InfoNet

Большие данные в статистике населения и их применение в госуправлении

Что такое статистика и большие данные

ЧТО такое статистика населения

СТАТИСТИЧЕСКИЕ ДАННЫЕ

цифровые сведения, собранные, обработанные и проанализированные с целью выявления закономерностей и тенденций массовых явлений

Ключевые характеристики данных

- ✓ **Массовость:** собираются не по одному человеку, а по группам
- ✓ **Системность:** данные собираются регулярно и по единым правилам
- ✓ **Измеримость:** все явления выражены в цифрах (количество, проценты, коэффициенты)

СТАТИСТИКА НАСЕЛЕНИЯ

раздел статистики, который изучает численность, состав, размещение и воспроизводство населения в количественном выражении

Что изучает статистика населения

- ✓ **Численность и динамика:** сколько людей на определенной территории, сколько прибыло и убыло
- ✓ **Состав населения:**
  - Половозрастной: сколько мужчин/женщин, детей/пожилых
  - Национальный: какие народы живут
  - Социальный: городское/сельское, образование, источники дохода
- ✓ **Воспроизводство (Естественное движение):**
  - Рождаемость
  - Смертность
  - Браки и разводы
- ✓ **Миграция (Механическое движение):**
  - Переезды внутри страны и между странами

ЧТО использовать для статистики населения: цифровые источники данных

Данные мобильных операторов

Данные видеоаналитики

Данные административных регистров

Данные банков

ПРЕИМУЩЕСТВА ЦИФРОВЫХ ИСТОЧНИКОВ ДАННЫХ

ОПЕРАТИВНОСТЬ

данные можно получать в режиме близком к реальному времени - время сбора информации о фактическом населении сокращается до нескольких дней

СНИЖЕНИЕ ЗАТРАТ

меньше персонала и бумажной документации

ОБЪЕМ И ОХВАТ

широкий охват, особенно в городах

ДЕТАЛИЗАЦИЯ

возможность анализа по времени, локации (гранулярность территорий возможна от 500*500 м до административных делений)

Зачем нужны большие данные в статистике

ЗАЧЕМ использовать большие данные для статистики населения

InfoNet

Планирование бюджета и инфраструктуры

Государству нужно знать сколько людей живёт в стране, чтобы грамотно распределять **финансы**, строить **школы, дороги, больницы**, развивать **транспорт и связь**

Анализ миграции и урбанизации

Помогает понять **куда и откуда переезжают люди**, где растёт население и где сокращается. Это важно для городов, где особенно заметна урбанизация

Оценка трудовых ресурсов

Сколько людей имеет **постоянное место работы,** как далеко и часто **они перемещаются**

Политика и управление

Определение избирательных округов, субсидий, дотаций

Примеры использования больших данных в различных отраслях

InfoNet

ГРАДОСТРОИТЕЛЬСТВО

- Повышение качества исходных данных для комплексного проектирования территорий
- Планирование и развитие инфраструктуры с учетом трендов в динамике изменения численности проживающих
- Оценка подвижности населения по районам на территории Московской агломерации
- Применение данных из цифровых источников на регулярной основе для планирования, мониторинга и исполнения Генерального плана города Москвы

ТРАНСПОРТ

- Мониторинг перемещений населения по территории города: использование данных из цифровых источников в собственных транспортных моделях Департамента в целях повышения их точности
- Использование данных из цифровых источников при планировании реконструкции отдельных участков улично-дорожной сети
- Оценка потенциального спроса на систему маршрутов Быстрого Автобусного Транспорта (Метробуса)
- Оценка изменения объема спроса на междугородние автобусные маршруты

ЭКОНОМИЧЕСКОЕ РАЗВИТИЕ

- Оценка распределения рабочих мест относительно районов проживания на территории Московской агломерации
- Мониторинг численности трудозанятого населения и потоков маятниковой трудовой миграции
- Выявление на территории Москвы потенциальных торговых и офисных объектов, уклоняющихся от уплаты налога на имущество

ТУРИЗМ И КУЛЬТУРА

- Оценка и изменение численности туристов и экскурсантов из различных стран и регионов РФ и их распределение по территории Москвы
- Получение детализированной информации о посещаемости туристами центральной части Москвы в целях развития туристической инфраструктуры на исследуемой территории
- Определение уникального количества болельщиков, посетивших Москву за все время ЧМ-2018
- Оценка динамики посещаемости территорий расположения московских библиотек в целях корректировки и оптимизации графиков их работы

ТОРГОВЛЯ И УСЛУГИ

- Анализ посещаемости городских ярмарок и фестивалей
- Определение наиболее востребованных ярмарок в будние и выходные дни
- Выявление наиболее посещаемых локаций на территории города как возможных мест проведения ярмарок
- Определение уровня востребованности ярмарок среди городского/областного населения и иногородних

СМИ И РЕКЛАМА

- Оценка численности проживающего, работающего и транзитного населения в местах расположения рекламных конструкций в Москве:
  - оценка потенциального охвата аудитории в локациях расположения существующих рекламных конструкций
  - оценка потенциала территорий для установки новых конструкций

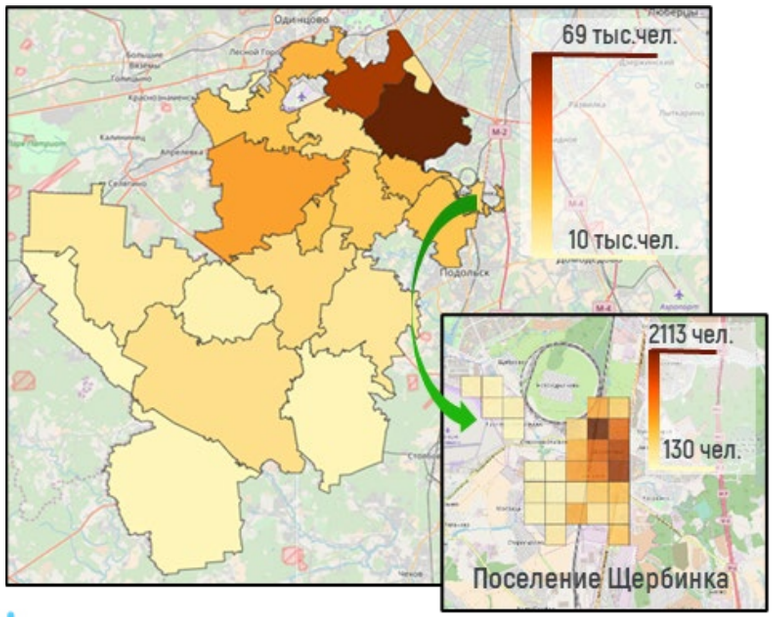
Коррекция планов комплексного проектирования территорий с учетом актуальной информации о фактически проживающем населении

InfoNet

Результат

- В план комплексного проектирования территорий Новой Москвы были внесены изменения на основе данных геоаналитики о численности фактически проживающего населения

Распределение численности фактически проживающего населения в Новой Москве



Этапы выполнения работы

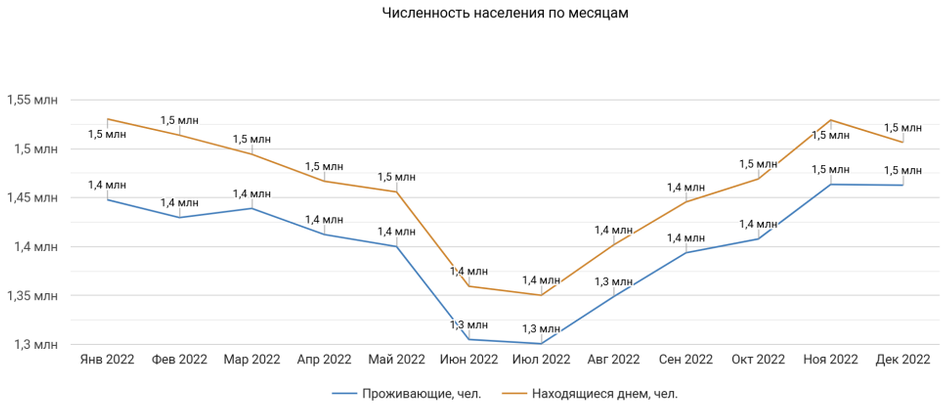
Использование геоданных позволило:

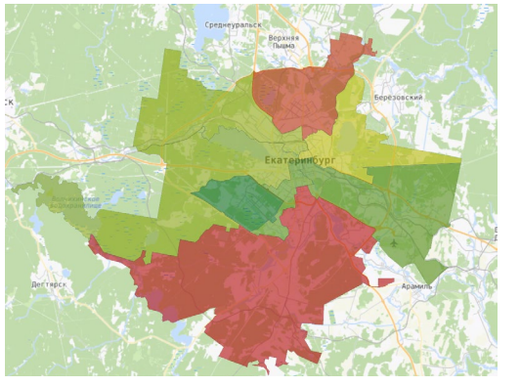
- Собрать необходимые данные за несколько недель
- Рассчитать численность фактически проживающего населения в Новой Москве с детализацией ее территории до секторов 500х500м.
- Сделать прогноз по приросту населения Новой Москвы с учетом сезонных трендов
- Скорректировать план по проектированию территорий Новой Москвы, реализация которого привела к значительному улучшению ситуации в социальной сфере и дорожно-транспортной системе Новой Москвы

Изменение численности проживающего населения в течение года

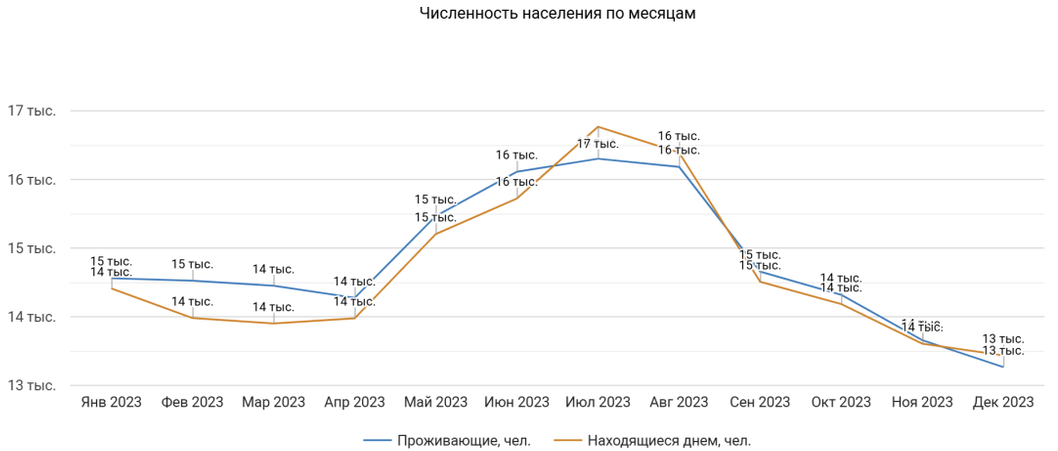
ПРИМЕРЫ ТРЕНДОВ

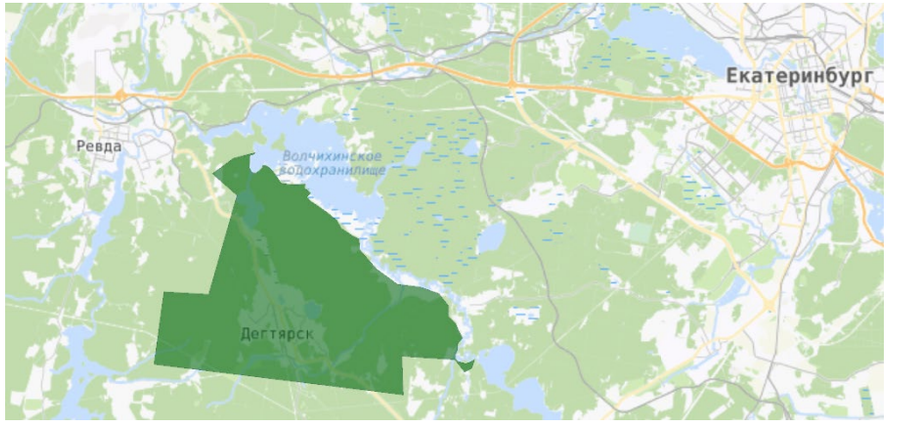
- Спад численности проживающих в летние месяцы за счет сезонных миграций (дачного сезона) – *Екатеринбург – спад численности в летние месяцы на 15%*





- Рост численности проживающих и дневного населения в летние месяцы в сельских районах - *ГО Дегтярск Свердловской области – прирост численности в летние месяцы более, чем на 20%*





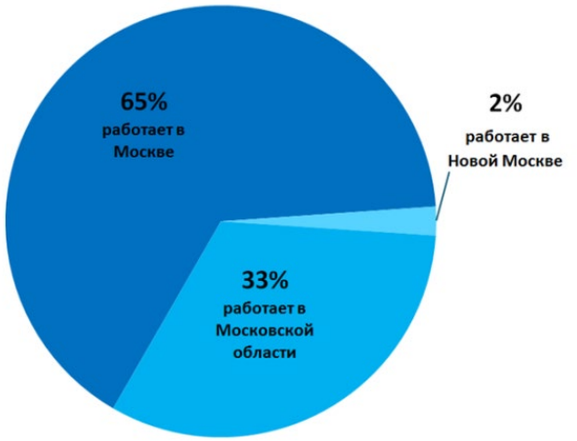
Мониторинг численности трудозанятого населения и потоков маятниковой трудовой миграции

InfoNet

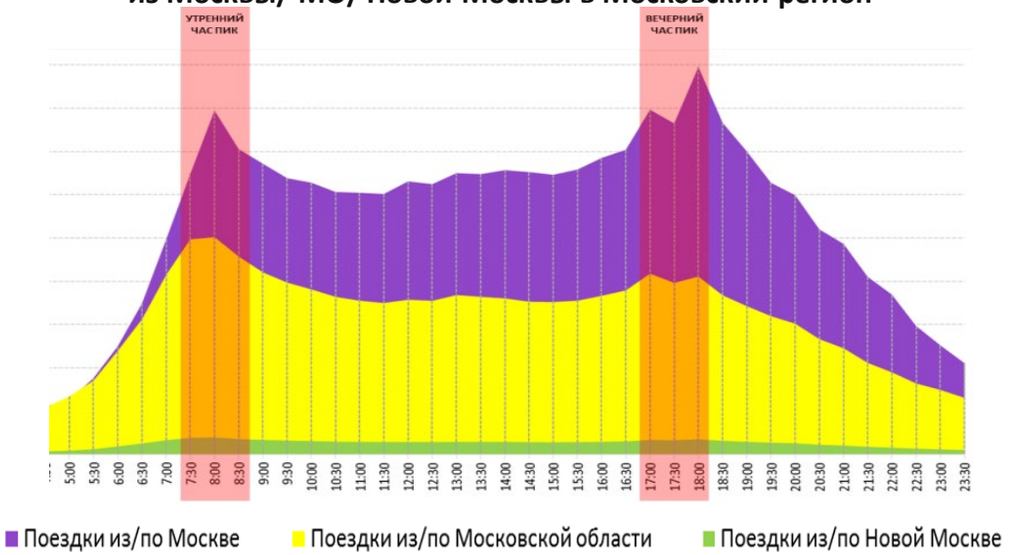
Ключевые выводы

- В Москве работает около 65% всего трудозанятого населения Московского региона
- Новая Москва – территория со значительным замещением контингента в структуре занятости:
  - в ней работает только 34% местного населения, имеющего работу
  - из всех работающих в Новой Москве более 54% составляют жители Москвы и области
- По будням утренний час пик начала поездок типа «дом-работа» из Москвы/МО/Новой Москвы приходится на 7:30-9:00, вечерний час пик – на 17:00-18:30.

Распределение трудозанятного населения Московского региона по месту работы



Изменение в течение суток средней по будням численности поездок из Москвы/ МО/ Новой Москвы в Московский регион

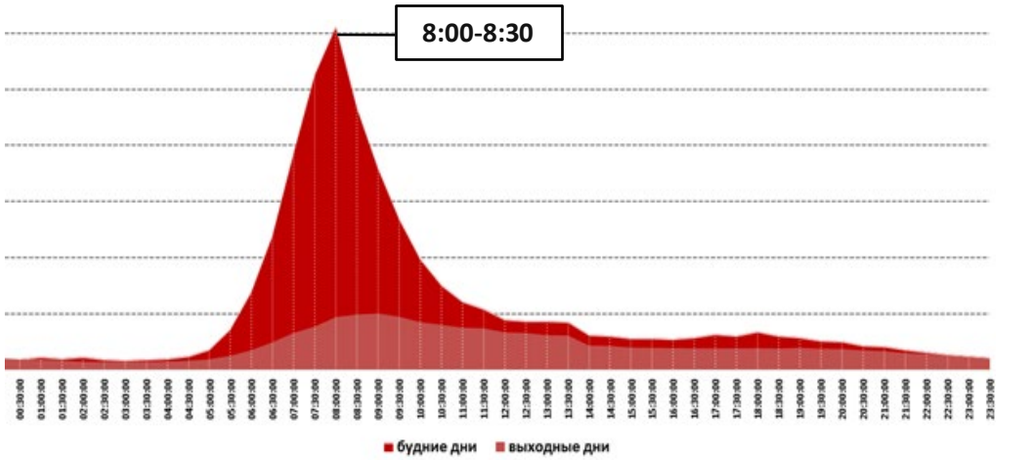


Связность территорий субъекта РФ на основе объемов перемещений населения

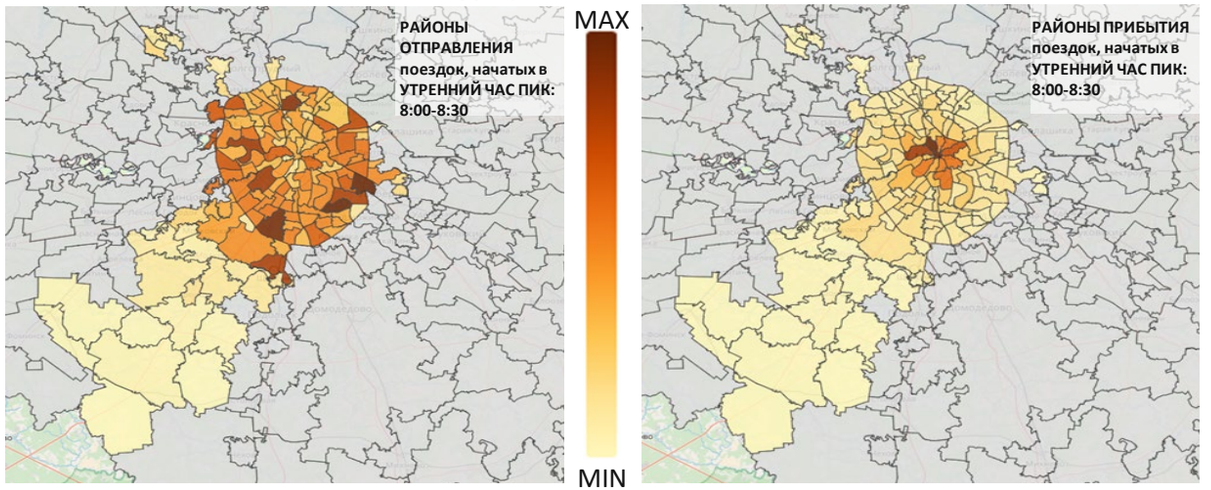
Описание данных

- Матрица корреспонденций содержит информацию о перемещениях населения между каждой парой районов города с разбивкой по получасовым интервалам каждой даты месяца
- В матрице отдельно выделены поездки типа «дом-работа» и «работа-дом»

Динамика изменения по получасовым интервалам количества поездок типа «дом-работа» (среднее отдельно для будних и выходных)



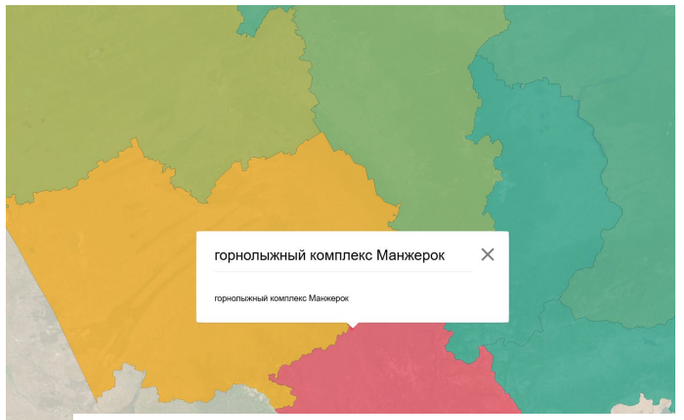
Пример распределения между районами отправления и прибытия в Москве объема поездок, начатых в утренний час пик



Мониторинг динамики численности туристов в горнолыжном комплексе Манжерок

Задача

- Выделение туристов среди других категорий населения
- Определение домашнего региона РФ или зарубежной страны
- Определение спроса на новый горнолыжный комплекс



Ключевые срезы данных

Анализ геоданных выполнен в следующих ключевых срезах:

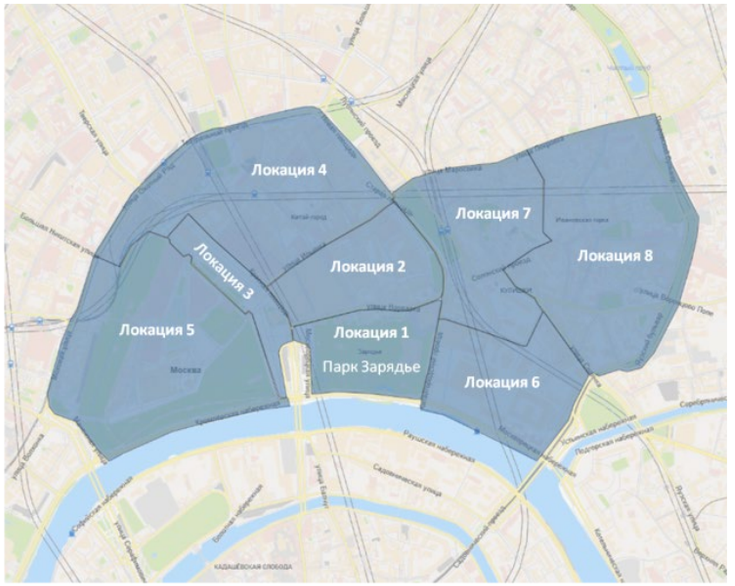
- Оценка общей численности туристов, посетивших Манжерок, с распределением их по странам и регионам РФ
- Распределение туристов по полу и возрасту
- Рейтинг туристов по домашнему региону/стране проживания
- Суммарные и средние траты туристов в Манжероке, в том числе в разбивке по домашнему региону туриста
- Обнаружено, что курорт более популярен не как горнолыжный, а как трекинговый для летнего посещения!

Оценка посещаемости туристами различных открытых городских локаций

Результат

- Проведен анализ посещаемости туристами центральной части Москвы на данных геоаналитики в целях развития туристической инфраструктуры на исследуемой территории

Разбиение центральной части столицы на отдельные локации



Ключевые показатели

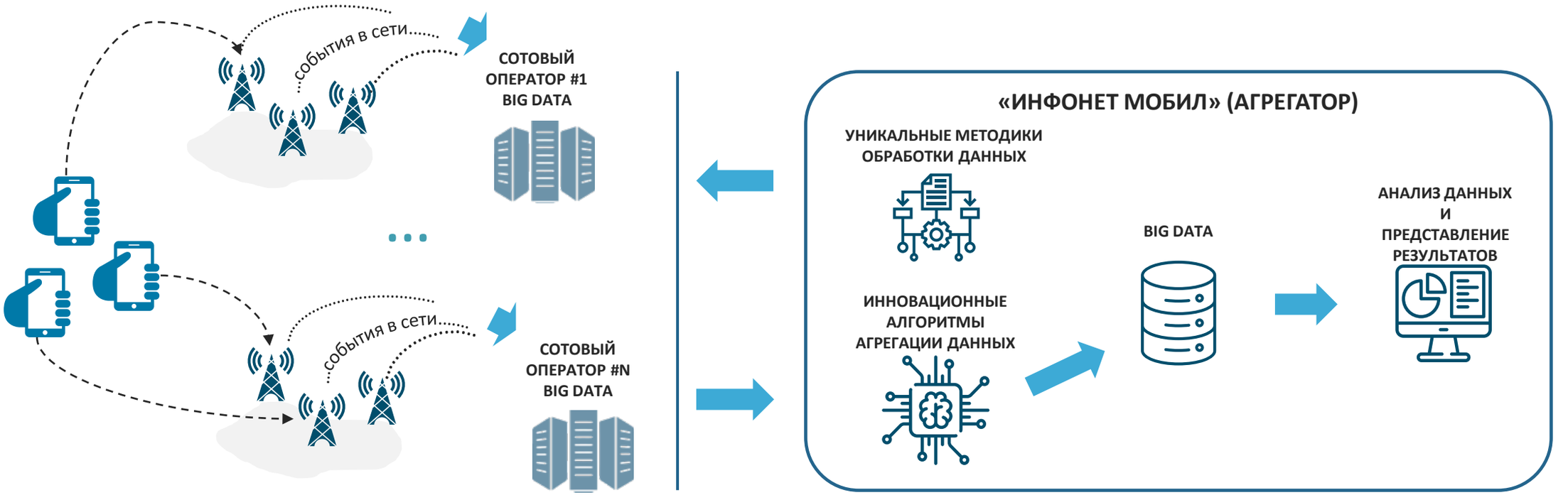
Анализ геоданных позволил определить:

- Объемы ежедневной посещаемости туристами центральных локаций Москвы с делением на категории (китайцы, иностранцы из ЕС, иностранцы из СНГ, остальные иностранцы, приезжие из регионов РФ)
- Динамику изменения по часам численности туристов в центральных локациях Москвы (в будни/выходные)
- Объемы и направления перемещений туристов по центральной части Москвы через локацию парка Зарядье

Как использовать большие данные в статистике

КАК происходит процесс сбора и обработки данных от мобильных операторов

InfoNet



Разработка уникального комплекса методов и алгоритмов в области получения агрегированных массивов данных с помощью технологических возможностей «Big Data»

Агрегированные наборы данных, получаемые от сотовых операторов, **полностью исключающие персональные данные**

Многоступенчатая верификация качества данных по критериям:

- проверка непротиворечивости значений
- перекрестная проверка значений
- соотнесение с внешними открытыми источниками информации

Поиск и ликвидация аномалий:

- ликвидация пропусков в данных
- ликвидация системных ошибок при перекрестном пересечении показателей

Агрегация данных - математические преобразования геоаналитических данных

Конвертация в геоаналитические отчеты, содержащие статистические показатели о населении

Участники процесса использования больших данных сотовых операторов для статистики населения

Национальная статистическая служба
формулирует задачи по расчету статистических показателей с использованием данных сотовых операторов и параметры их решения

Оператор сотовой связи
готовит исходные данные для расчета статистики и из них формирует предварительные агрегаты

Агрегатор
формулирует ТЗ для оператора сотовой связи, создает методики формирования предварительных агрегатов, методики верификации данных сотовых операторов, методики агрегации данных от разных операторов, методики досчета данных до полной статистики населения

Процесс использования данных сотовых операторов в статистике

InfoNet

Национальная статистическая служба устанавливает требования к выходным материалам и составляет ТЗ

Определяется подрядная организация (Агрегатор), заключается договор на выполнение работ по производству статистических оценок численности населения по данным операторов сотовой связи

Агрегатор составляет техническое задание для сотовых операторов на поставку предваритель-ных агрегатов в соответствии с требованиями и ТЗ

Агрегатор определяет необходимое и достаточное количество поставщиков статистических данных (операторов сотовой связи), которое обеспечит полноту покрытия исследуемой территории и охват населения, и заключает договор на поставку предваритель-ных агрегатов

Поставщик обрабатывает первичную информацию о событиях сотовой сети для определения местоположения абонентов, формирует предваритель-ные агрегаты, передает сформирован-ные предваритель-ные агрегаты Агрегатору

Агрегатор оценивает качество полученных от операторов сотовой связи предваритель-ных агрегатов, формирует итоговый агрегат – статистические данные о фактической численности населения

Агрегатор оценивает качество итогового агрегата и передает НСС подготовленные выходные материалы: статистические данные о численности населения на исследуемой территории; справочник территориаль-ного деления, справочник возрастно-половых групп и др.

Фактическая численность населения – базовый показатель

Методика расчета фактической численности населения – основа для составления других методик и расчета различных показателей статистики населения

Предоставление методики сотовым операторам

Агрегатор составляет техническое задание для оператора:

- методика определения места проживания абонента
- справочники территориальных зон и другие
- форматы данных
- регламент и сроки
- критерии оценки качества

Проверка данных от сотовых операторов

Агрегатор использует методики верификации данных сотовых операторов:

- проверка целостности данных
- соблюдение форматов данных
- проверка полноты данных
- проверка качества данных

Агрегация данных сотовых операторов

Агрегатор использует методики агрегации данных сотовых операторов:

- восстановление пропущенных значений на основе исторических данных и данных второго оператора
- исправление ошибок противоречивости данных и сглаживание аномальных значений на основе данных второго оператора
- агрегация

Досчет результатов

Агрегатор использует методики досчета данных до полной численности населения:

- досчет объединенных агрегированных данных для всех операторов связи на базе данных операторов, предоставивших данные
- досчет полученных данных в части групп населения, которые в меньшей степени используют или не используют мобильные устройства

Требования к Агрегатору при подготовке статистических отчетов на базе данных мобильных операторов

InfoNet

Требования к составу отчетов

- период учета населения
- исследуемая территория
- перечень выделяемых на исследуемой территории зон территориального деления
- перечень возрастно-половых групп

Какое население учитывать

- фактически проживающее население (граждане государства, иностранные граждане и лица без гражданства, при условии проживания в стране в течение последнего месяца)
- временно находящиеся на территории государства лица – отдельная категория
- учет – 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

Требования к формату

Статистические данные о численности населения - текстовый csv-файл, который содержит 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;
- количество лиц заданной возрастно-половой группы, временно проживающих на территории заданной зоны

Другие материалы:

- Справочник территориального деления
- Справочник возрастно-половых групп
- Отчет об оценке качества статистических данных
- Отчет об оценке качества предварительных агрегатов

Требования к исполнителю

- должен иметь успешный опыт (не менее 2 реализованных проектов) обработки данных сотовых операторов для целей статистики или анализа пространственных перемещений
- должен предоставить операторам связи методику формирование таблицы временных интервалов алгоритмы расчеты статистических метрик, а также методы верификации данных
- должен подготовить shape-файл исследуемой территории, обеспечив необходимое территориальное деление
- должен обеспечить верификацию данных каждого оператора, а потом провести агрегацию данных
- должен обеспечить аналитику и визуализацию полученных данных

Процесс формирования статистических отчетов операторами

InfoNet

Методики сбора и обработки первичной информации, регистрируемой системами операторов, сформированные на протяжении более 10 лет сотрудничества.
Методики учитывают разницу оборудования, ПО оператора и другие особенности сетей

Формирование ТЗ для оператора, которое включает терминологию, определения и четкий состав каждого отчета, предоставляемого оператором, в том числе форматы данных, описание полей отчета и разрешенных символов

Предоставление отчетов операторами, проверка корректности отчета и соответствия его состава техническому заданию, итерации обмена замечаниями и исправлениями

Пример формулировки определений для ТЗ

- **Проживающее население** - население, для которого существует локация (локация – ограниченная территория на которой человек находится в состоянии покоя согласно техническим событиям сети сотовой связи) в границах рассматриваемых территорий, на которой оно провело в календарном месяце наибольшее количество ночных часов с 23:00 до 06:00, но не менее 20% всех ночных часов календарного месяца. Такая

Пример описания формы отчета для ТЗ

Номер	Название поля	Описание	Формат
1	dt	Календарный месяц, за который собирались данные.	YYYY.MM.DD
2	zid	Идентификатор территории проживания или работы	Целое число
3	cnt_home	Количество проживающего населения	Целое неотрицательное число
4	cnt_job	Количество работающего населения	Целое неотрицательное число

Пример отчета оператора

```
month;zid;age;gender;cnt_home;cnt_work;cnt_day;cnt_night
2022-10-31;01;1;F;351146;154318;369986;369786
2022-10-31;01;1;M;330530;155913;373116;371143
2022-10-31;01;1;U;0;0;0;0
2022-10-31;01;2;F;339669;154684;341113;341354
2022-10-31;01;2;M;290941;126645;303971;302352
2022-10-31;01;2;U;0;0;0;0
2022-10-31;01;3;F;244317;101529;243364;244366
2022-10-31;01;3;M;189467;83761;190675;190073
2022-10-31;01;3;U;0;0;0;0
2022-10-31;01;4;F;238238;81912;238508;238238
```

Агрегатор: верификация статистических данных операторов

1. **Проверка целостности данных**
2. **Соблюдение форматов данных**
3. **Проверка полноты данных**
4. **Проверка качества данных**

Непротиворечивость показателей отчета
Пример:
Численность населения, находящегося себя дома, у себя на работе, не должно превышать численность всего населения, находящегося в этой зоне

Распределение показателей по зонам разбиения
Пример:
Визуальная проверка может помочь выявить такие изъяны в данных, как наличие населения, проживающего на территории парков, лесов, водоемов и т.п.

Сравнение данных с внешними источниками
Пример:
Распределение численности населения можно сравнить с данными Федеральной службы государственной статистики

Сравнение данных одного отчета по разным разбиениям
Пример:
Сравнение данных о численности населения территории в разбиении на ячейки 500*500 и на районы

Сравнение данных с предыдущими периодами
Пример:
Изменение численности во времени и выявление аномальных тенденций в сравнении с аналогичным периодом в прошлом, которые могут означать ошибку

Распределение показателей по времени
Пример:
Изменение плотности населения на территории во времени и внезапные повышения плотности (не всегда это ошибка, может быть массовое мероприятие)

Сравнение данных из разных отчетов
Все отчеты представляют собой связанные наборы данных, поэтому отдельным важным этапом проверки являются перекрестные проверки между отчетами

Агрегатор: поиск и ликвидация аномалий в данных

InfoNet

Технология формирования итоговых статистических показателей заключается в использовании **методов восстановления многопараметрического распределения всего населения по территории во времени на основе данных поставщиков** и напрямую зависит от качества проверки данных операторов, так как включает в себя исправление этих ошибок

Восстановление пропущенных значений на основе исторических данных и данных второго оператора для отдельных территорий с учетом различных групп населения	Такое возможно, потому что у операторов мобильной связи неравномерное покрытие территорий. Так, например, данные одного оператора дополняют данные второго в тех территориях, в которых у второго оператора нет покрытия с учетом его доли на данной территории
Восстановление пропущенных значений на основе исторических данных и данных второго оператора для отдельных временных сегментов	Например, при оценке численности населения на территории в определенный момент времени могут быть провалы в данных у некоторых операторов. В этом случае с учетом изменения во времени численности населения на заданной территории и доли оператора для второго оператора восстанавливаются данные в этот сегмент времени
Исправление ошибок противоречивости данных и сглаживание аномальных значений на основе данных второго оператора	Такое бывает, потому что, в системе сбора данных на стороне операторов возможны сбои. Например, колебания численности проживающего населения на территории должны оставаться в разумных пределах на протяжении всего года. Если же колебания сильные, то скорее всего они вызваны внешними воздействиями, поэтому должны совпадать у обоих операторов. В противном случае речь идет о наличии сбоя у одного из операторов.

Агрегатор: методики агрегации и досчета данных

Агрегация геоаналитических данных операторов сотовой связи, предоставивших данные	Агрегация геоаналитических данных в части численности населения проводится отдельно для постоянного и временного населения путем суммирования количества абонентов, для которых исследуемая территория является территорией постоянного проживания или временного пребывания (в том числе по отдельным категориям по длительности пребывания – для временного населения) соответственно. Агрегация по каждой категории населения осуществляется путем суммирования числа абонентов, относящихся к данной категории, каждого из представивших данных оператора сотовой связи.
Досчет объединенных агрегированных данных для всех операторов связи на базе данных операторов, представивших данные	Получение генеральной совокупности осуществляется путем деления агрегированных данных (по каждой группе абонентов по виду (постоянное, временное население, категории по социально-демографическим характеристикам), полученных на первом этапе, на суммарную долю операторов связи, геоаналитические данные которых агрегировались на первом этапе, на рынке услуг сотовой связи (по числу абонентов от общего числа абонентов).
Досчет полученных данных в части групп населения, которые в меньшей степени используют или не используют мобильные устройства	Экстраполяция показателей, полученных с использованием данных операторов сотовой связи об абонентах сотовой связи, осуществляется с учетом долей в численности населения детей и населения старше трудоспособного возраста, которые не используют сотовую связь. Для досчета указанных категорий населения осуществляется деление показателей, полученных для абонентов сотовой связи, на долю населения, не включающую детей до 14 лет и людей старше трудоспособного возраста.

InfoNet

info@infonet.ai

www.infonet.ai

ул. Россолимо, д. 17, стр. 3, г. Москва, РФ, 119021