жилых помещений

Шаг 3. Экстраполяция на генеральную совокупность

- Применить рассчитанные коэффициенты к данным по пользователям
- Рассчитать оценку численности населения для каждой ячейки
- Просуммировать оценки для получения общих показателей по территориям

Шаг 4. Калибровка по внешним данным

- Сравнить полученные оценки с данными текущего учета населения
- Выполнить итеративную калибровку модели для минимизации расхождений
- Учесть временной лаг между данными операторов и официальной статистикой

6.3.2 Требования к типовым способам «досчета» итоговых данных для полного охвата населения исследуемой территории с учетом:

Требование 6.3.2.1. Типовые способы «досчета» должны включать в себя алгоритм досчета отсутствующих категорий пола и возраста:

Шаг 1. Идентификация недостающих категорий

- Проанализировать полноту охвата различных демографических групп
- Выявить систематически недооцененные категории (дети, пожилые, с выделением пола и возрастной группы)
- Количественно оценить степень недоучета для каждой проблемной категории

Шаг 2. Разработка модели досчета

- Использовать данные переписи населения для определения реальной демографической структуры
- Построить регрессионные модели для прогнозирования численности недостающих категорий
- Учесть региональные особенности демографической структуры

Шаг 3. Применение досчета

- Рассчитать корректирующие коэффициенты для недостающих категорий

- Применить досчет с сохранением общей согласованности распределения
- Убедиться, что досчет не создает новых аномалий в данных

Шаг 4. Валидация демографического досчета

- Проверить, что скорректированное распределение соответствует известным демографическим паттернам
- Сравнить с альтернативными источниками данных о возрастно-половой структуре

Требование 6.3.2.2. Типовые способы «досчета» должны включать в себя алгоритм досчета территориальных делений с плохим покрытием сети:

Шаг 1. Идентификация проблемных территорий

- Проанализировать карты покрытия сотовой связи всех операторов
- Выявить территории с постоянным недостаточным покрытием или низким качеством сигнала

Шаг 2. Разработка методики территориального досчета

- Для сельских и удаленных территорий использовать экстраполяцию с аналогичных территорий с покрытием
- Для городских территорий с плохим покрытием применить методы малой области (small area estimation)
- Использовать геостатистические методы для пространственной интерполяции

Шаг 3. Интеграция данных досчета

- Объединить данные с территорий с хорошим покрытием и результаты территориального досчета
- Обеспечить плавные переходы на границах территорий с разными методами оценки
- Проверить согласованность суммарных оценок по крупным территориальным единицам

Шаг 4. Комплексная валидация территориальных оценок

- Сравнить полученные оценки с данными административных реестров
- Провести выборочные проверки на конкретных территориях
- Оценить погрешность оценок для территорий с досчетом

6.4 Требования к завершению сбора данных

Требования к типовым критериям и способам оценки финальной агрегации данных

Требование 6.4.1. Формат итоговых данных должен соответствовать требованиям технического задания. Проверка формата включает три ключевых аспекта: наименование файлов, структура данных внутри файлов (название и количество полей) и типы данных (форматы значений отдельных полей в отдельных записях).

Требование 6.4.2. Полнота данных должна соответствовать требованиям технического задания. Проверка данных на полноту включает следующие критерии:

- Временной охват. Наличие данных за все требуемые временные периоды. Отсутствие данных за иные временные периоды.
- Территориальный охват. Наличие данных по всем зонам территориального деления. Отсутствие данных по отсутствующим в справочнике территориального деления территориальным единицам.
- Содержательный охват. Наличие данных по всем возрастно-половым категориям населения в соответствии с выделенными в техническом задании категориям.

Требование 6.4.3. Проверка данных на внутреннюю непротиворечивость. Данные должны быть логически согласованы как внутри себя, так и с фундаментальными демографическими законами:

Балансовые проверки: Сумма населения по всем регионам должна равняться общей численности по стране. Сумма населения по возрастным группам должна равняться общей численности.

Проверка демографической пирамиды: Численность соседних возрастных категорий должна плавно изменяться (с учетом детской смертности и миграции). Резкие, необъяснимые "провалы" или "всплески" требуют объяснения.

Проверка половозрастной структуры: Соотношение полов в разных возрастных группах должно соответствовать известным биологическим и демографическим закономерностям (например, примерно 105 мальчиков на 100 девочек при рождении, превышение числа женщин в старших возрастах).

Требование 6.4.4. Верификация по отношению к данным официальной статистики о численности населения включает следующие критерии.

Количественное сравнение: Прямое сопоставление абсолютных значений численности населения на одну и ту же дату и для одной и той же территории. Расчет и анализ абсолютных и относительных расхождений.

Качественное сравнение: Выявление систематических расхождений (например, постоянное завышение или занижение в определенных регионах или категориях населения).

Анализ причин расхождений: Различие в методологии, в определении постоянного и временного населения, в сроках учета.

Требование 6.4.5. Валидация с использованием косвенных данных и альтернативных источников.

Для проверки данных в условиях неполной официальной информации могут быть использованы не прямые индикаторы, при условии наличия соответствующих данных. Ниже приводятся примеры некоторых возможных источников данных для косвенной оценки численности населения.

1. Данные о потреблении коммунальных ресурсов:

Электроэнергия: Анализ динамики потребления электроэнергии в жилом секторе. Рост/снижение потребления может косвенно свидетельствовать о росте/снижении реального населения. Используются удельные показатели (потребление на душу населения).

Водоснабжение: Анализ объемов потребления питьевой воды. Этот показатель часто считается более стабильным и менее подверженным влиянию экономической конъюнктуры, чем электроэнергия.

2. Данные из реестров и административных источников:

Бюро технической инвентаризации (БТИ) / государственный реестр недвижимости: Данные об общей жилой площади и количестве зарегистрированных жилых помещений. Сопоставление с данными о средней обеспеченности жильем на человека позволяет получить косвенную оценку численности.

Реестры избирателей: Количество зарегистрированных избирателей может служить индикатором для взрослого населения.

3. Прочие альтернативные источники:

Снимки ночной освещенности со спутников: Интенсивность освещения коррелирует с плотностью и численностью населения.

Данные о розничном товарообороте или объеме коммунальных отходов.

Важно: Данные из альтернативных источников требуют калибровки и учета специфики. Например, на потребление электроэнергии влияет не только численность, но и энергоэффективность.

Приложение 1

1. Методика формирования таблицы временных интервалов

Временной период для формирования отчетов составляет один календарный месяц. Если это специально не оговорено в методике формирования отчета, то для получения данных за отчетный календарный месяц используются лишь те устройства, суммарный голосовой трафик которых (входящий + исходящий) за этот месяц - не менее 10 минут.

Ниже описан алгоритм подготовки следующих таблиц (алгоритм описан для данных об одном абоненте):

1. *Таблица временных интервалов*, в которой содержится информация о сотах и зонах территориального разбиения, в которых находился каждый абонент в любой момент времени. Финальная таблица состоит из следующих полей:
 - a. MSISDN – идентификатор абонента
 - b. TimeInterval – интервал времени
 - c. CellList – набор сот
 - d. FirstEvent – время обработки первого события в интервале
 - e. LastEvent – время обработки последнего события в интервале
 - f. Status – статус состояния абонента (движение/остановка)
 - g. POI_Home – индикатор нахождения абонента дома
 - h. POI_Work – индикатор нахождения абонента на рабочем месте
 - i. POI_Dacha – индикатор нахождения абонента на даче
 - j. zid_dst – зона пребывания абонента по административному разбиению территории

Часть 1. Подготовка таблицы временных интервалов для определения места жительства абонента.

Этап 1. Формирование таблицы событий абонента.

В данный момент выделяются следующие типы событий:

1. Начало звонка

2. Окончание звонка
3. Пересечение границ сот во время звонка (хэндовер)
4. Начало пакетной сессии
5. Окончание пакетной сессии
6. Пересечение границ сот во время активной пакетной сессии (хэндовер)
7. Звонок
8. Смс
9. Пересечение границы LAC
10. Включение терминала
11. Выключение терминала
12. Регистрация терминала в сети после потери покрытия

Приведенный набор событий является минимально необходимым для работы по методике. Допускается расширение набора событий в дальнейшем с целью улучшения точности алгоритма.

Для каждого события известно:

1. MSISDN - идентификатор абонента
2. EventTime – дата и время обработки события (с точностью до секунды)
3. EventType - тип события
4. LAC, Cell ID - сота, в которой было обработано событие

Все события одного абонента за рассматриваемый период времени (как правило, это один или несколько месяцев) сортируются по времени обработки события.

В начало рассматриваемого периода времени добавляется вспомогательное событие, имеющие следующие свойства:

MSISDN	EventTime	EventType	LAC, Cell ID
Идентификатор рассматриваемого абонента	Начало рассматриваемого	Включение терминала	Сота, в которой было обработано первое событие

	временного интервала Например, 2024.02.01. 00:00:00		рассматриваемого абонента в рассматриваемый временной интервал
--	---	--	--

В конец рассматриваемого периода времени добавляется вспомогательное событие, имеющие следующие свойства:

MSISDN	EventTime	EventType	LAC, Cell ID
Идентификатор рассматриваемого абонента	Конец рассматриваемого временного интервала Например, 2024.08.31. 23:59:59	Выключение терминала	Сота, в которой было обработано последнее событие рассматриваемого абонента в рассматриваемый временной интервал

Этап 2а. Формирование таблицы временных интервалов.

Основываясь на таблице, полученной на первом этапе, формируется новая таблица, которая будет содержать интервалы времени пребывания абонента в каждой соте.

Таблица временных интервалов имеет следующий набор столбцов:

1. MSISDN - идентификатор абонента
2. TimeInterval – интервал времени между моментами времени начала интервала (TimeStart) и окончания интервала (TimeEnd). При записи интервала квадратная (круглая) скобка означает, что конечный момент времени входит (не входит) в интервал.

3. CellList - список сот, входящих в локацию пребывания абонента
4. FirstEvent - время первого события в интервале
5. LastEvent - время последнего события в интервале

Обработка таблицы событий идет построчно. Одновременно рассматривается две строки, т.е. два последовательных события:

Пусть для абонента зафиксировано 2 последовательных события, которые произошли в моменты времени T1 и T2 и были обработаны в сотах C1 и C2 соответственно.

А) Если выполнено хотя бы одно из следующих условий:

- а) между событиями прошло больше 24 (?) часов
- б) событие, произошедшее в момент времени T1, имеет тип «Выключение терминала»
- в) событие, произошедшее в момент времени T2, имеет тип «Включение терминала»

Тогда местоположение абонента в интервал времени (T1, T2) считается неизвестным.

В таблицу временных интервалов добавляются две строки:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	[T1, T1]	C1	T1	T1
Идентификатор абонента	(T1, T2)	Н/Д	Н/Д	Н/Д

Б) Если не выполнено ни одно из условий из пункта А), а событие, произошедшее в момент времени T2, имеет тип “Пересечение границы LAC”, тогда считается, что абонент все время (T1, T2) находился в соте C1.

В таблицу временных интервалов добавляется строчка:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	[T1, T2)	C1	T1	T1

В) Если ни одно из условий пункта А) и Б) не выполнено, тогда считается, что в течение промежутка $(T1, (T1+T2)/2]$ абонент находился в соте C1, а в течение промежутка $((T1+T2)/2, T2)$ – в соте C2.

В таблицу временных интервалов добавляется две строки:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	[T1, (T1+T2)/2]	C1	T1	T1
Идентификатор абонента	$((T1+T2)/2, T2)$	C2	T2	T2

Таким образом, в сформированной таблице интервалов содержится информация о наборе сот, в которых находился абонент в любой момент времени.

Этап 2б. Исключение непродолжительных периодов времени, во время которых местоположение абонента было не определено.

В случае, если временной интервал имеет CellList = Н/Д и его длительность меньше 10 минут, то этот временной интервал прибавляется к предыдущему, т.е. две последовательные строчки из таблицы интервалов:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	[T12, T13]	C8	T117	T118
Идентификатор абонента	(T13, T14)	Н/Д	Н/Д	Н/Д

Заменяются одной:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	[T12, T14]	C8	T117	T118

Этап 3. Поиск непродолжительных циклов.

На данном этапе происходит уточнение полученной таблицы временных интервалов путем объединения временных интервалов, составляющих непродолжительный цикл.

Назовем циклом последовательность, которая состоит из двух или большего количества временных интервалов, причем набор сот последнего временного интервала совпадает или полностью содержится в наборе сот первого временного интервала в последовательности:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	[T15, T16)	C9	T119	T120
Идентификатор абонента	[T17, T18)	C10	T127	T128
Идентификатор абонента	[T19, T20)	C11	T129	T130
Идентификатор абонента
Идентификатор абонента	[T21, T22)	C9' (причем $C9' \subseteq C9$)	T131	T132

При этом ни один из наборов сот в последовательности не должен быть равен Н/Д, а также ни одна из сот не должна относиться к подземным станциям метрополитена.

Длительность цикла равна времени, которое прошло между событиями, произошедшими в одном и том же наборе сот, т.е. для приведенного примера длительность цикла равна $T_{131} - T_{120}$.

В случае, если длительность цикла меньше 5 минут, входящие в цикл интервалы объединяются в один интервал:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	[T15, T22)	C9, C10, C11, ...	T119	T132

После этого продолжается построчная обработка таблицы, начиная с добавленной строки.

Этап 4. Поиск последовательных интервалов времени, в которые абонент пребывал в соседних сотах.

На данном этапе таблица интервалов времени уточняется при помощи справочника соседних базовых станций. Обработка таблицы идет построчно. Одновременно рассматривается две строки, т.е. два последовательных интервала времени, для которых местоположение абонента известно (т.е. значение столбца CellList не равно Н/Д) и кроме того, ни в одном из временных интервалов он не находился в подземной части метрополитена:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	[T6, T7)	C4 = {C4а, C4б, C4в, ...} – список из одной или нескольких сот	T113	T114
Идентификатор абонента	(T7, T8)	C5	T123	T124

Если среди сот из множества {C4а, C4б, C4в, ..., C5} есть хотя бы одна, для которой все соты из данного множества (кроме нее самой) являются соседними, то рассматриваемые две строки заменяются одной:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent
Идентификатор абонента	[T6, T8)	C4a, C4б, C4в,..., C5	T113	T124

После этого продолжается построчная обработка таблицы, начиная с добавленной строки.

Этап 5. Добавление в таблицу интервалов времени поля, показывающего, находился ли абонент в движении.

Для каждого временного интервала определяется, был ли абонент неподвижен или находился в движении. В таблицу, полученную на предыдущем этапе, добавляется столбец Status, значение которого для каждой строчки определяется следующим образом:

1. Если местоположение абонента неизвестно (значение столбца CellList = Н/Д), то Status = Н/Д.
2. Если местоположение абонента известно (CellList не равно Н/Д), тогда:
 - a. Если длительность временного интервала (TimeEnd – TimeStart) не меньше 60 минут, то абонент считается неподвижным и в Status записывается Stay.
 - b. Если длительность временного интервала меньше 60 минут, то абонент считается передвигающимся и в Status записывается Move.
 - c. Если все соты из CellList относятся к подземным станциям метрополитена, то вне зависимости от длительности временного интервала в Status записывается Move.

Этап 6. Добавление в таблицу вспомогательных передвижений.

Будем считать, что абонент, который был неподвижен сначала в одном наборе сот, а потом в другом наборе сот, обязательно совершает между этими неподвижными состояниями передвижение.

Таким образом, между каждыми двумя последовательными временными интервалами, имеющими Status = Stay:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent	Status
Идентификатор абонента	[T9, T10]	{C6a,C6б,...}	T115	T116	Stay
Идентификатор абонента	(T10, T11)	{C7a,C7б,...}	T125	T126	Stay

Добавляется вспомогательная строчка, имеющая Status = Move:

MSISDN	TimeInterval	CellList	FirstEvent	LastEvent	Status
Идентификатор абонента	[T9, T10)	{C6a,C6б,...}	T115	T116	Stay
Идентификатор абонента	[T10, T10]	Н/Д	T10	T10	Move
Идентификатор абонента	(T10, T11)	{C7a,C7б,...}	T125	T126	Stay

Таким образом, таблица интервалов каждого абонента содержит информацию о наборе сот, в которых находится абонент в любой момент за рассматриваемый период времени. Кроме того, в ней есть информация, находился ли абонент в неподвижном состоянии (Status = Stay) или совершал поездку (Status = Move). В дальнейшем, под местоположением абонента в любой момент времени будет пониматься набор сот из колонки CellList того временного интервала, который содержит рассматриваемый момент времени.

2. Методика составления отчета «Изменение численности населения»

Для формирования отчета используется разбиение территории Московской агломерации на административные районы.

При формировании отчета используются таблица временных интервалов, полученная согласно пункту 1. Методика формирования таблицы временных интервалов.

Алгоритм обработки данных для построения отчета состоит из следующих этапов:

1. Определение домашней соты для каждого абонента.
2. Перевод данных из соты в разбиение.
3. Формирование отчета на основе полученных на предыдущих этапах данных.

Подробное описание этапов:

Этап 1. Определение домашней соты для каждого абонента.

На основании таблицы временных интервалов определяется время, которое абонент провел в каждой соте ночью. Для определения времени пребывания абонента в соте $C1$, учитываются все временные интервалы из таблицы, для которых в наборе сот (CellList) содержится $C1$.

Этап 1. Определение основной домашней соты для каждого абонента.

Для определения времени *ночного* пребывания абонента в соте рассматривается время с 23:00 до 6:00 за все дни месяца (с учетом праздников и выходных). Из всех сот, в которых абонент был зафиксирован в ночное время, выбирается та, в которой он провел максимальное время ночью в рассматриваемом календарном месяце. Сот, удовлетворяющих этому условию максимальной, может быть несколько (в случае, если в нескольких удовлетворяющих указанному выше условию сотах абонент провел одинаковое время), тогда из них выбирается одна сота случайным образом. Время, которое абонент провел ночью в выбранной соте, будем называть *ночным временем* абонента, а саму выбранную соту – *ночной сотой*.

Если ночное время абонента больше, чем $(0.185 * \langle \text{количество дней в отчетном периоде} \rangle * 7)$ часов, то ночная сота будет называться *основной домашней сотой* абонента. В противном случае считается, что для данного

абонента основная домашняя сота не определена. Например, для отчетов за март 2024 года основная домашняя сота будет определена для абонентов, ночное время которых превышает $0.185 * 31 * 7 = 40.145$ часов.

Стоит обратить внимание, что согласно введенной терминологии для каждого абонента определено ночное время (оно может быть равно, например, нулю) и ночная сота. При этом для абонента может не существовать основной домашней соты.

Для каждого абонента, для которого была определена ночная сота, формируется *набор ночных сот*. Для этого перебираются все временные интервалы данного абонента за рассматриваемый календарный месяц. В случае, если для рассматриваемого временного интервала в поле CellList встречается ночная сота, то все соты из данного списка CellList включаются в набор ночных сот.

Для каждого абонента, для которого была определена основная домашняя сота, формируется *набор домашних сот*. Для этого перебираются все временные интервалы данного абонента за рассматриваемый календарный месяц. В случае, если для рассматриваемого временного интервала в поле CellList встречается основная домашняя сота, то все соты из данного списка CellList включаются в набор домашних сот.

Этап 2. Перевод данных из сот в разбиение.

На предыдущем этапе для каждого абонента были определены ночная и основная домашняя сота. Каждой ночной и основной домашней соте ставятся в соответствие зона разбиения по изложенному ниже алгоритму. Таким образом, для каждого абонента будет однозначно определена зона самого популярного ночного времяпровождения (зона ночевки) и зона проживания (при их наличии).

Этап 3. Определение зоны проживания/ночевки абонента.

Каждой соте из набора домашних/ночных сот ставится в соответствие область на плоскости – область покрытия соты – в соответствии с данными

радиопланирования. Назовем домашней/ночной локацией сумму областей покрытия сот, входящих в набор домашних/ночных сот.

Зона проживания/ночевки абонента для административного разбиения территории определяется следующим образом:

1. Фиксируется ближайший предшествующий отчетный период, для которого определена зона проживания/ночевки абонента. Назовем этот период *предыдущим отчетным периодом*.
2. Если домашняя/ночная локация в текущем отчетном периоде находится на расстоянии не более 500 метров от домашней/ночной локации из предыдущего отчетного периода, то будем считать, что зона проживания/ночевки абонента в текущем отчетном периоде совпадает с зоной проживания/ночевки абонента в предыдущем отчетном периоде. Расстояние между локациями равно минимальному расстоянию между точками, каждая из которых лежит в соответствующей локации. Например, расстояние между пересекающимися локациями равно 0.
3. В противном случае, если область покрытия основной соты не имеет общих точек ни с одной зоной территориального деления, то зона проживания/ночевки абонента не определена.
4. В противном случае, основная домашняя/ночная сота приписывается к зоне разбиения, на территории которой находится область покрытия соты. Если разные части области покрытия основной домашней/ночной соты находятся в разных зонах разбиения, то сота приписывается к зоне разбиения случайным образом, причем вероятность приписывания соты к зоне разбиения должна быть пропорциональная площади пересечения области покрытия с зоной.

Приведем пример:

Пусть у нас есть разбиение на зоны, в котором каждая зона является квадратом со стороной 1 км, и основная домашняя/ночная сота 1#54. Зона покрытия соты устроена так, как на рисунке ниже.



В таблице ниже приведены вычисления вероятности приписывания абонента к каждой из зон разбиения.

Номер зоны	2	3	4	9	10	11	19	20	21	Сумма
Площадь пересечения зоны с сотой 1#54, кв. км	0	0.5	0.5	0	1	0.5	0	0	0	2.5
Вероятность нахождения абонента в зоне	0	1/5	1/5	0	2/5	1/5	0	0	0	1
Нижняя граница P1	0	0	0.2	0.4	0.4	0.8	1	1	1	
Верхняя граница P2	0	0.2	0.4	0.4	0.8	1	1	1	1	

Для выбора одной зоны с нужной вероятностью осуществляется генерация случайного числа P в диапазоне от 0 до 1 ($0 \leq P < 1$). В зависимости от значения P выбирается та зона разбиения, для которой выполнено двойное неравенство: $P1 \leq P < P2$.

После того, как определена зона проживания/ночевки абонента, следует перейти к следующему абоненту, который нас интересует и повторить алгоритм выбора зоны сначала (даже в том случае, если основная домашняя/ночная сота нового абонента совпадает с только что рассмотренным абонентом).

Этап 4. Формирование отчета

При формировании отчета рассматривается административно-территориальное разбиения. В отчетах содержится информация за отчетный месяц.

В поле 1 отчета (dt) записывается номер отчетного месяца.

В поле 2 отчета (zid) записывается рассматриваемая зона разбиения.

В поле 3 отчета (gender) записывается пол абонентов.

В поле 4 отчета (age) записывается возрастная группа абонентов.

В поле 5 отчета (cnt_home) записывается число абонентов, для которых место проживания определено и находится на территории, заданной в поле 2, пол соответствует значению поля 3, а возраст соответствует значению поля 4.

В поле 6 отчета (cnt_night) записывается число абонентов, для которых место ночевки определено и находится на территории, заданной в поле 2, пол соответствует значению поля 3, а возраст соответствует значению поля 4.