

Руководство по использованию данных мобильных телефонов для целей официальной статистики

*Глобальная рабочая группа ООН по использованию больших данных
для целей официальной статистики*

Сентябрь 2019

Содержание

1. Введение.....	4
2. Список определений и сокращений	5
Определения, используемые в Руководстве.....	5
Сокращения для общих терминов.....	6
Сокращения для технических терминов, связанных с функционированием сетей мобильной связи	7
1. Применение.....	9
1.1. Использование данных мобильного позиционирования в производстве официальной статистики	9
Статистика путешествий в Эстонии	9
Статистика прибытий посетителей из-за границы в Индонезии	10
1.2. Данные мобильного позиционирования, используемые в исследованиях и пилотных проектах и применимые для целей статистики	10
Статистика туризма и статистика событий	11
Статистика населения.....	12
Статистика миграции	16
Статистика маятниковой миграции	17
Статистика транспортных потоков	18
Статистика занятости по приграничным и сезонным работникам.....	20
Другие применения.....	20
2. Источники данных	23
2.1. Данные из систем операторов мобильных сетей.....	23
Центральная система хранения	24
Данные зондирования и сигнализации	24
Данные активного позиционирования.....	26
2.2. Данные о событиях мобильных телефонов – данные пассивного позиционирования	27
Формы мобильных данных	27
Идентификационные данные абонента	28
Идентификационные данные оборудования	28
Временные характеристики	29
Характеристики местоположения	29
Идентификатор соты	32
Зона покрытия антенны.....	33
Дополнительные атрибуты данных о событиях	34
Дополнительные атрибуты сетевых данных.....	34

Дополнительные атрибуты для абонентов	34
2.3. Общий процесс извлечения данных	35
Подготовка данных	35
Анонимизация данных	36
Шифрование данных	39
Передача данных	39
Архивирование данных	40
Логический порядок шагов в процессе извлечения данных	40
3. Доступ к данным мобильных телефонов и модели партнерства	41
3.1. Введение	41
3.2. Обеспечение среды для доступа к данным мобильных телефонов для официальной статистики	44
Модели партнерств для использования данных мобильных телефонов для целей официальной статистики	45
Проект МСЭ по использованию больших данных для измерения информационного общества	53
Заинтересованные стороны: роли, возможности и полномочия	54
4 Методы	60
4.1. Понятия и определения	60
Понятия, связанные с поездками	61
Понятия, подходящие для нескольких областей статистики	62
Некоторые понятия, связанные со статистикой туризма	63
4.2. Методология обработки данных	64
4.3. Система обеспечения качества для производства статистики на основе данных сетей мобильной связи	67
4.4. Как справиться с недостаточным / избыточным охватом	69
Приложение 1 – Тематическое исследование: Франция	78
Приложение 2 – Тематическое исследование: Индонезия	85

1. Введение

В последние годы данные с мобильных телефонов стали одним из источников больших данных с хорошими перспективами для применения в официальной статистике. Ожидается, что данные с мобильных телефонов могут заполнить пробелы в данных, особенно в развивающихся странах, учитывая высокий уровень проникновения технологии. В своем докладе 2018 года¹ Международный союз электросвязи (МСЭ) показывает, что количество абонентов сотовой телефонной связи продолжает расти, несмотря на то, что абонентов уже больше, чем людей на Земле. Развивающиеся страны и наименее развитые страны будут продолжать сокращать отставание от развитых стран. Это демонстрирует, насколько всепроникающими являются мобильные телефоны. МСЭ уточняет, что сельские районы все еще отстают от городских, и это следует учитывать в исследованиях с использованием данных мобильных телефонов, но очевидно, что для этих данных характерен глобальный охват. Почти каждый человек в мире живет в пределах досягаемости сигнала мобильной сотовой связи.

Целевая группа по использованию данных мобильных телефонов изучала возможности использования этих данных для целей официальной статистики и подготовила это руководство, чтобы предоставить подробную информацию для стран, которые находятся в процессе разработки проекта по использованию данных мобильных телефонов для своей официальной статистики или находятся в процессе реализации этого проекта. Руководство будет опубликовано только в электронном формате на английском языке и, как ожидается, будет пересматриваться и изменяться по мере развития проектов.

Руководство состоит из четырех глав (Приложения, Источники данных, Доступ к данным мобильных телефонов и модели партнерства, Методы), за которыми следуют два тематических исследования по странам.

Это руководство составлено и отредактировано членами Глобальной рабочей группы по большим данным и, в частности, Целевой группы по использованию данных мобильных телефонов.

¹ Доклад об измерении информационного общества 2018: <https://www.itu.int/en/ITU-D/Statistics/Pages/publications/misr2018.aspx>

2. Список определений и сокращений

Определения, используемые в Руководстве

1. Данные мобильных телефонов или данные мобильного позиционирования (MPD) в контексте этого документа — это любой тип данных о событиях, связанных с использованием мобильного телефона, которые включают идентификатор абонента, атрибут времени и местоположение.
2. Оператор сети мобильной связи (МО), также известный как поставщик услуг беспроводной связи, оператор беспроводной связи, компания сотовой связи или оператор мобильной связи, является поставщиком услуг беспроводной связи, который владеет всеми элементами, необходимыми для продажи и предоставления услуг конечному пользователю, или контролирует их. Ключевой определяющей характеристикой оператора сети мобильной связи является то, что он должен владеть лицензией на использование радиочастотного спектра от регулирующего или государственного органа или контролировать ее применение. Второй ключевой определяющей характеристикой оператора сети мобильной связи является то, что он должен владеть элементами сетевой инфраструктуры, необходимыми для предоставления услуг абонентам в разрешенном радиочастотном спектре, или контролировать их.
3. Внутренние данные – любые события местоположения, происходящие внутри сети определенного оператора мобильной связи.
4. Данные входящего роуминга – любое событие местоположения абонента иностранного оператора мобильной связи в сети местного оператора
5. Данные исходящего роуминга - любое событие местоположения абонента местного оператора мобильной связи, совершенное в другой сети (обычно услуги роуминга иностранного оператора).
6. Подробная запись о вызове (CDR) – Запись метаданных телекоммуникационной операции (звонок, текстовое сообщение и т. д.), которая обычно используется для выставления счетов.
7. Подробная запись интернет-протокола (IPDR) или подробная запись о данных (DDR) – записи метаданных об обмене данными в мобильной сети.
8. Данные сигнализации – данные, полученные при помощи пассивного позиционирования (см. ниже)
9. Пассивное зондирование - мониторинг потоков данных между различными сетевыми объектами (в отличие от активного зондирования, которое генерирует трафик для мониторинга сети). Это предпочтительный способ сбора информации из сети мобильной связи.

10. Тип записи – каждое событие в сети можно различать по типам данных (CDR и IPDR), типам услуг (например, SMS, MMS, исходящий вызов и т. д.), и событиям, которые не создаются абонентом (например, обновления местоположения).
11. Хеширование, или маскирование, или псевдонимизация – процедура деидентификации, при которой поля персонально идентифицируемой информации в записи данных заменяются одним или несколькими искусственными идентификаторами или псевдонимами.
12. Сота (ячейка) сети – наименьший структурный элемент для систем сотовой связи. Может описываться азимутом, углом сектора, формой и размером зоны покрытия, типом антенны и местоположением. Каждая сота может обслуживать только ограниченное количество пользователей.
13. Базовые трансиверные станции (BTS) – узел, который поддерживает связь в пределах географической области и может иметь несколько связанных направленных антенн. В сетях LTE называется eNodeB.
14. Башня или мачта сотовой связи – совокупность передатчиков в одном конкретном месте, которое можно назвать объектом связи.
15. Объект сотовой связи – см. башня сотовой связи
16. Лонгитюдные данные – данные об абонентах за продолжительный период времени.
17. Обычная среда – географические границы, в которых человек перемещается в рамках своей обычной повседневной жизни

Сокращения для общих терминов

1. Национальный статистический институт (НСИ) – Национальная статистическая служба (НСС)
2. Международный союз электросвязи (МСЭ)
3. Орган, регулирующий системы электросвязи (TRA)
4. Национальное статистическое управление (НСУ)
5. Орган по защите данных (DPA)
6. Информационные и коммуникационные технологии (ИКТ)
7. Цели устойчивого развития (ЦУР)
8. Система обеспечения качества (QAF)
9. Межмашинное взаимодействие (M2M)
10. Интернет вещей (IoT)

Сокращения для технических терминов, связанных с функционированием сетей мобильной связи

1. Подробная запись о вызове (CDR)
2. Подробная запись интернет-протокола (IPDR) или подробная запись данных (DDR) – записи метаданных об обмене данными в мобильной сети
3. Регистр местоположения посетителей (VLR)
4. Глобальная система мобильной связи (GSM) – относится к сетям 2G
5. Универсальная система мобильной связи (UMTS) – относится к сетям 3G
6. Долговременное развитие (LTE) – относится к сетям 4G
7. Проект партнерства третьего поколения (3GPP)
8. Подсистема базовых станций (BSS)
9. Подсистема сетевой коммутации (NSS)
10. Центр коммутации мобильной связи (MSC)
11. Контроллер базовых станций (BSC)
12. Внутренний регистр местоположения (HLR)
13. Система поддержки операций/бизнеса (B/OSS)
14. Узел управления мобильностью (MME)
15. Международный идентификатор мобильного абонента (IMSI) – идентификатор SIM карты. Пока абонент не сменил SIM-карту, он останется прежним.
16. Номер мобильного абонента цифровой сети с интеграцией служб (MSISDN) – Идентификатор абонента. До тех пор, пока абонент не изменил свой номер телефона, он останется таким же.
17. Международный идентификатор мобильного оборудования (IMEI) – Идентификатор устройства. До тех пор, пока устройство используется, он будет таким же.
18. Международный идентификатор оборудования мобильной станции и версия программного обеспечения (IMEISV)
19. HLR
20. SGSN - узел обслуживания абонентов GPRS
21. Глобальный идентификатор соты (CGI)
22. Зона местоположения (LA)
23. Зона отслеживания (TA)
24. Международный идентификатор зоны местоположения (LAI) – временные данные абонента
25. Идентификация зоны отслеживания (TAI) – временные данные абонента
26. Код локальной зоны (LAC)/Код зоны отслеживания (TAC) – идентификатор локальной зоны в рамках сети мобильной связи

- 27. Сеть с коммутацией каналов (CS) – сеть, в которой группа сот образует локальную зону или зону отслеживания для сетей LTE
- 28. Мобильный код страны (MCC)
- 29. Мобильный код сети (MNC)
- 30. Область маршрутизации (RA)
- 31. Сеть с коммутацией пакетов (PA)
- 32. Идентификатор области маршрутизации (RAI) – сочетание LAI/TAI плюс RAC.

Временные данные абонента

- 33. Код области маршрутизации (RAC)
- 34. Идентификатор соты (CI, Cell ID) – уникален только внутри локальной зоны.
- 35. Глобальный идентификатор соты (CGI) – LAI/TAI/RAI и CI. Это уникальный идентификатор одной соты.

1. Применение

Предприятия и компании использовали большие данные уже в течение довольно продолжительного периода времени, в то время как национальные правительства, включая национальные статистические управления, а также международные организации, только начали изучать, как этот источник данных может быть использован для производства официальной статистики и, соответственно, для более эффективного принятия решений. За последнее десятилетие было разработано множество приложений, но цель этой главы состоит в том, чтобы рассмотреть некоторые из тех приложений, которые основаны на использовании данных мобильных телефонов и имеют отношение к производству официальной статистики, включая статистику туризма, статистику событий, статистику населения, статистику миграции, статистику поездок на работу, транспортные потоки/мониторинг, статистику транспорта, городское планирование, статистику занятости или статистику приграничных/сезонных работников.

1.1. Использование данных мобильного позиционирования в производстве официальной статистики

В настоящее время есть две страны, где производители статистики используют данные мобильного позиционирования (MPD) для регулярного производства официальной статистики – Эстония и Индонезия.

Статистика путешествий в Эстонии

Центральный банк Эстонии (Eesti Pank) с 2008 года сотрудничает с частными компаниями, чтобы получать большие данные для целей официальной статистики. Центральный банк использует данные мобильных телефонов от эстонских операторов мобильной связи для количественной оценки входящих и исходящих поездок, а также данные о платежах по кредитным картам для калибровки данных о расходах. Комбинация этих двух наборов данных и нескольких других наборов справочных данных позволяет своевременно и эффективно формировать статистику торговли услугами, используемую при составлении платежного баланса и национальных счетов.

Статистические данные о путешествиях на основе данных мобильных телефонов распространяются отдельно в качестве официальной статистики на веб-сайте Банка Эстонии ежеквартально с 2012 года, при этом временные ряды начинаются с 2008 года. Сотрудничество выдержало испытание временем в рамках процесса производства статистической информации центральным банком.

Статистика прибытий посетителей из-за границы в Индонезии

Чтобы расширить охват данных о количестве прибывших иностранных посетителей, особенно в приграничных районах, где их число не регистрируется на официальных пунктах пропуска иммиграционных служб, НСУ Индонезии (BPS) и Министерство туризма пытались улучшить методологию расчета количества иностранных посетителей, используя с октября 2016 года данные мобильного позиционирования.

Данные мобильного позиционирования используются в пунктах пересечения границы в районах, где иммиграционные проверки отсутствуют, а организация обследований в пунктах пересечения границы затруднена из-за географических условий.

Опубликованные НСУ Индонезии временные ряды данных о въездах посетителей, которые включают данные мобильного позиционирования, имеются с октября 2016 года. Модель сотрудничества с годами изменилась, как это отражено в Приложении 2.

1.2. Данные мобильного позиционирования, используемые в исследованиях и пилотных проектах и применимые для целей статистики

В настоящее время существует большой объем знаний, касающихся данных мобильного позиционирования, который увеличивается на сотни научных статей в год. Большинство исследований и пилотных проектов являются разовыми проектами, но все чаще появляются исследовательские организации, которые имеют регулярный доступ к данным мобильного позиционирования, что позволяет им выполнять работу, основанную на разработанных методах. В этой главе мы хотим дать представление о проделанной работе, и здесь обсуждаются только некоторые исследования и пилотные проекты с данными мобильного позиционирования, которые были проведены до настоящего времени.

Описание	Источник
Цель этого исследования состояла в оценке возможности использования данных мобильного позиционирования для формирования статистики по внутренним, выездным и въездным туристическим потокам, а также в рассмотрении сильных и слабых сторон, связанных с доступом, доверием, стоимостью, а также технологическими и методологическими проблемами, присущими использованию этого источника данных. Страна: Эстония	Возможность использования данных мобильного позиционирования для разработки статистики туризма. Консолидированный отчет по контракту Евростата No 3051.2012.001-202.452 (2014) Авторы: Rein Ahas Jimmy Armoogum, Siim Esko, Maiki Ilves, Epp Karus, Jean-Loup Madre, Ossi Nurmi, Françoise Potier, Dirk Schmücker, Ulf Sonntag, Margus Tiru Ссылка: http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Consolidated-report.pdf
Исследователи обнаружили, что национальное распределение телефонных звонков в Эстонии аналогично официальной статистике по средствам размещения. Они также обнаружили, что посетители из более отдаленных мест остаются в Эстонии на более длительный срок, и что для этих посетителей расхождение между статистикой телефонных вызовов и статистикой размещения меньше. Исследователи также изучили статистику событий и их влияние на статистику туризма. Они обнаружили, что сельскохозяйственные ярмарки и международные мероприятия очень сильно влияют на туризм. Страна: Эстония	<i>Mobile positioning data in tourism studies and monitoring: case study in Tartu, Estonia (2007)</i> Данные мобильного позиционирования для исследований и мониторинга туризма: пример из Тарту, Эстония (2007 г.) Авторы: Rein Ahas, Anto Aasa, Siiri Silm, Margus Tiru Ссылка: https://pdfs.semanticscholar.org/492b/53aa1c540ac9db823faefef3b1fef5f229b9.pdf

Исследователи изучили способность спортивных и культурных мероприятий формировать приверженность к посещению места и вызывать повторные посещения. В целом результаты показали, что все мероприятия являются хорошими генераторами новых посещений. Это имеет практические последствия для туристических организаций, поскольку дешевле сохранить существующих клиентов, чем пытаться найти новых. Таким образом, вместо того чтобы проводить дорогостоящие рекламные кампании для привлечения новых туристов, разумнее было бы использовать мероприятия и уделять больше внимания попыткам вернуть посетителей предыдущих мероприятий в Эстонию. Страна: Эстония	<i>The ability of tourism events to generate destination loyalty towards the country: an Estonian case study (2010)</i> Способность туристических мероприятий формировать приверженность к посещению страны: пример из Эстонии (2010 г.) Авторы: Andres Kuusik, Rein Ahas, and Margus Tiru Ссылка: https://ojs.utlib.ee/index.php/TPEP/article/download/878/855
Используя пассивное мобильное позиционирование, исследователи проанализировали пространственное поведение туристов в связи с погодой и климатом. Результаты исследования показали, что в Эстонии количество прибывающих туристов в летний период положительно коррелирует с температурой воздуха. Места с наибольшей корреляцией — это места для отдыха на побережье, где преобладает пляжный туризм. Поведение туристов, которые посещают внутренние районы с меньшим количеством туристических достопримечательностей, похоже, не подвержено влиянию погоды. Страна: Эстония	<i>Weather dependence of tourists' spatial behaviour and destination choices: study with passive mobile positioning data in Estonia (2007)</i> Зависимость от погоды пространственного поведения туристов и выбора места назначения: исследование с использованием данных пассивного мобильного позиционирования в Эстонии. (2007 г.) Авторы: Olle Järv, Anto Aasa, Rein Ahas and Erki Saluveer Ссылка: https://www.semanticscholar.org/paper/WEATHER-DEPENDENCE-OF-TOURIST'S-SPATIAL-BEHAVIOUR-J%C3%A4rvAasa/f2d5966c48a4a1db8f3a3f3828604737444371aa

Статистика населения

Описание	Источник
Для этой исследовательской работы данные обследования личных поездок, в основном, использовались для создания существенных ежедневных траекторий для системы моделирования и для валидации результатов оценки численности населения.	<i>A Study on Urban Mobility and Dynamic Population Estimation by Using Aggregate Mobile Phone Sources (Chapter 6)</i> Исследование городской мобильности и динамическая оценка населения с использованием агрегированных источников

Страна: Япония	данных мобильных телефонов (глава 6) Ссылка: http://www.csis.u-tokyo.ac.jp/dp/115.pdf
Автор предложил показатели статистики населения, которые могут быть получены на основе данных мобильного позиционирования, такие как: количество жилых домов, рабочих мест, школ, второго жилья и других регулярно посещаемых мест, смена рабочих мест, трансграничная миграция, сеточная статистика населения, статистика временного населения. Кроме того, была представлена оценка в реальном времени для конкретных мест и оценка рисков для правоохранительных органов. Страна: Эстония.	<i>Overview of the sources and challengers of mobile positioning data for statistics</i> Обзор источников и проблем использования данных мобильного позиционирования для целей статистики Автор: Margus Tiru Ссылка: https://unstats.un.org/unsd/trade/events/2014/Beijing/Margus%20Tiru%20-%20Mobile%20Positioning%20Data%20Paper.pdf
Исследовательская группа предложила систематическую методологическую основу для оценки плотности населения на основе данных из сети мобильной связи. В исследовании рассматривалась проблема использования сетевых данных (CDR и / или VLR) для оценки распределения плотности населения на общеевропейском уровне. Основная цель исследования состояла в том, чтобы разработать методологическую основу для сбора и обработки сетевых данных, которые могут быть убедительно применены ко многим операторам мобильной связи. Общеевропейский уровень	<i>Estimating population density distribution from network-based mobile phone data (2015)</i> Оценка распределения плотности населения на основе данных из сетей мобильной связи (2015 г.) Авторы: Fabio Ricciato, Peter Widhalm, Massimo Craglia, Francesco Pantisano Ссылка: https://ec.europa.eu/eurostat/cros/system/files/Final-%20jrc-AIT-MNO-study-compressed_1.pdf
Исследовательская группа разработала методы картирования населения для оценки плотности населения. Результаты исследования ясно показали оценку плотности населения в национальном масштабе, а также продемонстрировали сезонные изменения в распределении населения. Это исследование продемонстрировало, как анализ данных мобильных телефонов, которые ежедневно собираются операторами мобильной связи, может дополнять традиционные результаты переписи. Не только карты распределения населения могут быть такими же точными, как карты, составленные по данным переписи, а существующие методы	<i>Dynamic population mapping using mobile phone data (2014)</i> Динамическое картирование населения с использованием данных мобильных телефонов (2014 г.) Авторы: Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R. Stevens, Andrea E. Gaughan, Vincent D. Blondel, Andrew J. Tatem Ссылка: https://www.pnas.org/content/pnas/111/45/15888.full.pdf

локализации могут быть построены исключительно на основе данных мобильных телефонов, но эти данные предлагают дополнительные преимущества с точки зрения измерения динамики населения. Кроме того, сочетание данных мобильных телефонов и методов дистанционного зондирования способствует улучшению как пространственного, так и временного разрешения и демонстрирует, как можно создавать наборы данных о населении с высоким разрешением для любого периода времени. Страны: Португалия, Франция	
Группа исследовала огромные и постоянно обновляемые массивы информации, содержащейся в данных о вызовах по мобильным телефонам (CDR), для оценки статистики населения, связанной с местом проживания и перемещениями. В этом исследовании авторы предложили и опробовали процесс анализа, построенный на основе так называемого Социометра (<i>Sociometer</i>) – инструмента интеллектуального анализа данных для классификации пользователей по их привычкам совершения вызовов. Полученные результаты показывают, что оценки потоков с использованием метода, представленного в исследовании, более точны для крупных городов. Страна: Италия	<i>Use of mobile phone data to estimate mobility flows. Measuring urban population and inter-city mobility using big data in an integrated approach (2014)</i> Использование данных мобильных телефонов для оценки потоков перемещения. Измерение городского населения и перемещения между городами с использованием больших данных в рамках интегрированного подхода (2014 г.) Исследование было совместно проведено Istat, CNR и Университетом Пизы в рамках интересов комиссии по исследованиям, задачей которой был выбор для Istat направления работы с большими данными (“<i>Commissione di studio avente il compito di orientare le scelte dell’Istat sul tema dei Big Data</i>”) Авторы: Barbara Furletti, Lorenzo Gabrielli, Fosca Giannotti, Letizia Milli, Mirco Nanni, Dino Pedreschi, Roberta Vivio, Giuseppe Garofalo Ссылка: http://www.cisstat.com/BigData/CIS-BigData_06_Eng%20%20IT%20Mobile%20phone%20data.pdf
Понятие дневного населения относится к количеству людей, которые присутствуют на территории в обычные рабочие часы, включая работников. Это отличается от понятия «проживающего» населения, которое касается людей, проживающих в данном районе и обычно присутствующих в вечерние и ночные часы. Страна: Нидерланды	Визуализация дневного населения, проведенная НСУ Нидерландов Ссылка: https://www.cbs.nl/en-gb/our-services/innovation/project/how-many-people-here-

В последние годы муниципальное статистическое бюро Пекина в Китае создает платформу динамического мониторинга населения на основе данных мобильных телефонов и разрабатывает методы оценки численности населения. Платформа представляет собой новый источник данных и служит важным дополнением к традиционной статистике населения. Используя мобильные данные, муниципальное статистическое бюро Пекина ежемесячно получает данные о населении и вносит корректировки после согласования данных между муниципалитетами и районами. Платформа измеряет динамические потоки населения, отвечает потребностям органов государственного управления и, кроме того, является экономичной и обеспечивает своевременность данных. Страна: Китай	Нет данных
--	-------------------

Описание	Источник
Исследователи создали методологическую основу для изучения миграции населения и изменения климата при помощи использования данных мобильных телефонов. Исследование показало, что анализ, основанный на данных сети мобильной связи, может описать важные краткосрочные характеристики (часы-недели) передвижения людей во время и после экстремальных погодных явлений, которые чрезвычайно сложно количественно оценить с помощью стандартных исследований на основе обследований. Исследование также продемонстрировало, как можно анализировать взаимосвязь между фундаментальными параметрами моделей миграции в национальном масштабе на основе данных мобильных телефонов. В исследовании одновременно были количественно оценены частота, направление, продолжительность и сезонность эпизодов миграции в Бангладеш. Страна: Бангладеш	<i>Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh (2016)</i> Выявление скрытых моделей миграции и мобильности в регионах с неблагоприятным климатом: лонгитюдное исследование шести миллионов анонимных пользователей мобильных телефонов в Бангладеш (2016 г.) Авторы: Xin Lu, David J Wrathall, Pål Roe Sundsøy, Md. Nadiruzzaman, Erik Wetter, Asif Iqbal, Taimur Qureshi, Andrew Tatem, Geoffrey Canright, Kenth Engø-Monsen, and Linus Bengtsson Ссылка: http://www.flowminder.org/publications/unveiling-hidden-migration-and-mobility-patterns-in-climate-stressed-regions-a-longitudinal-study-of-six-million-anonymous-mobile-phone-users-in-bangladesh
Авторы описали, как мобильные телефоны могут стать новым источником данных о внутренней миграции. Анализ выявил более тонкие закономерности, которые не были обнаружены в государственном обследовании миграции. Исследование предоставило новый количественный взгляд на определенные модели внутренней миграции в Руанде, которые нельзя наблюдать при использовании стандартных методов обследования, и изучило способы использования новых форм информационных и коммуникационных технологий для лучшего понимания сути миграции в развивающихся странах. Страна: Руанда	<i>Inferring Patterns of Internal Migration from Mobile Phone Call Records: Evidence from Rwanda (2011)</i> Вывод о закономерностях внутренней миграции на основе данных о вызовах по мобильным телефонам: информация из Руанды (2011 г.) Автор: Joshua Blumenstock Ссылка: http://jblumenstock.com/files/papers/jblumenstock_itd2012_wp.pdf

Описание	Источник
Исследователи представили первые шаги в изучении моделей маятниковой миграции и функциональной регионализации с использованием данных о местоположении мобильного телефона. Основная цель заключалась в том, чтобы представить и обсудить потенциал данных о местоположении мобильных телефонов в качестве источника данных, альтернативного переписям, для картирования транспортных потоков при ежедневных поездках на работу и обратно с последующей функциональной регионализацией. Был предоставлен набор аналитических карт, отражающих различные аспекты регулярной маятниковой миграции населения и функциональной регионализации. Данные о местоположении мобильных телефонов могут предоставить информацию о маятниковой миграции для большинства населения с почти произвольно выбранными моментами времени и частотой повторения, с большей пространственной детализацией и намного дешевле, чем традиционные средства сбора данных. Более того, временной интервал между моментом обследования и моментом доступности данных можно сократить почти до режима реального времени. Страна: Эстония	<i>Application of mobile phone location data in mapping of commuting patterns and functional regionalization: a pilot study of Estonia (2013)</i> Применение данных о местоположении мобильных телефонов для картирования моделей маятниковой миграции и функциональной регионализации: пилотное исследование в Эстонии (2013 г.) Авторы: Jakub Novak, Rein Ahas, Anto Aasa & Siiri Silm Ссылка: http://www.tandfonline.com/doi/pdf/10.1080/17445647.2012.762331?needAccess=true
Исследователи проанализировали взаимосвязь между мобильностью людей и экономической активностью в городской среде, используя данные мобильных телефонов. В статье указано, что понимание этой взаимосвязи особенно актуально для крупных городов в развивающихся странах, которые являются динамичными, взаимосвязанными и плотно заселенными, но плохо охвачены традиционными методами сбора данных. Чтобы добиться прогресса в этом вопросе, в исследовании был разработан набор данных о потоках маятниковой миграции, извлеченных из данных об операциях мобильных телефонов на Шри-Ланке, и проанализированы эти данные с использованием простой модели городской экономики. Модель сопоставляет	<i>Estimation of urban commuting patterns using cellphone network data (2012)</i> Оценка городской маятниковой миграции с использованием данных сетей мобильной связи (2012 г.) Авторы: Vanessa Frías-Martínez, Cristina Soguero-Ruíz, Enrique Frías-Martínez Ссылка: https://pdfs.semanticscholar.org/4af0/36dcdbd2f0a26f01c0e1b90fedelc94821782.pdf

потоки поездок на работу и обратно с географическим распределением ключевых экономических показателей, таких как объем производства и доход домохозяйств. Исследование показало, что потоки маятниковой миграции, а также географическое распределение объема производства и дохода, оцененное с помощью модели, отражают основные черты городской структуры Коломбо. Исследование также показало, что полученный путем моделирования показатель дохода имеет высокую предсказательную силу для ночного освещения – замещающей переменной для дохода домашних хозяйств. Страна: Шри-Ланка	
---	--

Статистика транспортных потоков

Описание	Источник
В статье представлены новые взгляды на структуру транспортных потоков и объяснено, каким образом и в какой степени поездки жителей пригорода повлияли на движение в час пик. Они обнаружили, что ежедневные поездки на работу и обратно и жители пригородов повлияли на потребность в транспорте из-за увеличения трафика в вечерние часы пик, хотя ежедневные поездки на работу составляли только 31% от передвижений в этот период. В работе также отмечается, что вечернее движение в час пик по пятницам отличалось от движения в другие дни, и предполагалось, что это связано с внутренним туризмом и проведением досуга, что позволяет предположить, что общее перемещение людей, связанное с изменениями в социальном поведении, играет большую роль для условий дорожного движения в вечерние часы пик по сравнению с влиянием развития пригородов. Страна: Эстония	<i>Mobile Phones in a Traffic Flow: A Geographical Perspective to Evening Rush Hour Traffic Analysis Using Call Detail Records (2012)</i> Мобильные телефоны в транспортном потоке: географический взгляд на анализ дорожного движения в вечерний час пик с использованием подробных записей о вызовах по мобильным телефонам (2012 г.) Авторы: Olle Järv, Rein Ahas, Erki Saluveer, Ben Derudder, Frank Witlox Ссылка: https://doi.org/10.1371/journal.pone.0049171
Исследователи оценили время в пути на участках дороги, сопоставив последовательность зашифрованных сигнальных сообщений (для сохранения анонимности) для каждого мобильного устройства с физическим движением по дороге. В среднем их подход давал результаты на 3 минуты быстрее, чем традиционный подход мониторинга. Кроме того, их подход обеспечил	<i>The Cellular Network as a Sensor: From Mobile Phone Data to Real-Time Road Traffic Monitoring (2015)</i> Мобильная сеть в качестве датчика: от данных мобильных телефонов к дорожному движению в режиме реального времени (2015 г.)

улучшение на 25% по сравнению с наименьшей средней длиной участка дороги, которую наблюдали другие традиционные и существующие системы мониторинга дорог. Наконец, оценки времени поездки, полученные с помощью этого метода, можно было проверить вручную, чтобы предсказать сигналы о возможных эпизодах заторов.	Авторы: Andreas Janecek, Danilo Valerio, Karin Anna Hummel, Fabio Ricciato, Helmut Hlavacs Ссылка: https://www.researchgate.net/publication/275273390_The_Cellular_Network_as_a_Sensor_From_Mobile_Phone_Data_to_RealTime_Road_Traffic_Monitoring
Исследователи использовали анонимные данные мобильных телефонов, чтобы оценить изменения в структуре дорожного движения, вызванные праздничными днями, и обсудили, как модели дорожного движения меняются каждый день в праздничные периоды (перед праздником, во время праздника и после праздника) Ожидалось, что результаты этого исследования помогут транспортным инженерам и руководителям программ внедрить соответствующие контрмеры, для снижения заторов, для уменьшения пробок на проезжей части дороги во время самых загруженных праздничных периодов. Результаты показали, что праздничные дни повлияли на дневной и почасовой объем трафика и скорость движения, и что влияние было различным для движения в южном и северном направлениях на исследуемой территории. Страна: Китай	<i>Understanding Intercity Freeway Traffic Variability at Holidays Using Anonymous Cell Phone Data (2016)</i> Понимание изменчивости движения на междугородних автострадах в праздничные дни с использованием анонимных данных мобильных телефонов (2016 г.) Авторы: Gang Liu, Chenhao Wand, Tony Z. Qiu Ссылка: https://www.researchgate.net/publication/304620561_Understanding_Intercity_Freeway_Traffic_Variability_at_Holidays_Using_Anonymous_Cell_Phone_Data
Авторы признали, что заторы на дорогах становятся все более серьезной проблемой в странах, переживающих экономический рост. Данные необходимы для выявления узких мест и определения приоритетных направлений для улучшений. В этом исследовании авторы хотели понять, могут ли большие данные мобильной сети помочь в планировании транспорта в городе Коломбо, Шри-Ланка, путем создания матриц отправления и назначения, которые описывают поток дорожного движения между различными районами и определяют, где источники дневной маятниковой миграции в город Коломбо. Результаты их исследования показали, что в развивающихся странах с ограниченными сенсорными сетями, таких как Шри-Ланка, данные активного и пассивного позиционирования от операторов сетей мобильной связи могут дать актуальную	<i>The potential of mobile network big data as a tool in Colombo's transportation and urban planning (2016)</i> Потенциал больших данных сетей мобильной связи как инструмента городского и транспортного планирования в г. Коломбо (2016 г.) Авторы: Sriganesh Lokanathan, Gabriel Kreindler, Nisansa de Silva, Yuhei Miyauchi, DedunuDhananjaya, Rohan Samarajiva Ссылка: https://www.researchgate.net/publication/269465996_Using_Mobile_Network_Big_Data_for_Informing_Transportation_and_Urban_Planning_in_Colombo

информацию, которая может повысить эффективность и надежность транспортной системы. Страна: Шри-Ланка	
В этом документе представлен широкий обзор основных исследований и проектов, направленных на использование данных, полученных из сетей мобильной связи, для получения оценки местоположения и движения для отдельных лиц, в качестве отправной точки для дальнейших исследований по управлению аварийными ситуациями и дорожным движением. Авторы также презентовали результаты проекта Current City, который представлял собой тестовую систему в Амстердаме для извлечения данных с мобильных телефонов и анализа пространственных моделей активности в сетях. Страна: Нидерланды	<i>Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities (2013)</i> Данные мобильных телефонов из GSM сетей для оценки параметров дорожного движения и городских пространственных характеристик: обзор применения и возможностей (2013 г.) Авторы: John Steenbruggen, Maria Teresa Borzacchiello, Peter Nijkamp, Henk Scholten Ссылка: https://link.springer.com/article/10.1007/s10708-011-9413-y

Статистика занятости по приграничным и сезонным работникам

Описание	Источник
Исследование продемонстрировало, как данные мобильных телефонов могут быть использованы для быстрого и точного обнаружения, отслеживания и прогнозирования экономических изменений на нескольких уровнях. В этом исследовании показано, как данные мобильных телефонов могут быстро дать представление об уровне занятости именно потому, что модели общения людей меняются, когда они не работают. В частности, исследователи отметили, что их новые методы позволили им предсказать нынешние уровни безработицы за две-восемь недель до публикации традиционных оценок и спрогнозировать будущие уровни занятости вплоть до четырех месяцев раньше официальных отчетов. Их алгоритмы также предсказывали будущие состояния.	<i>Tracking employment shocks using mobile phone data (2015)</i> Отслеживание шоков в области занятости при помощи данных мобильных телефонов (2015 г.) Авторы: Jameson L. Toole, Yu-Ru Lin, Erich Muehlegger, Daniel Shoag, Marta C. Gonzalez, David Lazer Ссылка: http://rsif.royalsocietypublishing.org/content/ro-yinterface/12/107/20150185.full.pdf

Другие применения

Ряд исследовательских работ и примеров показывают, что данные мобильных телефонов играют очень важную роль в различных областях, таких как здравоохранение, социально-экономическая область, реагирование на стихийные бедствия, управление городским хозяйством и т. д. Список таких исследовательских работ приводится ниже:

Тема	Источник
Развитие	Данные сетей мобильной связи для развития, <i>Global Pulse</i>, 2013 г. Ссылка: http://www.unglobalpulse.org/sites/default/files/Mobile%20Data%20for%20Development%20Primer_Oct2013.pdf
Общественное благо	Данные мобильных телефонов для общественного блага, <i>Global Pulse, GSMA</i>, 2015 г. Ссылка: http://www.gsma.com/mobilefordevelopment/wp-content/uploads/2017/02/State-of-Mobile-Data-for-Social-Good-Preview-Feb-2017.pdf
Моделирование эпидемий	Об использовании показателей мобильности людей для моделирования эпидемий, 2014 г. Ссылка: http://perso.uclouvain.be/adeline.decuypere/docs/journal.pcbi.1003716.pdf
Передача малярии	Статистика миграции, имеющая отношение к передаче малярии в Сенегале, полученная на основе данных мобильных телефонов и использованная в агент-ориентированной модели миграции, 2016 г. Ссылка: http://www.geospatialhealth.net/index.php/gh/article/view/408/357
Вспышка лихорадки Эбола	В 2015 году МСЭ запустил проект больших данных, чтобы продемонстрировать потенциал больших данных для облегчения своевременного обмена информацией для борьбы с эпидемией лихорадки Эбола, охватившей Западную Африку в 2014 году, и будущими кризисами в области здравоохранения. В проекте использовались подробные данные о телефонных вызовах (CDR), которые включают информацию об использовании мобильных телефонов, в том числе местоположение, от операторов мобильной связи в Либерии, Гвинее и Сьерра-Леоне. Проект продемонстрировал, как анализ данных CDR может предоставить информацию о мобильности людей, включая трансграничные перемещения, и пространственно-временном распределении людей, обеспечивая при этом конфиденциальность личных данных. Веб-сайт проекта МСЭ по борьбе с лихорадкой Эбола доступен по адресу: http://www.itu.int/net4/ITU-D/CDS/projects/display.asp?ProjectNo=7RAF15087
Распространение вспышек холеры	Данные мобильных телефонов подчеркивают роль массовых скоплений людей в распространении вспышек холеры Ссылка: http://www.pnas.org/content/113/23/6421.full.pdf
Бедность и богатство	Прогнозирование бедности и богатства на основе метаданных мобильных телефонов Ссылка: http://www.uvm.edu/~cdanfort/csc-reading-group/blumenstock-science-2015.pdf
Сезонная мобильность	Анализ моделей сезонной мобильности с использованием данных мобильных телефонов, <i>Global Pulse project series, No. 15</i>, 2015 г. Ссылка: http://www.unglobalpulse.org/sites/default/files/UNGP_ProjectSeries_Mobility_Senegal_2015_0.pdf

Общественное благо/ЦУР	Мобильный доступ к данным на благо общества / ЦУР: примеры и предостережения, <i>Flowminder</i> Ссылка: http://www.oecd.org/std/Flowminder-OECD-2015-Dec.pdf
Землетрясение в Непале	Оперативные и близкие к реальному времени оценки перемещения населения с использованием данных мобильных телефонов после стихийных бедствий: землетрясение в Непале в 2015 году, 2016 г. Ссылка: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4779046/

2. Источники данных

2.1. Данные из систем операторов мобильных сетей

Извлечение данных из системы оператора мобильной сети (МО) зависит от конкретных технических решений, которые использует МО². Как правило, у каждого оператора мобильной связи есть биллинговый центр и, скорее всего, хранилище данных, где данные периодически хранятся для выставления счетов и дальнейшего анализа в целях планирования и управления сетью (рисунок 2.1). Помимо центральных систем хранения, где данные легко доступны, можно извлекать данные в процессе зондирования на уровнях подсистемы базовой станции (BSS) и подсистемы сетевой коммутации (NSS).

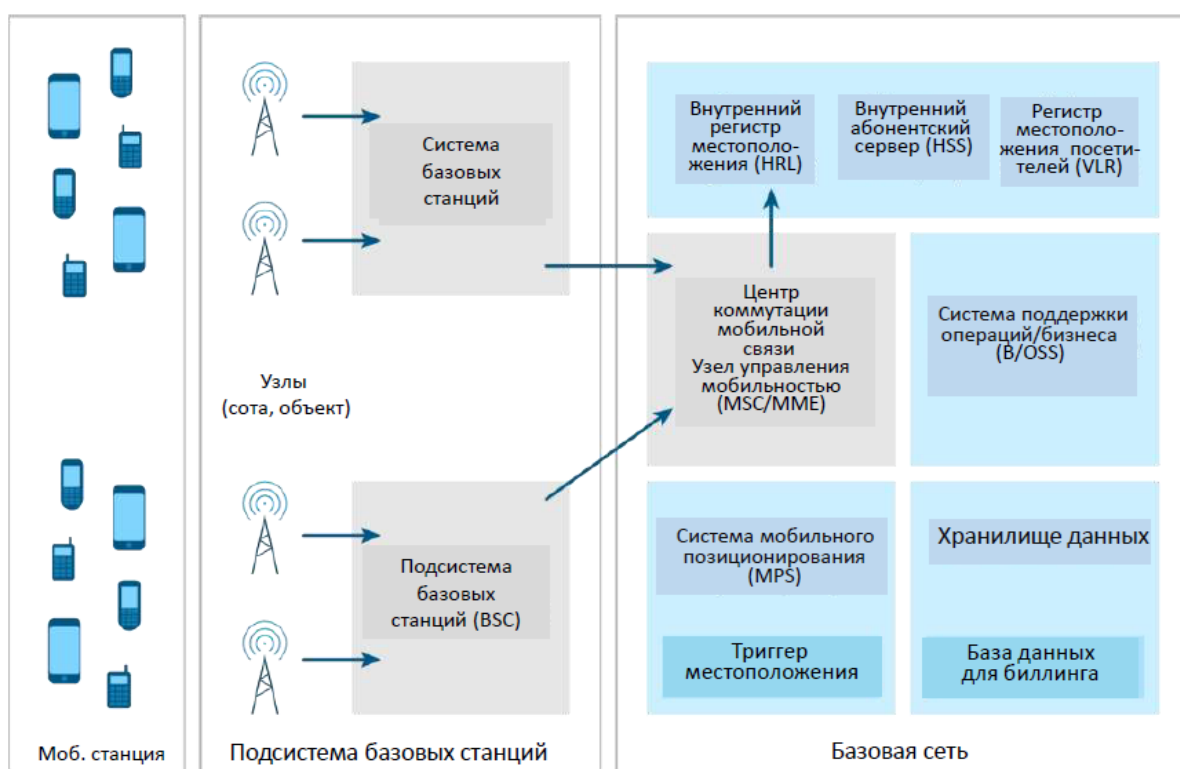


Рисунок 2.1 Источники данных мобильного позиционирования в системе оператора³

² Беспроводные технологии (такие как 2G, 3G и 4G), которые используются в мобильных телефонах и телекоммуникационных сетях по всему миру, соответствуют различным стандартам и спецификациям, разработанным и поддерживаемым такими организациями, как ETSI, 3GPP, Qualcomm и IEEE. В Европе (а также в Японии и Китае) наиболее распространенными технологиями мобильной связи сегодня являются GSM (2G), UMTS (3G) и LTE (4G), которые соответствуют техническим спецификациям Проекта партнерства третьего поколения (3GPP). По этой причине большая часть описаний сетевой архитектуры операторов мобильной связи и доступа к данным из систем операторов мобильной связи в этом документе будет основана на спецификациях 3GPP.

³ Tiru, M., Ahas, R. (2012) "Passive Anonymous Mobile Positioning Data for Tourism Statistics" (*Данные пассивного анонимного мобильного позиционирования для статистики туризма*). 11 Глобальный форум по статистике туризма <http://congress.is/11thtourismstatisticsforum/papers/Session3.pdf>

Центральная система хранения

Извлечение данных из центральных систем хранения — это самый простой способ получить данные для статистики на национальном уровне. Есть несколько мест, где можно найти такие данные, собранные для разных целей и содержащие различную информацию:

- В домене биллинга хранятся подробные записи о вызовах (CDR) и подробные записи интернет-протокола (IPDR) для целей начисления платы. Данные будут храниться здесь после успешной процедуры создания записи о начислении платы.
- Базы данных клиентов содержат информацию о пользователях, которая может повысить ценность данных CDR или IPDR — например, социально-демографические характеристики и место жительства.
- Хранилища содержат массивы данных из разных источников, но не всегда могут быть легко доступны из-за огромного количества хранящихся в них данных.

Базы данных в биллинговом домене обычно считаются наиболее доступными. Данные, хранящиеся в биллинговом домене, собираются промежуточной системой от различных сетевых объектов, ответственных за предоставление различных типов услуг.

Данные зондирования и сигнализации

Существует два типа зондирования — активный и пассивный — для мониторинга и сбора данных из телекоммуникационных сетей. Активное зондирование используется оператором сети для генерации трафика с целью мониторинга рабочих характеристик сети, тогда как пассивные устройства контролируют потоки данных между различными сетевыми объектами. В связи с этим пассивное зондирование является одной из возможностей для сбора информации из сетей операторов мобильной связи и обычно является предпочтительным вариантом.

Возможности пассивного зондирования обычно реализуются путем развертывания лицензионного программного обеспечения вместе с необходимым оборудованием в сети. Есть несколько возможных мест для таких зондов, но чаще всего они устанавливаются как можно более локально в системе. Эти приборы также могут запрашивать базы данных, такие как регистр местоположения посетителей (VLR), в которых хранятся соответствующие пользовательские данные.

С точки зрения увеличения детализации наблюдений в данных, использование данных от зондов дает множество преимуществ:

- Предоставление доступа к типам данных, которые обычно не хранятся в биллинговых центрах и системах хранилищ данных;
- Увеличение количества записей на одного абонента помогает минимизировать дневные и суточные различия в количестве записей, которые являются результатом структуры вызовов абонентов.

Однако использование данных от зондов (данные сигнализации) имеет свои недостатки:

- Ожидается, что системы мониторинга или сбора данных будут «онлайн», чтобы постоянно собирать и хранить полученные данные, которые являются временными по своей природе;
- Дополнительная нагрузка на сеть для операторов мобильной связи;
- Не вся информация является обязательной для хранения в сети мобильной связи во время зондирования (например, идентификатор зоны местоположения (LAI) является обязательным, а более точный глобальный идентификатор соты (CGI) – нет);
- Значительно увеличивает количество событий на одного абонента, а это означает, что необходимо выделить больше ресурсов для хранения и обработки таких данных;
- Вовлечение данных, генерируемых системой, в статистический анализ может привести к высокому уровню фоновых шумов, который необходимо устранить на этапе обработки данных.

Стоимость установки зондирующих систем в системах операторов мобильной связи обычно высока, и она не будет внедряться просто для того, чтобы генерировать статистические показатели для НСУ. Но если операторы мобильной связи уже внедрили какую-то систему зондирования и необходимое хранилище данных, это является хорошим источником данных для статистических показателей, поскольку включает гораздо больше данных по сравнению с «традиционно» собранными CDR из биллинговой системы и хранилища данных. Данные зондирования ограничиваются входящим роумингом и внутренними данными, обычно данные сигнализации не передаются между операторами мобильной связи, поэтому они исключаются из данных исходящего роуминга.

Данные сигнализации обычно относятся к прямому получению сигналов передачи из радиосети и сохранению их в базе данных. Как и данные зондирования, данные сигнализации также очень объемны и могут быть ограничены входящим роумингом и внутренними данными с исключением данных исходящего роуминга.

Данные активного позиционирования

Данные активного позиционирования генерируются путем определения местоположения мобильного абонента с помощью устройство-центричных и сетевых методов, а также через спутник (например, GPS). Эти методы используются либо для предоставления услуг в определенном месте, либо в ответ на национальные правила, требующие сбора высокоточных данных о местоположении (например, для целей служб спасения и полиции). Возможности активного позиционирования широко распространены в системах операторов в развивающихся странах и часто используются только на индивидуальной основе. Общие методы активного позиционирования включают триангуляцию в сети сотовой связи или технологию A-GPS (таблица 2.1).

Исследователи использовали активное мобильное позиционирование с небольшими выборками для сравнительного анализа и в исследованиях, требующих постоянного точного позиционирования.

Таблица 2.1. Сводка основных типов источников для данных мобильного позиционирования. *Источник: Авторы*

Источник данных		Насколько просто?	Что?	Сколько данных?	Насколько адекватно для статистики?
Пассивное	Операции сотовой связи	Просто, стандарт	Интенсивность использования телефона на уровне соты		
	Подробные записи о вызовах (CDR+DDR)	Нужны защита конфиденциальности и методы обработки	Операции сотовой связи		
	Зонды/сигнализация	Требуется сбор, защита конфиденциальности и методов обработки	Журналы операций сотовой связи и передачи данных; личные характеристики от оператора		
Активное	Системы мобильного позиционирования	Просто для малых выборок; обычно требует согласия	Точное позиционирование, разная частота; вопросник для возможных респондентов		

2.2 Данные о событиях мобильных телефонов – данные пассивного позиционирования

Событие на мобильном телефоне рассматривается как действие, инициированное абонентом или мобильным устройством или предназначенное для них. Большинство этих событий генерируют данные пассивного позиционирования. Когда эти события происходят, они регистрируются разными сетевыми объектами. Например, когда мобильное устройство инициирует обновление зоны местоположения, база данных, такая как VLR, хранит новый код зоны местоположения (LAC) для мобильного устройства. Аналогичным образом CDR и IPDR генерируются каждый раз, когда абонент использует такие услуги, как вызов (входящий / исходящий), обмен сообщениями (входящие / исходящие) или доступ к Интернету. Данные о событиях также могут поступать от приборов – источников зондирования или сигнализации и включать CDR и IPDR, а также от других технических событий, таких как обновления зоны местоположения или другие операционные действия, которые генерируют событие вместе со ссылкой на вышку сотовой связи. Для каждого такого события сеть мобильной связи собирает данные, относящиеся к событию, а это означает, что список характеристик часто зависит от события. Более того, только сеть мобильной связи может записывать некоторые характеристики, что также делает их наличие специфичным для конкретной мобильной сети.

Минимальные характеристики, необходимые для статистики населения, включают идентификатор абонента, атрибут времени и местоположение. Любые доступные дополнительные данные могут помочь повысить ценность и удобство использования данных мобильного позиционирования и потому являются желательными, если их возможно собрать.

Формы мобильных данных

1. **Внутренние данные:** любые события определения местоположения, происходящие в сети определенного оператора мобильной связи. Это CDR или другие события, в которых участвует домашний абонент (клиент) определенного оператора мобильной связи. Местный абонент, вызывающий другого местного абонента в сети того же оператора мобильной связи, сгенерирует по крайней мере два внутренних события (одно инициирование вызова, одно получение вызова).
2. **Данные исходящего роуминга:** любое событие определения местоположения абонентом местного оператора мобильной связи, проводимое в другой сети (обычно это услуга роуминга иностранного мобильного оператора). Эти данные обычно представляют

местных абонентов, использующих мобильные телефоны во время поездок за границу.

3. **Данные входящего роуминга:** любое событие определения местоположения абонентом иностранного оператора мобильной связи. Эти данные обычно представляют иностранных абонентов, пользующихся услугами местного роуминга. Эти данные также могут включать внутренних абонентов от другого оператора мобильной связи, использующего услугу роуминга, поскольку у их собственного оператора мобильной связи отсутствует прием.

Идентификационные данные абонента

Наиболее подходящие идентификаторы для использования в статистике связаны с абонентами. Эти идентификаторы помогают различать абонентов и обычно долговечны по сравнению с идентификаторами, связанными с оборудованием (люди, как правило, довольно часто меняют свои мобильные телефоны по сравнению с их номером телефона или SIM-картой). Наиболее важными характеристиками, которые следует учитывать при выборе идентификаторов для целей статистики населения, являются:

- Идентификатор должен быть уникальным в течение периода анализа;
- Продолжительность действия идентификатора в идеале должна равняться выбранному периоду анализа;
- В идеале у каждого человека должен быть только один идентификатор (на самом деле люди могут носить с собой несколько устройств – одно лицо может называться несколькими разными абонентами).

Существует ряд различных идентификаторов, связанных с абонентами, которые генерируются и пересылаются различным сетевым объектам. Некоторые идентификаторы используются в масштабе всей системы, а некоторые имеют только локальное значение в течение ограниченного периода времени. У каждого из этих идентификаторов есть разные ограничения относительно их использования в статистических целях. Обычно IMSI (международный идентификатор мобильного абонента), MSISDN (номер цифровой сети с интеграцией служб) или идентификатор клиента/абонента, специфичный для оператора мобильной связи, служат постоянными ключами для идентификации пользователя в базах данных. Пока абонент не сменил свою SIM-карту, IMSI останется прежним, а до тех пор, пока абонент не изменил свой номер телефона, MSISDN останется прежним.

Идентификационные данные оборудования

Каждое мобильное устройство уникально идентифицируется либо по его IMEI

(международный идентификатор оборудования мобильной станции), либо по IMEISV (международный идентификатор оборудования мобильной станции и номер версии программного обеспечения). Эти идентификаторы не связаны постоянно с абонентом, поскольку устройства сохраняют свои IMEI и IMEISV даже после передачи другому пользователю. По стандартам 3GPP IMEI и IMEISV считаются временной информацией, которая может храниться в HLR, SGSN или VLR.

Временные характеристики

Когда CDR/IPDR генерируются для целей биллинга, время события является еще одним обязательным полем, которое необходимо включить в соответствии со стандартом 3GPP. Каждый временной атрибут должен включать как минимум дату, часы, минуты и секунды. Стандарт 3GPP определяет три различных типа временных меток обработки вызовов:

- Время занятости – это время занятости ресурсов для ответа на вызов абонента. Это поле обязательно только для неуспешных вызовов.
- Время ответа – время ответа на вызов при успешном соединении. Время ответа – обязательное поле для успешных вызовов.
- Время освобождения – время, когда занятые ресурсы вновь освобождаются. Это необязательное поле.

Атрибуты времени генерируются объектами, которые отвечают за предоставление определенных услуг. Обычно они не хранятся в объектах, таких как HLR или VLR, а доступны во время процесса генерации CDR с помощью определенных функций, запускаемых при наступлении платного события.

Характеристики местоположения

Чтобы добавить пространственный аспект к статистическому анализу, в котором используются данные мобильного позиционирования, крайне важно иметь географическую привязку событий, отраженных в таких данных. По стандартам 3GPP (ETSI TS 132 250) идентификаторы, связанные с местоположением, обычно являются обязательными (или должны быть включены при соблюдении определенных условий) в процессе генерирования информации о начислении платы. Роуминг — это один из случаев, когда добавление идентификатора местоположения не является обязательным по стандарту. По этой причине данные об исходящих вызовах за пределы сети могут не включать идентификаторы местоположения, кроме кода страны.

Для внутренних данных и входящего роуминга географической привязкой является местоположение и/или зона покрытия соты сети (первоначально идентификатор соты). Для данных исходящего роуминга исходной географической привязкой является страна партнера по роумингу оператора мобильной связи. В некоторых случаях данные о вызовах за пределы сети также включают идентификатор соты сети события исходящего роуминга.

Для систем сотовой связи самым маленьким структурным элементом является сота сети. Каждую соту можно описать рядом показателей, таких как азимут, угол сектора, форма и размер зоны покрытия, тип антенны и местоположение. Сеть мобильной связи может содержать тысячи отдельных сот. Поскольку каждая сота может обслуживать только ограниченное количество пользователей, они неравномерно распределены по зоне обслуживания оператора мобильной связи – более плотно в городских районах, и гораздо менее плотно в сельских районах.

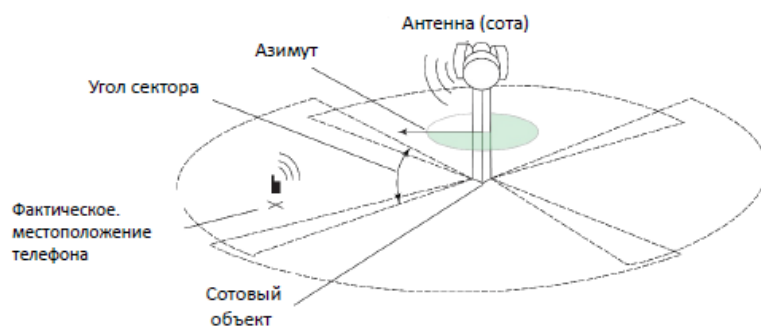


Рисунок 2.2 Соты, объект сотовой связи и зона местоположения и их отношения друг с другом

Вышка или мачта сотовой связи — это совокупность передатчиков в одном конкретном месте, которое можно назвать объектом сотовой связи (рисунок 2.2). Это означает, что в одном месте может иметься покрытие нескольких разных сот, когда мы сгруппируем их по координатам точек.

Соседние соты, обычно 30 или 40, сгруппированы в одну зону местоположения (LA⁴). Это делается для балансировки нагрузки сигнализации, поэтому устройству необходимо отправлять информацию о местоположении пользователя (обновления зоны местоположения) только тогда, когда происходит перемещение в другую зону

⁴ В сетях LTE используется термин «зона отслеживания» (TA) вместо термина «зона местоположения» (LA).

местоположения (LA)⁵, а также периодически с заранее заданными интервалами времени.

Идентификатор зоны местоположения и зоны маршрутизации

‘Зона местоположения’ – это термин, который относится к сети с коммутацией каналов (CS). В этой сети группа сот формирует зону местоположения (LA) или зону отслеживания (TA) для сетей LTE. Чтобы иметь возможность различать между несколькими LA/TA, назначается идентификатор зоны местоположения (LAI) / идентификатор зоны отслеживания (TAI). LAI/TAI состоит из трех элементов, где первые две части кода всегда одинаковы для оператора мобильной связи.

- Мобильный код страны (MCC) — это трехзначный идентификатор страны, в которой расположена зона местоположения (LA). MCC определяется Рекомендацией МСЭ-Т E. 212
- Код мобильной сети (MNC) — это двух- или трехзначный идентификационный код для оператора мобильной сети в соответствии с Рекомендацией МСЭ-Т E. 212
- Код зоны местоположения (LAC)/код зоны отслеживания (TAC) — это идентификатор зоны местоположения в сети оператора мобильной связи. Эта часть кода может быть представлена шестнадцатеричными значениями длиной два октета.

Например, LAI / TAI для компании Telia в Эстонии может быть 248010001, где:

- 248 – Мобильный код страны (MCC) для Эстонии;
- 01 – Мобильный код сети (MNC) для Telia;
- 0001 — LAC/TAC.

Что касается формата данных, то LAI/TAI может быть добавлен как одно поле (MCC+MNC+LAC/TAC) или как три поля (MCC; MNC; LAC/TAC). Если первые два поля отсутствуют, но источником данных является один конкретный мобильный оператор, то MCC и MNC могут быть получены из Рекомендации МСЭ-Т E.212. Поскольку LAC является уникальным только внутри пула зон местоположения одного конкретного оператора мобильной связи, наборы данных, содержащие данные нескольких операторов мобильной связи (например, исходящие), должны содержать коды MCC и MNC (или эквивалентные атрибуты, позволяющие различать операторов и страны).

Зона маршрутизации (RA) является аналогом LA в сетях с коммутацией пакетов. RA обычно имеет меньшую площадь по сравнению с LA, потому что использование мультимедийных услуг требует более частых пейджинговых сообщений. Уменьшение зоны, для которой необходимо передавать сообщения поискового вызова, помогает

⁵ Sauter, M. (2010) From GSM to LTE: An Introduction to Mobile Networks and Mobile Broadband (*От GSM к LTE: введение в мобильные сети и мобильный широкополосный доступ*). John Wiley & Sons, 414 стр.

снизить количество отправляемых пейджинговых сообщений. Зоны маршрутизации определяются при помощи идентификатора зоны маршрутизации (RAI), который представляет собой комбинацию ранее описанного LAI/TAI плюс код зоны маршрутизации (RAC), который представляет собой код длиной в один октет.

LAI, TAI и RAI являются временными данными абонента и сохраняются в VLR и SGSN, соответственно.

Идентификатор соты

Внутри каждой LA, TA и RA есть ряд сот, которые можно определить по идентификатору соты (CI, Cell ID). Этот идентификатор уникален только внутри зоны местоположения. По этим причинам был введен термин «глобальный идентификатор соты» (CGI). CGI — это комбинация LAI/TAI/RAI и CI, где CI имеет длину два октета и может быть закодирована с использованием шестнадцатеричного представления аналогично LA (рисунок 2.3). CGI — это уникальный идентификатор отдельной соты.

Как видно, для идентификации отдельной соты внутри сети требуется довольно длинный идентификационный код. По этой и нескольким другим причинам операторы мобильной связи могут создавать свои собственные коды идентификации соты.

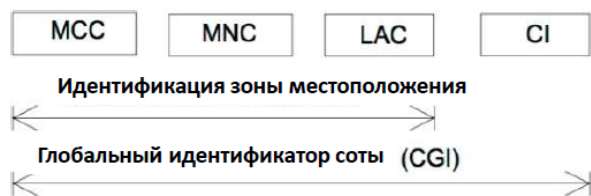


Рисунок 2.3 Сравнение идентификатора зоны местоположения (LAI) и глобального идентификатора соты (CGI).

Источник: ETSI TS 123 003⁶

Последний известный идентификатор соты временно сохраняется вместе с идентификатором абонента, например, в VLR и SGSN. Эти реестры используются системой-посредником для создания файлов CDR, которые будут передаваться в биллинговую систему и хранилище данных.

Уникальные идентификаторы соты (CGI или созданные оператором мобильной связи) обычно также хранятся в базе данных на уровне OSS с несколькими другими атрибутами, описывающими соту. Поскольку сеть мобильного оператора может часто меняться (например, при добавлении новых антенн, изменении направления для отдельных сот),

⁶ Sauter, M. (2010) From GSM to LTE: An Introduction to Mobile Networks and Mobile Broadband (*От GSM к LTE: введение в мобильные сети и мобильный широкополосный доступ*). John Wiley & Sons, 414 стр.

важно установить понимание между мобильным оператором и статистическим управлением относительно того, как часто должны передаваться такие данные. В идеале файлы информации о сотах следуют за данными о событиях.

Зона покрытия антенны

Обычно минимальная географическая информация, предоставляемая операторами мобильной связи, — это местоположение сетевых антенн (пара координат для точки антенны). Если географическая разбивка и точность выходных статистических показателей имеют низкое разрешение (карты на уровне страны, NUTS 3, LAU 1), достаточно географического представления сетевых антенн при помощи географических координат. Однако, если планируемая географическая разбивка должна быть более точной (LAU 2, LAU 3, муниципалитет, деревня, городские районы), важна информация о зонах покрытия антенн, поскольку зона покрытия антенн может охватывать несколько муниципалитетов (то есть антенны расположены в одном муниципалитете, но большая часть зоны покрытия охватывает другой муниципалитет). Поэтому важно, чтобы либо операторы мобильной связи также указывали географическую зону покрытия каждой антенны (как часть подготовки данных), либо теоретическая зона покрытия сот была бы рассчитана во время предварительной обработки данных (рисунок 2.4, рисунок 2.5).

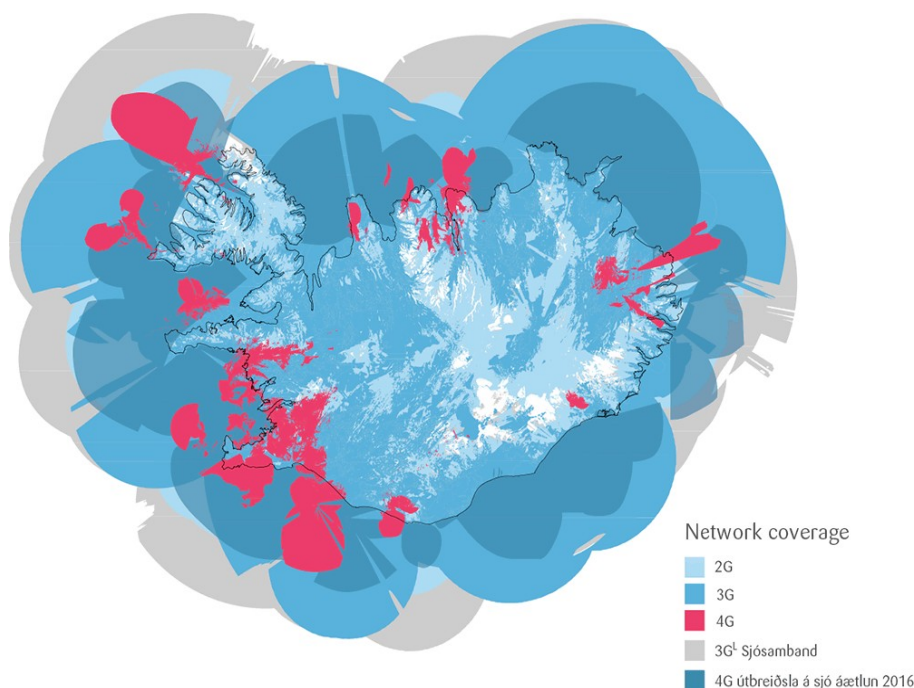


Рисунок 2.4 Иллюстрация зоны покрытия антенн сети оператора мобильной связи.
Источник: SIMINN⁷

⁷ SIMINN в Исландии — это оператор мобильной связи, который публикует онлайн карту покрытия сети: 33



Рисунок 2.5 Пример двух разных методов для расчета теоретической зоны покрытия антенны в сети: а) полигоны Вороного/Тиссена; б) теоретические секторы зоны покрытия

Дополнительные атрибуты данных о событиях

Список возможных атрибутов в этих спецификациях является исчерпывающим и, наряду с полезными характеристиками, содержит множество элементов, мало пригодных для обогащения данных для статистического анализа населения. Наиболее распространенные признаки, включение которых следует рассмотреть:

- Тип записи – помогает различать типы данных (CDR и IPDR), типы услуг (например, SMS, MMS, исходящий вызов и т. д.) и идентифицировать события, которые не генерируются абонентами (например, обновления местоположения);
- Продолжительность звонка или объем данных для анализа моделей использования мобильного телефона;
- Идентификатор абонента или оборудования для принимающих/отправляющих сторон.

Дополнительные атрибуты сетевых данных

Доступ к дополнительным параметрам для сети обычно осуществляется из OSS, где оператор мобильной сети хранит данные, важные для целей обслуживания сети. Для повышения точности определения местоположения событий большое значение имеют атрибуты сот.

Дополнительные атрибуты для абонентов

Операторы мобильной связи хранят информацию о своих клиентах в базах данных абонентов (для клиентов, совершающих внутренние и исходящие вызовы). Собираемая и хранимая информация сильно зависит от оператора сети, но обычно включает:

- Социально-демографические характеристики абонента (владельца контракта), например возраст, пол, предпочитаемый язык и т. д.
- Подробная информация о контракте и услугах:
 - частный или корпоративный клиент
 - адрес счета
 - средняя стоимость услуги
 - тип контракта (SIM-карта с предоплатой, SIM-карта с постоплатой, SIM-карта для межмашинного взаимодействия (M2M))

Эту информацию следует использовать с большой осторожностью, поскольку она очень чувствительна и не всегда точна (пользователь телефона может отличаться от держателя контракта). Ценность этих атрибутов заключается в возможности добавления дополнительных размерностей в статистический анализ (например, пол), а также в возможности выполнения процессов очистки данных (например, удаление данных M2M).

Социально-демографические характеристики подписчиков обычно собираются в системе управления взаимоотношениями с клиентами оператора мобильной связи и связаны с системой профилирования клиентов. Социально-демографические характеристики в основном собираются для частных (некоммерческих) абонентов с постоплатой, поскольку часто отсутствует информация о профилях клиентов с предварительной оплатой или информация ограничена (если только они не обязаны регистрироваться при покупке SIM-карты, но даже тогда информация о клиенте не всегда бывает в наличии). Для корпоративных контрактов, часто невозможно узнать, какой конкретный человек пользуется телефоном (и, следовательно, характеристики человека неизвестны). В случае с частными клиентами с оплатой по факту данные также часто искажены в случае семейных тарифов: социально-демографические данные человека, подписавшего контракт (отец, мать), распространяются на всю группу (то есть возраст и пол детей, пользующихся телефоном, принимают те же значения, что у отца/матери).

2.3. Общий процесс извлечения данных

Отдельные шаги процесса извлечения данных зависят от того, кто отвечает за обработку данных и где она выполняется. Если данные обрабатываются оператором мобильной связи или на территории оператора с использованием платформы, подобной сетевой песочнице (Sandbox-like platform), несколько этапов извлечения, таких как шифрование данных, можно пропустить.

Подготовка данных

Этап подготовки данных обычно включает создание скрипта для извлечения необходимых

данных из блока хранения. Это должно выполняться оператором мобильной связи, скорее всего, с использованием языка связи с базами данных, такого как SQL. Предпочтительно, чтобы скрипт извлекал данные автоматически через фиксированные интервалы, предварительно согласованные между оператором мобильной связи и статистическим управлением. Таким образом можно свести к минимуму задержки, которые могут быть вызваны человеческим фактором. Скрипт для подготовки данных должен учитывать такие характеристики, как тип данных, которые должны быть отправлены (например, CDR, IPDR или и то, и другое), период времени и атрибуты, которые необходимы, но также формат файла и необходимость анонимизации (для ясности об этом пойдет речь в отдельной главе). Скрипт для подготовки данных также включает в себя некоторые основные этапы обработки данных, которые оператор мобильной связи может предпринять для предоставления высококачественных данных. Эти шаги будут включать удаление нерепрезентативных данных или абонентов, занесенных в черный список по соображениям безопасности. Удаление нерепрезентативных данных будет включать исключение абонентов, которые не представляют собой реальных людей.

Как было сказано ранее, на этапе подготовки данных оператор мобильной связи должен назначить специалиста для написания этого скрипта и, в зависимости от исполнения, либо запускать его в определенное время вручную, либо убедиться, что автоматические процессы работают должным образом. Например, если есть задержки или данные отсутствуют, этот специалист должен уметь объяснить, почему возникли эти проблемы.

Анонимизация данных

Целью процесса анонимизации данных является защита конфиденциальности. Во время этого процесса может быть изменен личный идентификационный код абонента, или данные могут быть агрегированы для обеспечения анонимности субъектов. Хотя есть несколько способов сделать это, в настоящее время не существует хорошего метода для сохранения всех аспектов данных, необходимых для более длительных периодов исследования, при одновременном обеспечении их анонимности в требуемой степени. В то время как получение данных, в которых подписчик идентифицируется с помощью IMSI, было бы идеальным случаем, различия в законодательстве и политике операторов мобильной связи ограничивают методы, которые фактически могут быть использованы. Иногда операторы мобильной связи предпочитают вычислять «хэш» значение идентификатора и, таким образом, маскировать его для прямой идентификации (псевдонимизация). Если возможно, это хэширование должно позволять отслеживать идентификатор в течение длительного периода времени, но в таблице 2.2 перечислены

ограничения более коротких частот обновления.

Наряду с представленными вариантами, выборка (уменьшение возможности включения человека в набор данных) и обфускация (маскирование / скрытие исходных данных) могут использоваться в качестве альтернативы или применяться для дальнейшего повышения уровня защиты конфиденциальности.

При рассмотрении этих вариантов следует подробно обсудить масштаб проекта, в котором будут использоваться данные, чтобы понять, какой метод анонимизации будет приемлем и для статистического управления. Анонимизация (если имеется) должна быть частью скрипта подготовки данных. Если будут использоваться псевдонимные уникальные коды, оператору мобильной связи необходимо создать и поддерживать хэш-функции, чтобы гарантировать те же пары ключей и значений в течение согласованного периода.

Таблица 2.2 Ограничения, которые накладывает анонимизация на полезность набора данных для статистики туризма⁸

Уровень анонимизации идентификатора абонента	Уровень защиты персонально идентифицируемой информации (анонимность)	Возможности для статистики								
		Объединение идентификатора с др. операторами	Распознавание активности во внутреннем пространстве	Обычная среда	Идентификация поездки	Посещения в тот же день / с ночевкой	Приграничные искажения	Долгосрочные посетители	Транзитные посещения	Повторные посещения
Постоянный уникальный код (IMSI)	Не защищено совсем									
Постоянный уникальный код псевдоним	Очень слабо защищено. Псевдонимизация, но не анонимизация. Косвенная идентификация возможна									
Временный уникальный код-псевдоним на одну неделю	Слабо защищено. Косвенная идентификация возможна в большинстве случаев									
Временный уникальный код-псевдоним на 24 часа	Очень хорошо защищено. Косвенная идентификация возможна, но в редких случаях									
Агрегированные первичные данные	Идентификационные данные абонентов полностью защищены, если также используется порог (например, не отображаются агрегаты ниже 10).									
	Хороший уровень детализации									
	Ограниченный уровень детализации									
	Слишком ограничено									

⁸ Eurostat (2014) Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics (Евростат (2014) (Исследование возможности использования данных мобильного позиционирования для статистики туризма) (Евростат, 2014 г.)

Шифрование данных

Шифрование данных — это небольшой, но важный шаг, который необходим, когда обработка данных происходит за пределами территории оператора мобильной связи. Его цель — гарантировать, что неуполномоченные стороны не смогут прочитать данные в случае нарушения безопасности во время передачи данных.

Во время шифрования данных генерируется ключ, который определяет, как данные будут зашифрованы в зашифрованный текст. В случае симметричного шифрования тот же ключ будет использоваться для дешифрования данных. Для асимметричного шифрования создается пара ключей, в которой открытый ключ будет использоваться для шифрования данных, а закрытый ключ — для их дешифрования. Обычно рекомендуется использовать асимметричное шифрование, если обмен ключами будет осуществляться через Интернет.

Существует несколько криптографических библиотек, которые можно использовать для шифрования и дешифрования данных. Примеры включают OpenSSL и GNU Privacy Guard, которые могут использоваться в коммерческих и некоммерческих целях.

Передача данных

Во время передачи данные будут доступны для статистического управления либо для внутренней обработки, либо для обработки вне помещений оператора мобильной связи (централизованно). В первом случае данные обычно переносятся в базу данных, предназначенную для этой цели, или становятся доступными в форме файла данных на сервере, к которому может иметь доступ принимающая сторона. Если используется централизованная обработка данных, самыми простыми способами передачи данных являются:

- Оператор мобильной связи выделит пространство на сервере для извлеченных данных и сделает его доступным для принимающей стороны. Принимающая сторона подключится к серверу и транспортирует файлы данных в свое выделенное пространство на сервере.
- Принимающая сторона выделит пространство на сервере для извлеченных данных и сделает его доступным для оператора мобильной связи. Оператор подключится к серверу и передаст файлы данных.

Как видно из вышеизложенного, в любом случае обеим сторонам потребуется выделить некоторое пространство на сервере для передачи данных.

Архивирование данных

В случае случайной потери данных из-за технических проблем или по любой другой непредвиденной причине важно, чтобы исторические данные можно было повторно получить и использовать для пересчета. Для операторов мобильной связи это просто означает повторное извлечение исторических (архивных) данных. Однако, если данные передаются на внешнее средство обработки после извлечения, все извлеченные необработанные данные должны храниться в выделенном блоке хранения, который не подключен к тому же блоку, где происходит обработка данных. Эти архивы должны соответствовать законодательству, касающемуся сохранения данных, и храниться только в течение периода времени, разрешенного нормативными актами.

Логический порядок шагов в процессе извлечения данных

1. Где можно обрабатывать данные? У оператора мобильной связи или вовне.
2. Каково происхождение данных? Зондирование или база данных.
3. Какие источники данных доступны? Входящие, исходящие, внутренние, CRM, географическая справочная база данных.
4. Каков охват типов данных? CDR, IPDR, прочее.
5. Каковы поколения системы сотовой связи? 2G, 3G, 4G.
6. Какие типы записей имеются? Вызов, SMS, MMS, трафик данных, обновления зоны местоположения, другое.
7. Каков охват абонентов? Все или выборка.
8. Каков охват пользователей по типам оплаты? Предоплата или постоплата.
9. Каков охват данных по типу биллинговой системы? Оффлайн или онлайн биллинг.
10. Какие атрибуты доступны в данных? Идентификатор абонента, местоположение, время, другие атрибуты.
11. Какие форматы идентификаторов используются?
12. Как будут обрабатываться ссылки на страны (входящие и исходящие данные)? MCC из идентификатора абонента или отдельное поле.
13. Каков метод защиты конфиденциальности? Агрегирование данных, уникальный псевдонимный идентификатор, выборка, обфускация, без анонимизации.
14. Как данные будут организованы в файлы?
15. Каковы спецификации элементов файла данных? Имена атрибутов, типы, имена файлов, тип файла.
16. Каковы спецификации скрипта для извлечения данных? Удаление нерепрезентативных данных (M2M / IoT, дубликаты), удаление абонентов,

включенных в «черный список».

17. Как будет осуществляться шифрование? Программное обеспечение для шифрования или обмен ключами.
18. Как часто можно извлекать данные из систем хранения? Каждый день, месяц, другое.
19. Как следует передавать данные? В соответствии с технологиями PUSH или PULL.
20. Где следует архивировать необработанные данные? Используя мощности мобильного оператора или статистического управления, либо отсутствие архивирования.

3. Доступ к данным мобильных телефонов и модели партнерства

3.1. Введение

В настоящее время данные мобильных телефонов уже имеют более широкие возможности для использования в развивающихся и наименее развитых странах по сравнению с другими хорошо известными источниками больших данных, такими как социальные сети, транзакции электронной торговли и другие виды деятельности в Интернете (рис. 3.1). Их широкая доступность делает их желательным источником для производства статистических данных, сопоставимых на глобальном уровне. Однако по ряду причин считается, что получение доступа к данным мобильных телефонов для использования в официальной статистике представляет собой большие сложности.



Рисунок 3.1 Области использования больших данных (BIG DATA) для целей официальной статистики

Во-первых, существуют пробелы в потенциале: различные заинтересованные стороны,

такие как национальные статистические управления, регулирующие органы электросвязи, министерства ИКТ, органы по защите конфиденциальности данных и операторы телекоммуникационных сетей, могут обладать разными полномочиями и разными уровнями технических возможностей и ресурсов. Несовпадение полномочий, ресурсов и опыта приводит к нехватке мощностей, что потенциально затрудняет доступ к данным мобильных телефонов. Государственные учреждения, отвечающие за сбор и публикацию официальной статистики на основе данных мобильных телефонов, могут не иметь вычислительной инфраструктуры и технического персонала для передачи, хранения, обработки и анализа больших данных, особенно в развивающихся странах. Обработка больших объемов данных, как ожидается, входит в сферу деятельности национальных статистических управлений, но они могут не обладать специфическими знаниями в области данных мобильных телефонов и не иметь постоянного контакта с операторами телекоммуникационных систем. Министерства ИКТ и регулирующие органы в области электросвязи, как правило, имеют больше возможностей для переговоров о доступе к данным мобильных телефонов, но у них может не быть технических знаний для облегчения передачи и хранения данных. Скорее всего, именно операторы связи, которые уже собирают, хранят, обрабатывают и анализируют данные мобильных телефонов в рамках своих повседневных деловых операций, обладают инфраструктурой и ноу-хау для работы с большими данными. Эти пробелы в потенциале существуют в различных формах и в различной степени как в развитых, так и в развивающихся странах, но в последних могут возникнуть более серьезные проблемы.

Во-вторых, существует пробел в отношении обеспечения конфиденциальности, возникающий из-за недостатков в правовой базе, регулирующей обработку, использование и передачу данных. Во многих странах законы и нормативные акты, касающиеся кибербезопасности, конфиденциальности данных или защиты данных, либо отсутствуют, либо не соблюдаются надлежащим образом. Отсутствие надлежащих правовых гарантий делает использование данных мобильных телефонов для официальной статистики более рискованным; затронутые стороны, будь то индивидуальные клиенты или предприятия, не будут защищены законом в случае намеренных или ненамеренных вторжений. Справедливо или ошибочно операторы сетей связи могут ссылаться на повышенный риск как на причину отказа в доступе к данным мобильных телефонов для целей, выходящих за рамки их собственных деловых операций. Тем не менее, введение в силу законов о кибербезопасности, конфиденциальности данных или защите данных не обязательно гарантирует быстрый и легкий доступ к данным мобильных телефонов для

официальной статистики. Скорее, основная роль этих законов состоит в том, чтобы снизить риски, связанные с использованием данных, и предоставить гарантии заинтересованным сторонам.

Налаживание партнерских отношений имеет решающее значение для устранения пробелов в мощностях и обеспечении конфиденциальности, которые препятствуют доступу к данным мобильных телефонов для использования в официальной статистике. Данные мобильных телефонов собираются, хранятся и обрабатываются операторами связи, основной интерес которых заключается в использовании данных в коммерческих целях. Государственным учреждениям, заинтересованным в использовании данных мобильных телефонов для производства официальной статистики по различным аспектам национальной жизни, будь то туризм, образование, миграция, бедность и т. д., потребуется сотрудничество с операторами систем связи и другими сетевыми провайдерами. Такому сотрудничеству можно способствовать, в первую очередь, путем создания правовой основы для операторов и других сетевых провайдеров, позволяющей предоставлять доступ к данным безопасным способом и с уважением права на неприкосновенность частной жизни отдельных лиц. В идеале это означает, что правила по защите конфиденциальности личных данных должны быть сначала законодательно закреплены и их исполнение обеспечено, если этого не было с самого начала. Для облегчения сотрудничества следует работать в тесном контакте с регулирующими органами в области телекоммуникаций и органами по обеспечению конфиденциальности данных. Во многих странах регулирующие органы в области электросвязи наделены полномочиями регулярно запрашивать данные у операторов сетей в ходе своей нормативно-правовой деятельности и, следовательно, имеют хорошие возможности для получения доступа к данным мобильных телефонов. С другой стороны, органы по обеспечению конфиденциальности данных там, где они существуют, в основном несут ответственность за обеспечение соблюдения национальных законов о конфиденциальности или защите данных. Они также являются важными партнерами в обеспечении доверия заинтересованных сторон к использованию данных мобильных телефонов для официальной статистики. Правительственные агентства и национальные статистические управления получают большие преимущества от тесного партнерства с регулирующими органами в области электросвязи и органами по защите данных в их стремлении использовать данные мобильных телефонов.

В этой главе мы приводим самые существенные аспекты построения партнерских отношений для получения доступа к данным мобильных телефонов для производства

официальной статистики. Мы подробно описываем основные способы формирования этих партнерских отношений, проблемы, которые необходимо решить, и роли четырех основных заинтересованных сторон: национальных статистических управлений, регулирующих органов электросвязи, операторов электросвязи/сетевых провайдеров и органов по обеспечению конфиденциальности данных. Мы также приводим примеры такого партнерства, используя опыт Международного союза электросвязи (МСЭ), специализированного агентства ООН по информационным и коммуникационным технологиям. МСЭ активно изучает способы использования больших данных из отрасли ИКТ для улучшения и дополнения существующих статистических данных и методологий измерения информационного общества с помощью пилотных проектов в различных странах. Большая часть обсуждений в этой главе основана на опыте МСЭ по осуществлению пилотных проектов в странах.

3.2. Обеспечение среды для доступа к данным мобильных телефонов для официальной статистики

Партнерские отношения, независимо от их формы, с большей вероятностью будут активно развиваться при наличии благоприятных институтов и поддерживающего законодательства. Для обеспечения продуктивного партнерства между ключевыми заинтересованными сторонами также необходимо адаптировать модель партнерства так, чтобы каждая заинтересованная сторона играла роль, дополняющую друг друга с точки зрения возможностей и полномочий. В случае операторов телекоммуникационных систем крайне важно понимать их рыночные стимулы, и, когда это возможно, использовать их для мотивации сотрудничества. Часто упускаемый из виду, но не менее важный компонент продуктивного партнерства, — это поддержка со стороны органа по защите конфиденциальности данных, если он существует. Партнерства по обеспечению доступа к данным мобильных телефонов для официальной статистики с большей вероятностью достигнут своих целей без непредвиденных задержек (и затрат), когда на ранних этапах партнерства имеется вовлеченность и участие органа по защите данных. Если органы по защите конфиденциальности данных участвуют в работе с самого начала, они могут помочь создать атмосферу доверия между различными заинтересованными сторонами и вселить уверенность в безопасном, надежном и социально выгодном использовании данных мобильных телефонов. В следующих разделах мы развиваем эти идеи, описывая различные типы моделей партнерства и обсуждая роли, возможности, полномочия и мотивы четырех основных заинтересованных сторон. Мы также исследуем проблемы использования данных мобильных телефонов в обычном процессе производства

официальной статистики, особенно те, которые связаны с конфиденциальностью и безопасностью данных мобильных телефонов, а также с их постоянной доступностью и пригодностью для статистических целей.

Модели партнерств для использования данных мобильных телефонов для целей официальной статистики

Партнерские отношения для обеспечения доступа к данным мобильных телефонов для официальной статистики могут принимать различные формы. Не существует универсальной модели партнерства; у каждой есть свои сильные и слабые стороны. Партнерства могут различаться с точки зрения управления, финансирования, целей и распределения задач по обработке данных. Когда цели являются долгосрочными и определены широко, то есть набор официальных статистических данных характеризует большой круг секторов и будет периодически извлекаться из данных мобильных телефонов в течение многих лет в будущем, национальные статистические управления, вероятно, будут руководить процессом. В партнерствах, цели которых определены узко и сосредоточены на формировании официальной статистики по конкретным секторам, государственное учреждение, ответственное за сектор, также может взять на себя руководящую роль и финансировать большую часть деятельности. Например, когда целью является производство официальной статистики для сектора электросвязи и ИКТ, регуляторы электросвязи и министерства ИКТ часто берут на себя инициативу по налаживанию партнерских отношений, финансируют большую часть деятельности и управляют результатами.

Финансирование, конечно же, является важнейшим элементом эффективных партнерских отношений. Существует множество различных подходов к финансированию деятельности, направленной на использование данных мобильных телефонов для целей официальной статистики, каждый из которых имеет свой набор проблем и преимуществ. Опять же, не существует единого правильного подхода для всех обстоятельств в разных странах. Независимо от выбранного подхода, в начале создания партнерства рекомендуется составить список и предусмотреть все прямые и косвенные затраты, с которыми может столкнуться каждая заинтересованная сторона на протяжении всего периода своего участия. Необходимо учитывать не только новые капитальные затраты и административные расходы, но и альтернативные издержки времени существующего персонала и оборудования, которые будут участвовать в партнерской деятельности. Во многих моделях партнерства ответственность за финансирование деятельности неизбежно будет разделяться между всеми заинтересованными сторонами, будь то в денежной или натуральной форме. Хотя

партнерство может иметь доступ к выделенным средствам из государственного бюджета или из других источников, которые будут использоваться для капитальных затрат или для найма узких специалистов, каждая заинтересованная сторона неизбежно будет вынуждена отвлекать время и ресурсы своего персонала на работу в сфере деятельности партнерства. Признание таких альтернативных издержек с самого начала очень поможет заинтересованным сторонам в планировании своих результатов и эффективном управлении ими.

Однако независимо от того, кто берет на себя руководство или финансирует большую часть деятельности, и от того, каковы статистические цели, ключевой отличительной чертой модели партнерства является распределение ответственности за обработку данных. В частности, какая из заинтересованных сторон является ответственной за обработку первичных данных, например: подробных записей о вызовах (CDR) или подробных записей интернет-протокола (IPDR), и получение статистических показателей? Где происходит обработка данных? Основываясь на распределении задач по обработке данных, мы можем выделить два основных типа моделей партнерства: а) за обработку данных, в основном, отвечают операторы телекоммуникационных сетей, или б) за обработку данных, в основном, отвечают государственные ведомства / не операторы. Выбор распределения задач обработки данных повлияет на распределение затрат, а также на необходимые меры для обеспечения безопасности, конфиденциальности и защиты данных. Чтобы лучше понять различия между этими двумя моделями партнерства, нам сначала необходимо определить разные уровни данных, соответствующие различным этапам обработки данных, от извлечения необработанных данных до расчета показателей, пригодных для публикации.

1. **Уровень I** - данные состоят из исходных, необработанных и неагрегированных данных, содержащих конфиденциальную коммерческую информацию и информацию, позволяющую идентифицировать личность. Структура первичных данных уровня I будет зависеть от конкретной базы данных, из которой они были извлечены, и будет различаться для разных операторов мобильной связи. Данные уровня I от разных операторов еще не готовы к объединению. Их необходимо будет дополнительно очистить, отформатировать и объединить в данные уровня II.

2. **Уровень II** - данные представляют собой первоначально агрегированные данные без личной информации и с некоторой деловой конфиденциальной информацией. Данные уровня II уже подготовлены таким образом, что данные от разных операторов готовы к объединению.

3. **Уровень III** - данные состоят из агрегированных показателей, которые могут быть опубликованы, и больше не содержат никакой личной или деловой конфиденциальной информации.

В **модели партнерства типа А**, как показано на рисунке 3.2, большая часть обработки данных выполняется операторами мобильной связи на их мощностях. Операторы несут основную ответственность за извлечение необработанных данных и их обработку до уровня II. Затем данные уровня II от разных операторов объединяются и обрабатываются далее в данные уровня III за пределами территории оператора ответственным государственным ведомством или консорциумом государственных ведомств и неправительственных организаций. В некоторых случаях, когда в партнерстве участвуют международные организации или иностранные субъекты (и это зависит от требований в отношении суверенитета данных), обработка данных уровня II - III может происходить даже за пределами национальных границ.

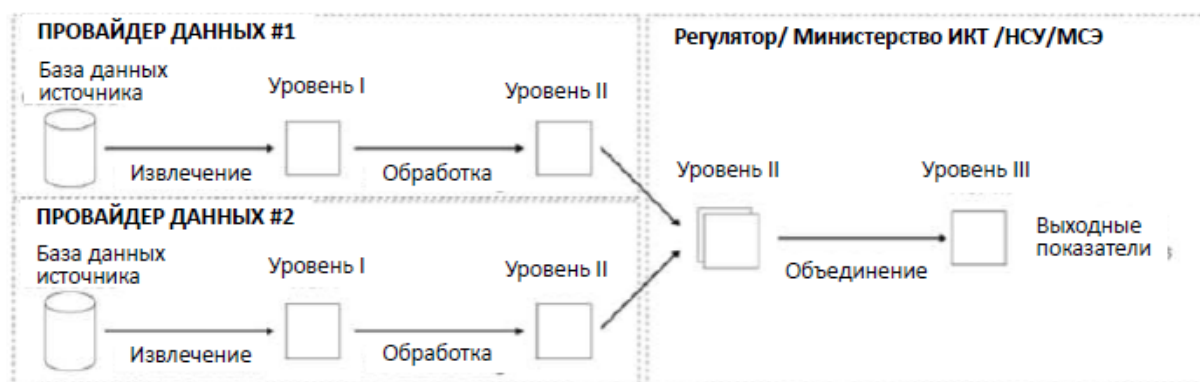


Рисунок 3.2 Модель партнерства типа А – Операторы мобильной связи в основном ответственны на обработку данных ⁹

Основным преимуществом модели партнерства типа А является то, что она дает провайдером данных (в данном случае, операторам мобильной связи) большую уверенность в том, что деловая конфиденциальная и личная информация не будет случайно раскрыта посторонним третьим лицам. Учитывая, что большая часть обработки данных выполняется их собственным персоналом в их помещениях, операторы будут иметь больший контроль над рисками нарушения конфиденциальности и безопасности данных. По этой причине они могут предпочесть модель партнерства типа А. Следовательно, с помощью этой модели может быть проще обеспечить участие операторов мобильной связи. Обратной стороной для операторов является то, что им придется нести большие расходы и уделять больше

⁹ Этот и следующий рисунок взяты из материалов МСЭ по проекту «Большие данные для измерения информационного общества»: <https://www.itu.int/en/ITU-D/Statistics/Pages/bigdata/default.aspx>

времени своего персонала и ресурсов оборудования задачам обработки данных. Однако, поскольку операторы чаще всего уже занимаются обработкой больших данных в рамках своих обычных деловых операций, их предельные затраты, вероятно, будут ниже по сравнению с затратами государственных ведомств, которым, возможно, потребуется осуществить новые вложения в вычислительную инфраструктуру и повышение технической квалификации персонала.

Что касается статистических результатов, то есть достоверности и точности вычисленных показателей, модель партнерства типа А создаст дополнительные трудности для верификации данных. Прежде чем показатели, рассчитанные на основе больших данных, можно будет принять в качестве официальной статистики, необходимо будет проверить скрипты и алгоритмы, используемые для извлечения первичных данных и преобразования их в данные уровня II. Это означает, что государственное ведомство, получающее данные уровня II, должно будет обладать необходимыми техническими знаниями для выполнения таких проверок и оценки полноты и точности первоначально агрегированных данных. Для модели партнерства типа А особенно важно будет внедрить механизмы, гарантирующие, что процедуры обработки данных, от извлечения до уровня II, являются прозрачными, документированными и воспроизводимыми.

Специалисты по обработке и анализу данных, не работающие на оператора сети, могут получить доступ только к агрегированным данным уровня II в рамках модели типа А, так что степень детализации данных недостаточна для выполнения подробного анализа на основе определения местоположения. Например, CDR выявляют закономерности в перемещениях населения, что полезно для прогнозирования вспышек опустошительных эпидемий в определенных местах. Однако эта потенциально жизненно важная информация теряется при агрегировании данных.

В **модели партнерства типа В**, как показано на рисунке 3.3, большая часть обработки данных выполняется государственными органами вне помещений операторов мобильной связи. Операторы несут ответственность за извлечение первичных данных и передачу их в качестве данных уровня I ответственным государственным ведомствам. Государственное ведомство (или консорциум государственных и неправительственных организаций, в зависимости от обстоятельств) будет нести ответственность за очистку и агрегирование данных уровня I до данных уровня II, объединение данных уровня II от нескольких операторов и, наконец, расчет агрегированных показателей, содержащихся в данных уровня III. В некоторых случаях, когда в партнерстве участвуют международные организации или иностранные субъекты (и это соответствует требованиям в отношении

суверенитета данных), обработка данных уровня I до уровня III может происходить даже за пределами национальных границ. Однако, учитывая потенциально чувствительную и частную информацию, содержащуюся в данных уровня I, и, в некоторой степени, в данных уровня II, обработка за пределами национальных границ потребует особой осторожности в отношении редактирования, анонимизации и безопасности.

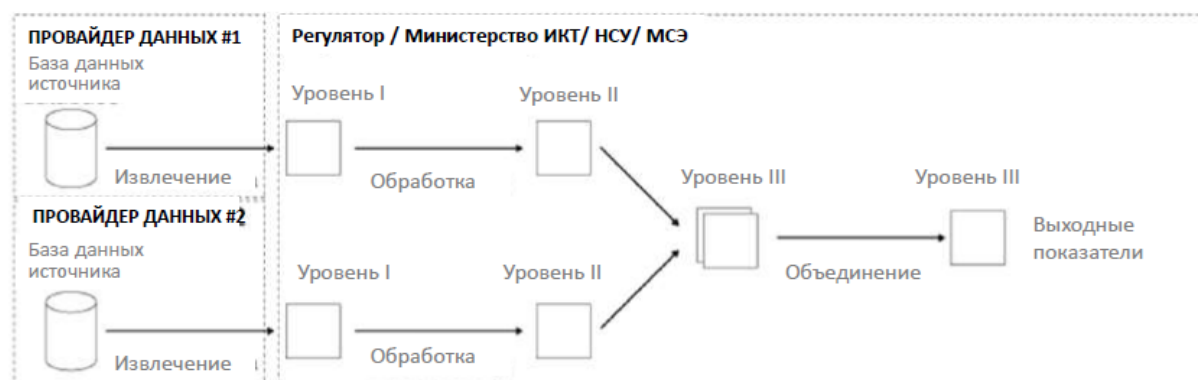


Рисунок 3.3 Модель партнерства типа В – Государственные ведомства ответственны за обработку данных

С точки зрения обеспечения достоверности и непрерывности получения официальной статистики на основе данных мобильных телефонов, модель партнерства типа В имеет множество преимуществ. Во-первых, государственные ведомства, публикующие официальную статистику на основе данных мобильных телефонов, будут иметь больше возможностей для проверки и оценки правильности расчета агрегированных показателей. Поскольку они сами обрабатывают данные, они смогут не только исследовать содержание алгоритмов и скриптов, чтобы определить, соответствуют ли они предписанным методологиям, но они также могут непосредственно повторно отслеживать, как выполнялись скрипты с данными для расчета показателей. Учитывая, что большая часть обработки данных осуществляется заинтересованными государственными организациями, также потенциально может быть меньше сбоев в производстве официальной статистики, когда операторы мобильной связи входят на рынок и выходят с него. Поскольку функция операторов состоит только в извлечении необработанных данных, нет необходимости ориентировать новых участников рынка на использование предписанной методологии, кроме указания необходимых полей данных.

Модель партнерства типа В потребует от заинтересованных государственных организаций иметь инфраструктуру и технические знания для обработки больших данных, а в противном случае – тратить значительно больше на создание вычислительной инфраструктуры и на приобретение знаний и опыта. Необходимые дополнительные

расходы со стороны государственных организаций не обязательно являются недостатком. Эти затраты, понесенные стратегически и с расчетом на будущее, можно рассматривать как долгосрочные государственные инвестиции, которые могут принести выгоды, выходящие за рамки их непосредственных целей. Например, партнерства типа В, изначально нацеленные на производство официальной статистики для конкретных секторов, могут распространить тот же подход на другие секторы, используя недавно приобретенную инфраструктуру и квалификацию персонала. Распространение на другие секторы может быть осуществлено легче, если задачи обработки данных, инфраструктура и экспертные знания будут сосредоточены в национальных статистических управлениях, поскольку они являются государственным органом, обычно уполномоченным собирать данные и производить официальную статистику по широкому кругу вопросов в экономической и социальной сферах.

Независимо от правовой юрисдикции и применяемой парадигмы конфиденциальности, модель партнерства типа В создаст дополнительную проблему с точки зрения соблюдения требований конфиденциальности данных, поскольку она требует передачи записей данных мобильных телефонов, содержащих личную информацию, от одной организации к другой. В большинстве юрисдикций передача таких данных является чувствительным и строго регулируемым действием. Это особенно актуально для стран с европейскими правилами защиты данных, в которых объем регулируемых данных и регулируемых операций по обработке определен более широко.

Согласно европейским правилам, предприятиям и другим организациям (включая государственные учреждения), как правило, запрещается обрабатывать персональные данные, где под персональными данными в широком смысле понимается любая информация, относящаяся к идентифицируемому человеку, независимо от того, позволяют ли сами данные идентифицировать субъект данных¹⁰. Точно так же обработка данных также определяется широко, и включает любую деятельность, связанную с персональными данными, от сбора до уничтожения и всего, что между этим. Обработка разрешена только в том случае, если субъекты данных дают свое согласие или когда применяются установленные законом исключения (например, в случаях, представляющих государственный интерес). При таких режимах конфиденциальности данных заинтересованным сторонам в модели партнерства типа В, возможно, потребуется добиваться исключений. Если исключения не предоставляются, заинтересованным сторонам потребуется предпринять дополнительные шаги для редактирования данных

¹⁰ Агентство Европейского Союза по основным правам, 2013 г.

уровня I, чтобы они больше не являлись персональными данными. В зависимости от правил конкретной страны редактирование может включать анонимизацию или агрегирование, а не только псевдонимизацию. Как минимум, необходимо будет принять меры для обеспечения безопасной передачи и хранения данных уровня I.

В большинстве случаев предпочтительным режимом работы, на самом деле, является гибридный режим - тип АВ, при котором данные остаются в помещениях оператора мобильной связи, но полностью доступны для назначенных органами государственного управления специалистов по обработке и анализу данных. Этот подход обычно используется в телекоммуникационном секторе, где операторы мобильной связи предоставляют доступ к данным сторонним лицам либо физически на месте, либо через удаленное безопасное соединение. Такой режим доступа использовался для многих успешных проектов, связанных с данными мобильного позиционирования, особенно в тех случаях, когда данные использовались в официальной статистике (см. Пример Индонезии в Приложении). Сравнение типов партнерств можно найти в таблице 3.1.

Таблица 3.1 Сравнение моделей партнерства для получения доступа к данным мобильных телефонов для целей официальной статистики

Характеристики	Тип А	Тип В	Тип АВ
Местоположение данных	Мобильный оператор	Орган государственного управления	Мобильный оператор
Ответственный за обработку данных	Мобильный оператор	Орган государственного управления	Орган государственного управления
Детализация данных	Низкая детализация	Высокая детализация	Высокая детализация
Восприятие риска раскрытия персональных данных и нарушения конфиденциальности коммерческих данных, возникающего при использовании данных для официальной статистики	Низкий риск	Высокий риск	Средний риск
Диапазон потенциальных статистических показателей, которые могут быть получены	Заранее определенные статистические показатели	Широкий диапазон возможных статистических показателей	Широкий диапазон возможных статистических показателей
Простота проверки и уровень контроля качества статистики	Низкий уровень контроля	Высокий уровень контроля	Высокий уровень контроля
Влияние на вход/выход новых операторов на рынок на затраты на производство официальной статистики	Высокие затраты при добавлении операторов	Низкие затраты при добавлении новых операторов	Средние затраты при добавлении операторов

Проект МСЭ по использованию больших данных для измерения информационного общества

МСЭ предпринял пилотный проект по использованию больших данных из отрасли телекоммуникаций для улучшения и дополнения существующих статистических данных и методологий измерения информационного общества. Ожидается, что результаты этого проекта помогут странам и МСЭ в производстве официальной статистики по ИКТ и разработке новых методологий, сочетающих новые и существующие источники данных¹¹. Следующие шесть стран участвовали в пилотном проекте: Колумбия, Грузия, Кения, Филиппины, Швеция и Объединенные Арабские Эмираты. Доступ к данным уже был успешно осуществлен через модель партнерства типа А или типа В, как показано в таблице 3.2 (тип АВ не использовался).

Таблица 3.2 Доступ к данным мобильных телефонов в шести пилотных странах: Проект МСЭ по использованию больших данных для измерения информационного общества

Страна	Модель партнерства	Ответственный за обработку данных
Колумбия	В	Национальное статистическое управление
Грузия	В	Регулятор телекоммуникаций
Кения	А	Операторы сетей
Филиппины	А	Операторы сетей
Швеция	В	МСЭ
ОАЭ	А	Операторы сетей

Было извлечено несколько уроков в отношении доступа к данным и подготовки поля для проекта по данным мобильного позиционирования.

1. Во-первых, необходимо подготовить все административные и юридические процедуры для доступа к данным до начала проекта. Это означает, что между заинтересованными сторонами существует соглашение о модели обработки для расчета (поставщиками данных или регулирующий орган/ НСУ). Независимо от того, где происходит обработка, метод передачи данных также должен быть согласован заранее – см. Раздел 2.3, где приведен логический порядок тем для обсуждения до того, как

¹¹ Последнюю информацию по проекту можно найти на вебсайте МСЭ «Большие данные для измерения информационного общества»: <https://www.itu.int/en/ITU-D/Statistics/Pages/bigdata/default.aspx>

передача данных может произойти.

2. Во-вторых, методика должна быть стандартной, четкой и однозначной. Сюда входит подробное описание источника данных (входные данные, см. Раздел 2.3). Должна быть четкая концепция методологии расчета, которая может быть основана на опыте других стран. Группам, занимающимся обработкой данных, также требуются примеры алгоритмов расчета.

3. В-третьих, для успеха проекта должна быть обеспечена инфраструктура и кадровые ресурсы для обработки данных (специалисты по обработке и анализу данных, инженеры по работе с данными и с системами, а также менеджеры проектов).

Кроме того, проект также показал, что должна постоянно поддерживаться координация со всеми заинтересованными сторонами (по доступу к данным, валидации результатов, анализу), что будет рассмотрено далее.

Заинтересованные стороны: роли, возможности и полномочия

Эффективное и устойчивое партнерство потребует вовлеченности и взятия на себя обязательств со стороны четырех групп ключевых заинтересованных сторон: национальных статистических управлений и/или отраслевых министерств, регулирующих органов электросвязи, операторов мобильной связи и других сетевых провайдеров, а также органов по защите конфиденциальности данных (там, где они существуют). Заинтересованные стороны имеют взаимодополняющие роли в зависимости от их полномочий или организационных целей и их технических возможностей. Что касается операторов, то рынок и нормативно-правовая среда будут влиять на их стимулы к обмену данными. Государственным ведомствам рекомендуется твердо понимать, что это за стимулы, и, когда это возможно, формировать эти стимулы с помощью нормативно-правового регулирования, чтобы обеспечить сотрудничество всех потенциальных поставщиков данных. Ниже обсуждается потенциальная роль каждой группы заинтересованных сторон в зависимости от спектра их полномочий и ответственности или целей, а также их технических возможностей.

Национальные статистические управления (НСУ): в эпоху больших данных официальная статистика находится на перепутье, как и НСУ, которые являются ее основными производителями¹². Ряд НСУ во всем мире сейчас осознают потенциал больших данных для повышения своевременности и экономической эффективности производства официальной статистики, а также для предоставления новых статистических

¹² Florescu, Karlberg, Reis, Castillo, Skaliotis and Wirthmann (2014) Will 'big data' transform official statistics? *(Изменяют ли большие данные официальную статистику?)*

продуктов и услуг. Согласно последнему обзору Глобальной рабочей группы Организации Объединенных Наций по большим данным для официальной статистики, в настоящее время существует более 180 проектов как в развитых, так и в развивающихся странах, которые исследуют или экспериментально используют большие данные для целей производства официальной статистики. Многие НСУ уже имеют хорошие возможности для использования источников больших данных, поскольку они обычно обладают юридическими полномочиями для востребования информации от поставщиков данных и для интеграции конфиденциальных данных¹³. У них также есть опыт и знания в области сбора и обработки больших объемов данных. И наконец, что не менее важно, НСУ обладают уникальной квалификацией для оценки качества и репрезентативности источников больших данных, поскольку они часто уже обладают необходимой информацией для определения статистических контрольных показателей.

Однако не все НСУ созданы равными. В некоторых странах были образованы НСУ с широкими и четкими полномочиями по сбору данных и беспристрастному и независимому производству официальной статистики. Такие НСУ обычно наделены правом доступа к данным как из государственного, так и из частного секторов при условии обеспечения конфиденциальности. Однако, есть страны, где НСУ имеют право на доступ к данным только от государственных органов. К сожалению, чаще всего право НСУ на доступ к данным не сопровождается конкретными обязательствами по предоставлению данных для их владельцев и адекватными штрафами за несоблюдение этих обязательств. Следовательно, НСУ может обнаружить, что его законодательные полномочия не поддерживаются на практике. Возможно, сначала потребуется укрепить законодательную базу для НСУ, чтобы они могли играть ведущую роль в использовании данных мобильных телефонов для производства официальной статистики.

Обработка данных мобильных телефонов уровня I создаст технические проблемы для НСУ из-за огромного объема массива данных. Одна запись о вызове имеет средний размер 200 байт. Если в стране 6 миллионов абонентов мобильной связи, из которых около 20 процентов звонят каждый день, общий размер файла для записей за один месяц будет составлять как минимум около 200 ГБ. Следовательно, данные уровня I потребуют использования более высокопроизводительных систем баз данных и вычислительного оборудования, чтобы обеспечить обработку в приемлемые сроки. Это, в свою очередь, также потребует новых навыков в области

¹³ Tam, S.-M. and Clarke, F. (2014) Official Statistics and Some Initiatives by the Australian Bureau of Statistics (*Официальная статистика и некоторые инициативы Австралийского бюро статистики*). Представлено на Международной конференции по использованию больших данных в официальной статистике, Пекин, Китай, 28–30 октября 2014
<http://unstats.un.org/unsd/trade/events/2014/Beijing/documents/other/Australia%20Bureau%20of%20Statistics>

инженерии данных. Для сравнения: обработка данных уровня II может выполняться с использованием традиционных компьютеров и систем баз данных. Важно отметить, что технические проблемы преодолимы. Существующий технический потенциал НСУ может быть увеличен за счет инвестиций в оборудование, программное обеспечение и персонал. Для наименее развитых стран ограничивающим фактором может быть нехватка квалифицированных кадров в области инженерии данных на местном рынке труда при ставках заработной платы в государственном секторе¹⁴. В таких обстоятельствах модель партнерства типа В может быть более осуществимой, поскольку тяжелая работа будет выполнена операторами, которые, скорее всего, уже обладают техническими возможностями.

Конкретная роль национальных статистических управлений в партнерстве по доступу к данным мобильных телефонов будет зависеть от их полномочий и возможностей. В идеале НСУ должно взять на себя инициативу по налаживанию партнерских отношений и управлению статистическими результатами. Для устойчивого использования данных мобильных телефонов в фактическом производстве официальной статистики решающее значение будут иметь вклад и статистический надзор со стороны НСУ. Как минимум, НСУ должно быть проинформировано о деятельности партнерства и полученных результатах, поскольку данные мобильных телефонов являются потенциально актуальным источником статистики для широкого круга секторов, находящихся в сфере компетенции НСУ.

Операторы сетей мобильной связи: для отслеживания использования мобильных телефонов и выставления счетов операторы мобильной связи создают CDR, содержащие данные о операциях мобильных телефонов своих абонентов. Эти CDR содержат информацию, по которой можно идентифицировать личность, а также могут использоваться для получения конфиденциальной деловой информации (например, финансового профиля, профиля абонента). Если данные мобильных телефонов будут использоваться для официальной статистики, крайне важно будет получить CDR от всех операторов. Если CDR будут получены не от всех операторов, данные будут отражать только характеристики абонентов участвующих операторов и непреднамеренно раскрывать конфиденциальную бизнес-информацию не участвующим конкурентам. В случае показателей ИКТ частичные данные не подходят для производства официальной статистики.

¹⁴ Нехватка таких навыков может служить основанием для международного сотрудничества и наращивания потенциала.

Готовность операторов предоставлять доступ к CDR будет зависеть от ситуации на рынке и от нормативно-правовой среды. Ключевые факторы могут включать:

- Законодательные требования к обмену данными
- Государственная собственность
- Барьеры для входа
- Общий размер рынка
- Количество предприятий
- Относительные доли рынка

Если по закону операторы уже обязаны предоставлять данные в регулирующий орган по условиям их лицензии или франшизы, тогда будет легче получить доступ к данным мобильных телефонов как от существующих операторов, так и от новых участников рынка (если таковые имеются). В противном случае может возникнуть необходимость рассмотреть вопрос о возможности введения требования о представлении данных. Это будет особенно важно на рынках телекоммуникаций, где государственная доля участия минимальна, поскольку будет меньше каналов для влияния на обмен данными. На рынках, где телекоммуникационные услуги приносят высокие инвестиционные доходы, при этом участие государства в капитале минимально, то есть имеются значительные барьеры для входа, приводящие к меньшему количеству операторов по сравнению с общим размером рынка, сопротивление обмену данными может быть сильным. Следовательно, такое сопротивление необходимо будет предвидеть и преодолевать в начале создания партнерства, в противном случае оно станет серьезным препятствием для доступа к данным с мобильных телефонов для целей официальной статистики. Необходимы четкие механизмы, гарантирующие операторам, что данные не будут использоваться ни для каких других целей, кроме производства официальной статистики. Также необходимы четкие протоколы и соглашения для обеспечения безопасности, защиты персональных данных и конфиденциальности при использовании данных мобильных телефонов как для операторов, так и для государственных заинтересованных сторон.

Предоставление доступа к данным мобильных телефонов для производства официальной статистики может быть выгодным для операторов. Оно может способствовать укреплению доброй воли регулирующих органов, а также помочь продемонстрировать соблюдение нормативных требований и корпоративную социальную ответственность. Его можно использовать как возможность укрепить внутренний потенциал операторов мобильной связи для анализа больших данных для принятия бизнес-решений. Это может улучшить своевременность и качество общедоступных наборов данных и статистических

показателей, которые могут использоваться в качестве исходных данных для определения бизнес-стратегии. При правильном подходе предоставление доступа к данным мобильных телефонов может стать выигрышным сценарием для поставщиков данных.

Орган, регулирующий системы электросвязи (TRA): Партнерские отношения для доступа к данным мобильных телефонов будут более эффективными, когда регуляторы телекоммуникационных систем будут поддерживать связь с операторами и облегчать доступ к данным мобильных телефонов. В некоторых случаях они могут уже обладать полномочиями запрашивать данные мобильных телефонов в качестве необходимого условия для мониторинга лицензионных соглашений, и они также уже могли инвестировать в оборудование и персонал для хранения и обработки этих записей. В таком случае регулирующие органы станут для НСУ важным каналом доступа к данным мобильных телефонов для производства официальной статистики за пределами секторов электросвязи и ИКТ. Даже если они еще не собирают данные мобильных телефонов, они являются государственными учреждениями, которые часто взаимодействуют с операторами в ходе своей работы регуляторов. Таким образом, они, вероятно, будут иметь хорошие возможности для переговоров и посредничества в отношении доступа к данным мобильных телефонов.

Соответствующая роль регулирующих органов в партнерстве будет зависеть, среди прочего, от объема ресурсов, имеющихся в их распоряжении, их относительного уровня технических знаний и опыта, а также их юридических полномочий по принуждению к представлению данных мобильных телефонов. В странах, где ресурсы НСУ перенапряжены, а технические возможности ограничены, но где регулирующий орган уже активно собирает и обрабатывает данные мобильных телефонов, инициатива по созданию партнерства может принадлежать регулирующему органу. Однако после создания партнерства и завершения пилотных проектов ответственность за руководство и управление может быть передана НСУ, основной деятельностью которого является производство официальной статистики. Ключевой тезис заключается в том, что партнерство является более устойчивым и продуктивным, когда регулирующие органы являются ответственными участниками, независимо от того, играют они ведущую роль или нет.

Орган по защите данных (DPA): Страны все чаще законодательно устанавливают правила безопасности данных и создают национальные органы по защите конфиденциальности данных. По состоянию на 2017 год около 90 стран обнародовали по крайней мере некоторые правила в отношении конфиденциальности данных, соблюдение

которых поддерживается органами по обеспечению конфиденциальности данных или другими уполномоченными государственными органами¹⁵. Хотя конкретное содержание законов о защите данных в разных странах различается, их можно отнести к одной из двух правовых парадигм: законы о защите данных в континентальном европейском стиле или законы о персональных данных в англосаксонском стиле¹⁶. Первые часто являются более всеобъемлющими и ограничительными по сравнению со вторыми. Согласно законам континентального европейского типа, подход по умолчанию заключается в том, чтобы запретить и свести к минимуму обработку данных, касающихся идентифицируемых лиц, за исключением случаев, когда есть обоснованное исключение. Обработка данных в статистических, научных и других целях, представляющих общественный интерес, как правило, разрешена при соблюдении мер безопасности, включая, когда это возможно, псевдонимизацию или анонимизацию. Законы же в англосаксонском стиле сосредоточены на защите людей от перехвата конфиденциальных сообщений. Это означает, что, если нет разумных ожиданий в отношении конфиденциальности, то нет и явных ограничений на обработку данных.

Партнерские отношения для доступа к данным мобильного телефона должны будут учитывать конкретные действующие правила и работать вместе с органом по обеспечению конфиденциальности данных, если таковой имеется. Предоставляя рекомендации и надзор за законной обработкой данных, орган по обеспечению конфиденциальности данных должен соблюдать баланс между значением общественных интересов при статистическом использовании данных мобильных телефонов и предполагаемыми рисками для конфиденциальности.

¹⁵ Фирма DLA Piper имеет регулярно обновляемую карту правил конфиденциальности по всему миру: <https://www.dlapiperdataprotection.com/>

¹⁶ Determann (2015) Determann's Field Guide to Data Privacy Law: International Corporate Compliance (*Полевое руководство Детерманна по закону о конфиденциальности данных: международное корпоративное соответствие*)

4 Методы

4.1. Понятия и определения

Данные мобильных телефонов, в частности данные мобильного позиционирования, могут использоваться в нескольких статистических областях, которые выиграют от использования общей концептуальной основы. Статистические области, ранее определенные как возможные для использования этого источника данных, включают статистику туризма, «статью путешествия» платежного баланса, спутниковый счет туризма, статистику пассажирского транспорта, населения, миграции и поездок на работу (маятниковой миграции) (рис. 4.1).

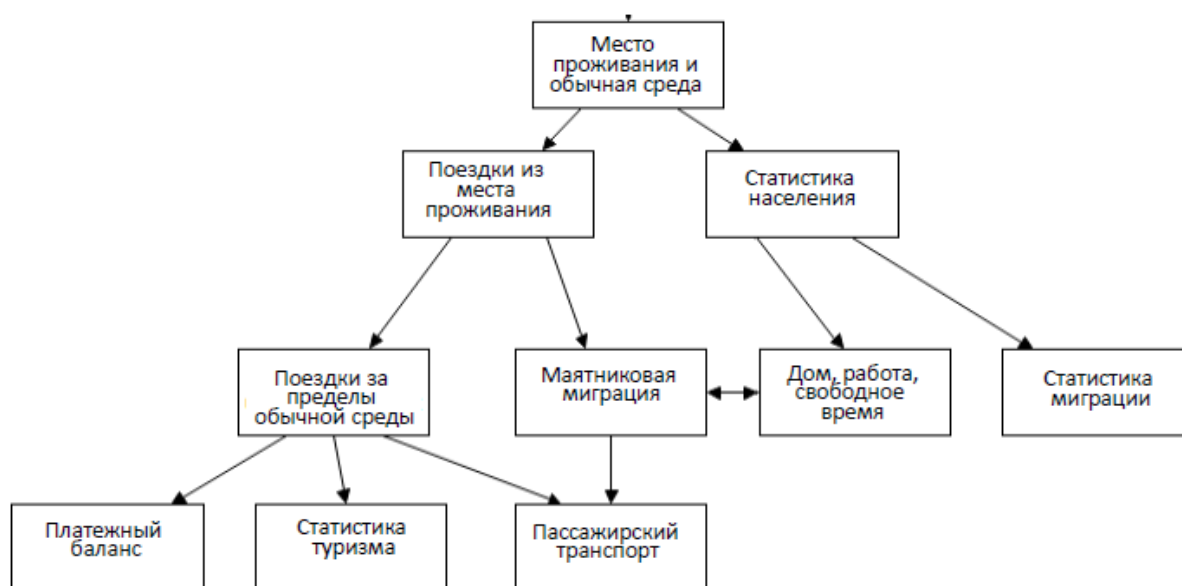


Рисунок 4.1 Связь разных областей статистики¹⁷

В статистике экономики и финансов для оценки синергизма со статистикой туризма в основном важен платежный баланс (ПБ). Статистически «туризм» — это подмножество «путешествий», и, следовательно, (туристы) «посетители» — это подмножество «путешественников». Более того, вопросы туристической деятельности, а именно проблема точного определения обычной среды, не относится к статье «путешествия» в платежном балансе. Методологически при обработке данных может быть достигнуто несколько целей в зависимости от настройки системы. Например, хотя внутренний туризм концентрируется на путешествиях за пределами обычной среды, тот же процесс может быть расширен для выявления любых поездок, совершаемых в обычных условиях.

¹⁷ Евростат (2014) Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics. Report 3a (Исследование возможности использования данных мобильного позиционирования для статистики туризма. Отчет 3 а)

Для калибровки спроса на транспортные услуги и организации предложения услуг транспорта очень важно иметь точные оценки матриц происхождения-назначения. Однако получить эти матрицы обычными методами обследования довольно сложно и очень дорого. Операторы мобильной связи могут предоставлять менее затратные и гораздо более точные матрицы, основанные на данных мобильного позиционирования. Однако данные не покажут вид транспорта или цель поездки.

В статистике населения будут иметь значение пространственное распределение населения (проживающее, работающее) и аспекты мобильности, такие как маятниковая миграция. В отличие от статистики на основе переписи, данные мобильного позиционирования всегда будут недостаточно точными и не будут обеспечивать требуемый уровень детализации. Однако эти данные своевременны и могут предоставить информацию для общих показателей для оценки поездок на работу, миграции и внутренней миграции.

Внедрение системы производства статистики на основе данных мобильного позиционирования довольно дорого; однако, если система внедряется для нескольких областей (туристическая деятельность, включая платежный баланс, транспорт и население), дополнительные затраты на добавление компонентов обработки будут относительно ниже.

В следующих главах будут представлены некоторые полезные понятия для методологии производства различных статистических данных о мобильности людей и их применения при использовании данных мобильного позиционирования.

Понятия, связанные с поездками

Поездка: Поездка: относится к путешествию человека с момента его отъезда из места проживания до его возвращения; поэтому это поездка туда и обратно. Поездка состоит из посещения разных мест. Поездки состоят из одного или нескольких посещений в течение одного и того же путешествия (аналогично официальному определению).

Посещение / пребывание / проживание: означает пребывание в месте, которое посетили во время туристической поездки. Пребывание не обязательно должно быть с ночевкой чтобы считаться туристическим посещением. Тем не менее, понятие пребывания предполагает остановку. Въезд в географический район без остановки не считается посещением этого района. В сфере туризма рекомендуется, чтобы страны определяли минимальную продолжительность остановок для того, чтобы поездка считалась туристическим посещением. Концепция посещения зависит от географического уровня, на

котором она используется. Это может означать либо всю поездку, связанную с туризмом, либо только ее часть, в зависимости от точки зрения (на основе происхождения или места назначения) (аналогично официальному определению).

Пребывание с ночевкой: Понятие, которое используется в качестве критерия для того, чтобы отличать туристов (посетителей с ночевкой) от однодневных посетителей. Считается, что посетитель пребывал в каком-то месте с ночевкой, если предполагается, что он останавливался там во время смены календарных дат (место, в котором прошла ночь, независимо от фактического места отдыха). Если во время смены дат посетитель находится в процессе перемещения между точками А и В в пределах страны, в которой ведется наблюдение, ночевка может быть определена (в зависимости от национальных критериев конкретной страны) в точке А или в точке В, или может вообще не определяться. Однако с точки зрения рассматриваемой страны (местом является указанная страна) посетитель проводит ночь внутри страны (что отличается от официального определения).

Раздел поездки: поездка состоит из разделов пребывания и передвижения. Все разделы (пребывание и передвижение) объединяются в разделы пребывания, если смотреть с более высокого географического уровня. Раздел пребывания в пункте А; раздел движения между А и В; и раздел пребывания в месте В в стране Х объединяются в одно пребывание, если смотреть на уровне страны (новая концепция).

Продолжительность поездки/посещения/пребывания: Мобильное позиционирование позволяет измерить продолжительность посещения в часах, днях и ночах. Продолжительность путешествия в пункт назначения и обратно можно определить и исключить. Общее количество часов и ночей, потраченных на поездку, можно суммировать для всех уровней агрегирования; однако дни присутствия не могут быть суммированы (что отличается от официального определения).

Понятия, подходящие для нескольких областей статистики

Страна (обычного) проживания: Страна, в которой лицо проводит большую часть года.

Место (обычного) проживания: Географическое положение места проживания лица. В случаях, когда это зарубежная страна, местом проживания является страна, поскольку при использовании данных мобильного позиционирования невозможно определить более точное/конкретное расположение места жительства в другой стране. В случае рассматриваемой страны местом проживания является конкретное местоположение в пределах этой страны с точностью, зависящей от метода определения фактической точки

местоположения (например, наименьший административный уровень, наименьшая идентифицируемая единица сетки или географическая точка).

Обычная среда местонахождения: у каждой формы туризма есть свое определение и метод определения места проживания и обычной среды местонахождения. По умолчанию местом проживания и обычной средой для абонентов с входящими данными, является зарубежная страна происхождения абонента, если не указано иное. Для внутренних и исходящих данных обычная среда абонентов может быть определена с точностью до страны, округа, муниципалитета или некоторых других географических районов или административных единиц. Используемый уровень детализации зависит от имеющихся данных и потребностей производителей данных (что отличается от официального определения).

Некоторые понятия, связанные со статистикой туризма

Туризм: деятельность, осуществляемая посетителями, которые совершают поездку в основное место назначения, находящееся за пределами их обычной среды, которая длится менее года и предназначена для любой основной цели, включая бизнес, отдых или другие личные цели, кроме занятости юридическим лицом-резидентом в месте, которое было посещено. Основные характеристики аналогичны официальному определению, хотя лица, которые наняты лицом-резидентом в посещаемом месте, не могут быть исключены из данных мобильного позиционирования.

Основной пункт назначения: Основной пункт назначения туристической поездки определяется как посещаемое место, которое имеет решающее значение для принятия решения о поездке. Однако, если посетитель не может определить такое место, основным пунктом назначения считается место, в котором он проводил большую часть своего времени во время поездки. Опять же, если посетитель не может определить такое место, то основным пунктом назначения считается место, наиболее удаленное от места проживания. В данных мобильного позиционирования необходимо различать основной пункт назначения поездки (который аналогичен официальному определению) и вторичный пункт назначения (новая концепция). Основной пункт назначения поездки можно определить по официальным критериям; однако во время поездок с ночевками каждый день может иметь другое основное место назначения (обычно это то место, в котором проводится ночь). В однодневной поездке основным пунктом назначения является место, в котором была проведена большая часть времени. Посетитель может посетить одно основное место назначения и несколько второстепенных мест назначения в

течение одного дня.

Вторичный пункт назначения/посещения: в отличие от основного пункта назначения, вторичный пункт назначения – это место, которое посещает посетитель (пребывает) в дополнение к основному пункту назначения в течение времени, превышающего минимальную продолжительность остановки для того, чтобы считаться туристическим посещением (новая концепция).

Транзитный проезд: в отличие от посещения основного пункта назначения и вторичного пункта назначения, при транзите посетители проезжают без остановки или останавливаются в течение периода времени, который короче минимальной продолжительности, при которой остановка считается туристическим посещением. Транзитный проезд не считается туристическим визитом. На уровне страны транзитный проезд или транзитные поездки/посещения рассматриваются как поездки, целью которых является проезд через эту страну по пути в страну или из страны, которая является их основным пунктом назначения (аналогично официальному определению).

4.2. Методология обработки данных

Методология состоит из следующих этапов: дополнительная подготовка данных о событиях, формирование совокупности, разработка и оценка данных. Исходные данные, которые извлекаются и подготавливаются операторами мобильной связи, основаны на сетевых событиях, которые определяют присутствие конкретного абонента во времени и пространстве. Дополнительная подготовка может включать географическую привязку, устранение мобильных устройств, не управляемых человеком, проверку времени и зоны покрытия данных, работу с пропущенными значениями и т. д. После того, как данные были подготовлены операторами мобильной связи, устанавливаются следующие этапы обработки (Рисунок 4.2):

- Извлечение данных и обеспечение качества первичных данных
- Формирование совокупности:
 - Применение алгоритмов идентификации поездки – определение индивидуальной поездки каждого абонента в рассматриваемую страну, с указанием времени начала и окончания каждой поездки.
 - Выявление интересующей группы населения (отличие туристической деятельности от нетуристической деятельности):
 - Определение роуминговых абонентов, которые фактически не пересекают границу и не въезжают в страну (входящие, исходящие).

- Определение резидентов (входящие, исходящие данные)
 - Определение места проживания и обычной среды (внутренние)
 - Определение транзитных поездок по стране (входящие)
 - Определение места назначения и транзитных стран (исходящие)
- Разработка данных:
 - Пространственная детализация (посещения на самом мелком административном уровне)
 - Определение переменных (количество посещений, продолжительность поездок, классификация и пр.)
 - Распространение выборочных данных для операторов мобильной связи на всю интересующую совокупность включает:
 - Агрегирование данных во времени и пространстве (день, неделя, месяц, квартал, по сетке (например, один км2), LAU-2, LAU-1, страна)
 - Объединение данных от различных операторов мобильной связи и расчет окончательных статистических показателей

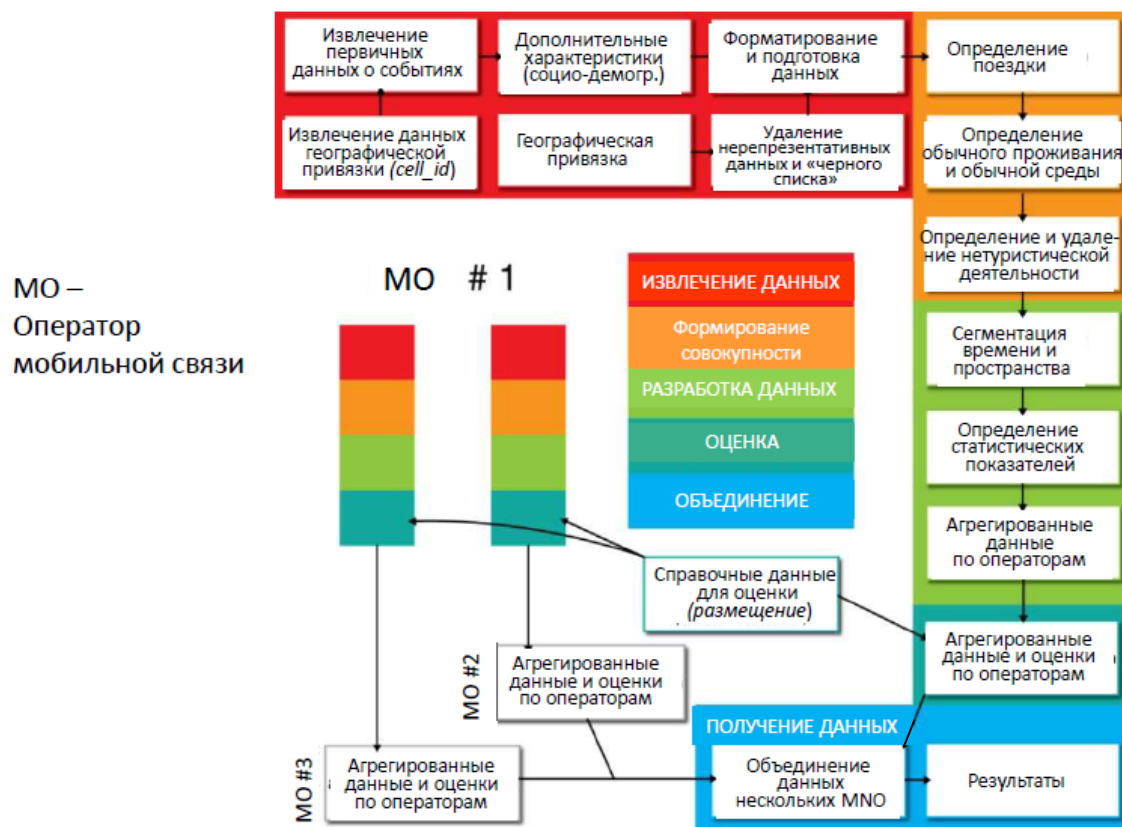


Рисунок 4.2 Этапы обработки данных для производства статистики туризма на основе данных мобильных телефонов ¹⁸

Конкретная методология может зависеть от характеристик и происхождения данных

¹⁸ Евростат (2014) *Исследование возможности использования данных мобильного позиционирования для статистики туризма. Отчет 3 а- Возможность использования, методологические вопросы*

(частота в пространстве и во времени, географическая точность событий и доступные атрибуты событий или абонентов). Процессы, описанные в *Отчете За Евростата об исследовании возможности использования данных мобильного позиционирования для целей статистики туризма* (расчет поездок, определение обычной среды, нерезиденты, транзитные поездки, продолжительность пребывания и т. д.) предполагают, что лонгитюдные расчеты для отдельных абонентов возможны.

Важнейшие процедуры, выполняемые в процессе формирования совокупности, включают определение страны проживания и обычной среды абонентов, чтобы иметь возможность определять туристические поездки (то есть поездки за пределы обычной среды). Для этих расчетов требуются исторические временные ряды и взаимосвязь между различными наборами данных (один и тот же идентификатор абонента во внутренних и исходящих данных) для абонентов, чтобы иметь возможность определять часто посещаемые места и страны.

Одним из важнейших методических вопросов является определение обычного места проживания. Применить официальные определения к мобильным данным сложно, поскольку необходимо определить несколько критериев для построения алгоритма обнаружения домашнего адреса. Согласно исследованиям, проведенным во Франции¹⁹, выбор критериев в алгоритмах обнаружения домашнего адреса существенно влияет на результаты - до 40%. При определении алгоритмов для конкретной страны критерии зависят от источника данных и местных условий. В Эстонии Тартуский университет и компания Positium разработали модели якорных точек для домашних адресов²⁰, точность которых достигает 99% для уровня районов и более 90% на более высоких уровнях административных единиц.

Исследование возможности использования данных мобильного позиционирования для статистики туризма, проведенное по заказу Евростата, содержит описание методологии в *Отчете За*, которое можно использовать в качестве пошагового руководства для

¹⁹ Vanhoof, M et al (2018) Assessing the quality of home detection from mobile phone data for official statistics (*Оценка качества обнаружения домашнего адреса по данным мобильных телефонов для официальной статистики*). Это валидационное исследование, в котором результаты различных алгоритмов обнаружения домашнего адреса сравнивались с результатами переписи на уровне вышек сотовой связи. Исследование доступно на <https://arxiv.org/abs/1809.07567>

²⁰ Ahas et al (2010) Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. (*Использование данных мобильного позиционирования для моделирования местоположений, значимых для пользователей мобильных телефонов*). В статье проверяется точность алгоритма на агрегированном уровне для уровня административных единиц. Также было проведено еще одно валидационное исследование, совместно со НСУ Эстонии, в котором для добровольцев было проведено сравнение якорных точек по их мобильным данным, рассчитанных компанией Positium, данных реестра населения и истинных домашних адресов волонтеров.

производства статистики туризма, поскольку оно достаточно подробное, но в то же время достаточно общее для широкого использования. Предполагается, что набор данных включает все данные, необходимые для построения переменных статистики туризма, например за деятельностью анонимных абонентов можно следить в течение более длительного периода времени, чтобы установить их место проживания и/или обычную среду. Наличие лонгитюдных данных имеет решающее значение для производства статистики туризма, поскольку алгоритмы для определения обычной среды и страны обычного проживания, зависят от наличия исторических данных. Если доступные данные описывают только короткий период времени, возникает проблема ошибок обработки, и, как показывает моделирование, качество данных может быть слишком низким для получения надежных результатов. Такие ограниченные данные могут использоваться в качестве сравнительных показателей для некоторых неофициальных статистических данных (например, количество уникальных иностранных абонентов на какой-то площадке или на концерте) и для относительных сравнений.

4.3. Система обеспечения качества для производства статистики на основе данных сетей мобильной связи

Качество играет важную роль в укреплении доверия к производимым статистическим продуктам. Обеспечение качества является неотъемлемой частью процесса производства в любой статистической организации.

В проектах с использованием данных мобильного позиционирования обеспечение качества стало вопросом первостепенной важности. Система обеспечения качества (QAF) поддерживает обеспечение качества на каждом этапе обработки: 1) обеспечение качества входных данных, 2) обеспечение качества статистических процессов и 3) обеспечение качества продукта (результатов). Кроме того, необходимо создать институциональную среду, способствующую повышению качества статистики. Благодаря этим трем этапам качество производимых статистических данных будет поддерживаться на должном уровне, поскольку оно всегда проверяется на каждом этапе процесса статистического производства.

Хорошая система обеспечения качества (QAF) отражает концепции качества, связанные с производством официальной статистики (например, ESS 2012²¹, UN 2015²², UNECE

²¹ EU, (2012), European Statistical System “Quality Assurance Framework of the European Statistical System” (http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/QAF_2012/EN/QAF_2012-EN.PDF) (ЕС (2012), Европейская статистическая система. «Система обеспечения качества Европейской статистической системы»)

²² UNECE (2014) A Suggested Framework for the Quality of Big Data (*ЕЭК ООН (2014) Предлагаемая основа для оценки качества больших данных*)

2014²³). Система обеспечения качества (QAF) должна рассматривать следующие аспекты:

1. Институциональная среда – соглашения на институциональном уровне между заинтересованными сторонами, подробное описание коммуникации, вопросы конфиденциальности, правила, роли и другие общие вопросы.
2. Качество входных данных – общие руководящие указания в отношении качества на этом этапе обеспечивают готовность исходных входных данных к обработке. При первом доступе или в связи с важными изменениями в исходных входных данных собираются общие метаданные и показатели качества (включая показатели точности и избирательности). Далее на этом этапе основное внимание уделяется метаданным, сопровождающим регулярную передачу данных.
3. Качество обработки – подробное описание статистической обработки в основном касается качества методологии и алгоритмов обработки – проверки алгоритмов, достоверности смоделированных данных, метаданных, протоколирования и отчетности на протяжении всего процесса обработки.
4. Качество продукта – Качественные аспекты результатов обработки данных, правильность переданных файлов данных, соответствие результатов принципам качества, обычно измеряемое степенью согласованности (сопоставимость во времени, по географическому признаку и со справочными данными).

Последний этап – измерение качества продукта – лучше всего знаком статистическим управлениям. Однако, если качество данных обеспечивается на предыдущих этапах профессионально и последовательно, показатели качества выходных данных не должны преподносить никаких сюрпризов. Это первые этапы обработки данных мобильного позиционирования, на которых должно быть обеспечено качество, и выходные данные отражают это.

Стандартов для создания системы обеспечения качества нет, но в ЕЭК ООН продолжается работа по созданию системы обеспечения качества (QAF) для больших данных²⁴, которая применима к данным мобильного позиционирования с некоторыми изменениями. В Индонезии Статистическое управление Индонезии (BPS) имеет комплексную систему обеспечения качества методологии оценки числа прибывающих трансграничных туристов на основе данных мобильного позиционирования (см. соответствующее тематическое

²³ United Nations (2015) Fundamental Principles of Official Statistics: Implementation guidelines (ООН (2015) *Руководство по реализации Основных принципов официальной статистики*)

²⁴ UNECE (2014) A Suggested Framework for the Quality of Big Data (ЕЭК ООН (2014) *Предлагаемая основа для оценки качества больших данных*)

исследование в Приложении).

4.4. Как справиться с недостаточным / избыточным охватом

В случае использования данных мобильного позиционирования есть одно несоответствие, которое сразу становится очевидным при сравнении определения целевой совокупности населения и определения исследуемой выборочной совокупности населения. В выборочной совокупности представлены все те абоненты, которые пользуются мобильными телефонами, в то время как целевая совокупность включает всех лиц, проживающих в стране.

На рис. 4.3 представлена взаимосвязь между целевой совокупностью и полученной окончательной выборкой. Здесь мы различаем население с телефонами (любого типа) и население с мобильными телефонами. Может случиться так, что для всех единиц из наблюдаемой выборки у нас нет информации о целевой переменной (например, о местонахождении жилья), которая должна быть импутирована. В этом случае мы могли бы рассмотреть проблему скрытых переменных, которые наблюдаются не прямо, а косвенно (например, путем анализа местоположения при помощи базовых передающих станций).

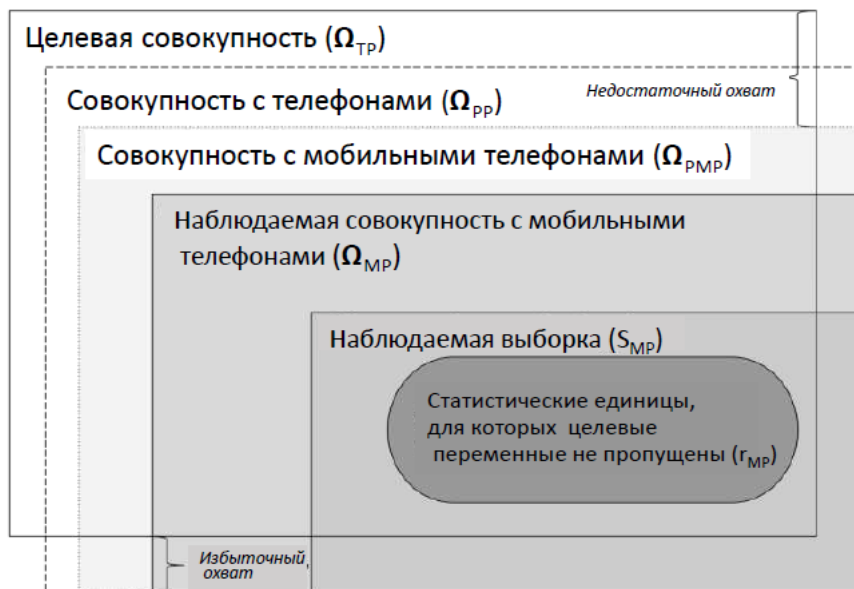


Рисунок 4.3 Движение от целевой к наблюдаемой совокупности при использовании данных мобильных сетей²⁵

Почему самоотбор (селективность) представлен с точки зрения подмножества целевой

²⁵ Maciej Beręsewicz, Risto Lehtonen, Fernando Reis, Loredana Di Consiglio, Martin Karlberg (2018) An overview of methods for treating selectivity in big data sources (Обзор методов обращения с избирательностью в источниках больших данных: <https://ec.europa.eu/eurostat/documents/3888793/9053568/KS-TC-18-004-EN-N.pdf>

совокупности? Независимо от определения целевой совокупности (например, поездки, люди), некоторые единицы всегда будут исключены. Например, инфраструктура мобильной сети может ограничивать точную идентификацию поездок. Однако это следует рассматривать как ошибку охвата. Наблюдаемые поездки — это подмножество совокупности всех поездок (например, ограничивается одним провайдером, только теми, у кого есть мобильные телефоны), и поездку можно идентифицировать, что также связано с охватом инфраструктуры.

Две основные причины проблем с охватом и селективностью проистекают из следующего:

- Охват целевой совокупности (вне зависимости от ее определения). Следует учитывать, что наблюдаемые совокупности абонентов мобильных операторов включают предприятия и физических лиц, и они могут частично совпадать (например, одно лицо с частным и служебным мобильным телефоном).
- Второй вопрос касается селективности станций и того, на каком уровне агрегирования можно использовать эти данные. Конечно, базовые станции располагаются в соответствии с плотностью населения, и в литературе в основном представлены с помощью диаграмм Вороного²⁶.

Одним из наиболее важных мероприятий в начале любого проекта по использованию данных мобильного позиционирования является обеспечение качества входных данных, чтобы иметь очень хорошее описание набора данных и аспектов охвата с точки зрения профиля абонента, данных о событиях и покрытия соты.

Подводя итог, можно сказать о следующих проблемах в отношении выводов на основе мобильных данных:

- Проблемы, связанные с определением и получением целевых совокупностей населения, наблюдаемых в данных мобильной сети.
- Ограниченный доступ к справочной информации о пользователях мобильных устройств (место жительства, пол, возраст, семейное положение) - необходимость импутирования (или профилирования, как описано в литературе)
 - Идентификация избыточного охвата в совокупности;
 - Идентификация статистических единиц;
 - Определение точного местоположения;
 - Включение неопределенности, связанной с импутированием, в оценки.

²⁶ Ricciato et al., 2015

Некоторых проблем с охватом можно избежать во время формирования совокупности или разработки данных (до проведения оценки) путем выявления тех наблюдений, которые не являются частью целевой совокупности, и путем исключения их из совокупности (таблица 4.1).

Таблица 4.1 Список возможных вопросов, касающихся охвата, в отношении данных от мобильных операторов²⁷

Вопрос	Недостаточный или избыточный охват	Возможное решение
1. Данные только от выбранных мобильных операторов (МО), то есть не все данные от операторов имеются	Недостаточный охват	Оцените доли других МО, используя внешнюю информацию о степени проникновения и профилях клиентов.
2. Лица, которые не используют мобильные телефоны	Недостаточный охват	На этапе оценки используйте информацию из дополнительных источников, чтобы сделать и проверить предположения о поведении недостаточно охваченной группы, а затем примените соответствующие модели для корректировки оценок. Могут использоваться оценки как на основе моделей, так и полученные при помощи моделей.
3. Использование более одного мобильного устройства: • Все устройства используют одну и ту же сеть, вызывая дублирование в одном из наборов данных МО • Устройства используют разные сети, вызывая дублирование данных, предоставляемых разными МО.	Избыточный охват	На этапе оценки используйте информацию из дополнительных источников - если таковая имеется - чтобы сделать и проверить предположения о количестве людей, имеющих несколько мобильных телефонов, а затем применить соответствующие модели для корректировки оценок. Корректировка зависит от частоты явления.
4. Разные степени проникновения зарубежных абонентов конкретных стран в сети мобильных операторов.	Недостаточный или избыточный охват	Предполагая, что имеется информация, позволяющая создавать отдельные модели на основе МО с параметрами для конкретной страны, можно вычислить оценки с поправкой на вес.

²⁷ На основе: Евростат (2014) Исследование возможности использования данных мобильного позиционирования для статистики туризма. Отчет 3 а – Возможность использования, методологические вопросы

5. Разные региональные и социально-демографические показатели проникновения для МО	Недостаточный или избыточный охват	Обследования, посвященные региональным и социально-демографическим уровням проникновения МО, если таковые имеются, могут использоваться для оценок.
6. Устройства межмашинной связи (M2M), которые не были удалены МО в процессе подготовки данных.	Избыточный охват	Возможное удаление на основе моделей событий. На основании данных о входящем роуминге в Эстонии процент таких устройств в первичных данных составляет от 0,04% до 2% в зависимости от МО и используемого алгоритма. По данным эстонского исходящего роуминга, доля таких устройств очень мала, менее 0,01%.
7. Входящие: национальные роуминговые абоненты.	Избыточный охват	Может быть полностью исключен, так как страна роуминг-партнера совпадает с рассматриваемой страной
8. Входящие: посетители, фактически не въезжающие / не выезжающие из рассматриваемой страны (не пересекающие границу), но использующие услугу роуминга МО в рассматриваемой или зарубежной стране.	Избыточный охват	Может быть частично исключен на основе алгоритмов распознавания искажения границы. По данным для Эстонии, процент исключенных поездок среди входящего роуминга составил 10,4% от всех поездок. Всего 58% таких поездок классифицируются как поездки только с одним событием.
9. Туризм: однодневные посещения представлены избыточно по сравнению с посещениями с ночевками. Это происходит с данными CDR, когда посетители не используют свои мобильные телефоны каждый день в поездке.	Избыточный охват более коротких посещений за счет более длительных посещений. Неполный охват продолжительности поездок.	Смоделируйте вероятность того, что одна поездка будет однодневной поездкой без ночевки, и используйте смоделированные значения для оценки. В случае наличия надежных справочных данных о разнице между однодневными и ночующими посетителями, такие данные следует использовать для корректировки данных мобильных телефонов
10. Входящие: посетители, которые по какой-то причине используют местные SIM-карты и не представлены в реестре входящего / исходящего роуминга (см. также п. 18 в этой таблице).	Недостаточный охват	Калибровка или аналогичный подход, при котором известные итоговые значения из других источников используются для корректировки недостаточно охваченной части генеральной совокупности. Используйте данные о происхождении владельцев SIM-карт, если в стране действуют обязательные требования к регистрации.

11. Технологические ограничения на использование мобильных телефонов из / в определенных странах из-за технологических барьеров, ограничений на услуги роуминга (отсутствие соглашения о роуминге между МО), высокий уровень роуминга	Недостаточный охват	При оценке необходимо учитывать различия в конкретных странах, компенсируя технологические ограничения. В Эстонии абоненты из США и Японии находятся в такой ситуации и поэтому требуют более высоких поправочных коэффициентов.
12. SIM-карты для международного роуминга (путешествия), которые продаются по всему миру, и там, где страна МО не совпадает с фактической страной происхождения абонента	Избыточный охват в стране МО, недостаточный охват абонентов из стран, использующих эту услугу	Исключение таких абонентов должно быть возможным для исходящего роуминга, поскольку МО обычно знают, кто их абоненты, использующие их услугу международного роуминга для путешествий. Может потребоваться возможная компенсация за счет фактических жителей рассматриваемой страны. Количество таких роуминговых SIM-карт сильно различается по странам и его очень трудно измерить
13. Различия в способах использования телефона в зависимости от особенностей местоположения, в связи с чем генерирование разных объемов событий.	Недостаточный охват в сельской местности	На этапе оценки используйте информацию из дополнительных источников, чтобы сделать и проверить предположения о поведении недостаточно охваченной группы, а затем применить подходящие модели для корректировки оценок
14. Ограничения на использование мобильных телефонов в определенных регионах одной страны с плохим покрытием сети.	Недостаточный охват	Если территория большая, используйте соответствующую информацию из дополнительных источников, чтобы сделать и проверить предположения о поведении недостаточно охваченной группы и применить соответствующие модели для корректировки методов оценки малых районов, например может быть применена синтетическая оценка
15. Чем ниже географический уровень, тем менее репрезентативными будут данные мобильных телефонов, когда речь идет о поездках в эти места или через них.	Недостаточный охват на более низких административных уровнях	Более высокие поправочные коэффициенты для более низких административных уровней по сравнению с более высокими административными уровнями

16. Перекрестный роуминг, то есть одно устройство, использующее несколько роуминговых услуг во время поездки, вызывая дублирование данных для одного абонента в наборах данных двух или более МО	Избыточный охват	Перекрестный роуминг можно оценить, используя методы сохранения конфиденциальности и сравнивая списки роуминговых IMSI между операторами — это было сделано в Индонезии для определения соотношений долей рынка.
17. Входящие: резиденты страны, использующие зарубежные телефоны	Избыточный охват	Может быть частично исключен в зависимости от продолжительности пребывания в рассматриваемой стране. По данным Эстонии, около 0,2–0,4% от общего количества поездок можно идентифицировать как поездки, совершаемые местными жителями, а не как туризм.
18. Туризм: Транзитные посетители в стране.	Избыточный охват	Может быть частично исключен на основании определения коротких поездок в транзитных коридорах. В зависимости от месяца доля транзитных поездок в данных по Эстонии колеблется от 2% до 10% от общего количества въездных поездок, а 9,2% пребывания в зарубежных странах можно рассматривать как транзитные.
19. Туризм: абоненты-нерезиденты, чья обычная среда и место проживания находятся за пределами рассматриваемой страны, но которые по разным причинам используют местные SIM-карты с предоплатой (см. также пункты 10 и 20).	Избыточный охват	Исключите краткосрочных абонентов или, если возможна привязка идентификатора к исходящим данным, сравните продолжительность пребывания в рассматриваемой стране и за границей для исключения из внутренних и исходящих данных и включения во входящие данные. По данным Эстонии, 2,9% всех внутренних абонентов проводят большую часть своего времени за границей, что составляет 24% всех выездных поездок, и поэтому их следует рассматривать как иностранных резидентов и исключать из набора данных по внутреннему туризму.

20. Абоненты-резиденты, которые очень часто меняют свои предоплаченные карты (короткий срок действия <i>идентификатора абонента</i>) и чье место жительства и / или обычная среда не могут быть определены (см. также п. 19).	Недостаточный или избыточный охват	Исключите краткосрочных абонентов
21. Туризм: абоненты, чье место жительства и / или обычная среда в стране рассчитывается неправильно.	Недостаточный или избыточный охват	Обычное место жительства и обычная среда «компенсируются» аналогичными некорректными расчетами для других абонентов. Количество неверных вычислений зависит от уровня пространственного разрешения.
22. Исходящие: жители зарубежных стран, использующие SIM-карты рассматриваемой страны	Избыточный охват	По данным по Эстонии, примерно 3% от суммы внутренних и исходящих абонентов проводят за границей больше времени, чем в рассматриваемой стране. Если связь идентификатора с исходящими данными возможна, сравните продолжительность пребывания в рассматриваемой стране и за границей для исключения из внутренних и исходящих данных и включения во входящие данные. Долю таких лиц также можно оценить.

Есть много факторов, вызывающих смещение охвата, но из-за их совместного влияния некоторые факторы взаимно уничтожаются (чрезмерный охват по сравнению с недостаточным охватом), некоторые влияют незначительно, а другие могут оказывать большое воздействие. Однако многие проблемы присущи самим данным мобильного позиционирования, и поэтому их нельзя избежать. Более того их общее воздействие, то есть степень смещения охвата для интересующей оценки необходимо оценить, или вычислить оценки с поправкой на смещение.

Некоторая информация об использовании мобильных телефонов во время путешествий доступна для Европы. Европейская комиссия регулярно проводит исследования «Евробарометр», и в некоторых из них изучается поведение при нахождении в роуминге. В выпуске «Флэш-Евробарометр» 468²⁸ было показано, что четыре пятых от путешественников теперь пользуются роумингом в других странах Европы после того, как роуминг стал бесплатным в ЕС (рис. 4.4). Ранее в Специальном выпуске «Евробарометра»

²⁸ «Флэш-Евробарометр» 468 (2018) отчет об исследовании по окончании взимания платы за роуминг внутри ЕС: <https://ec.europa.eu/digital-single-market/en/news/eurobarometer-survey-report-end-roaming>

414²⁹ было показано, что 28% путешественников в ЕС выключают свои мобильные телефоны при посещении другой страны ЕС (на Рисунке 4.5 показаны данные по странам).

В 2а Последний раз, когда вы посещали другую страну ЕС, как часто вы пользовались следующими услугами на вашем мобильном телефоне

% - ЕС, 2018

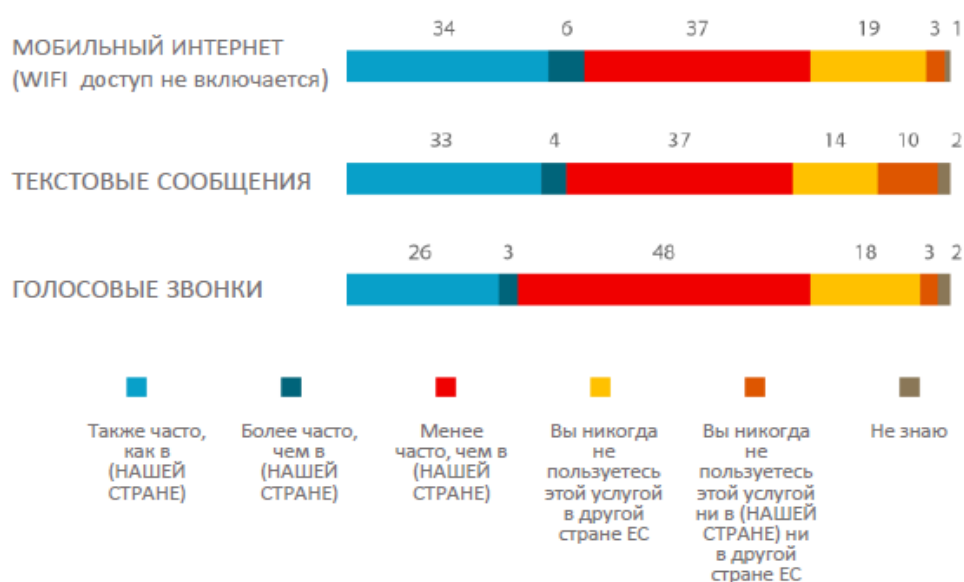


Рисунок 4.4. Процент лиц, опрошенных в ЕС в 2018 г. о том, как часто они используют разные услуги роуминга при поездках в другие страны ЕС (одна пятая не используют).

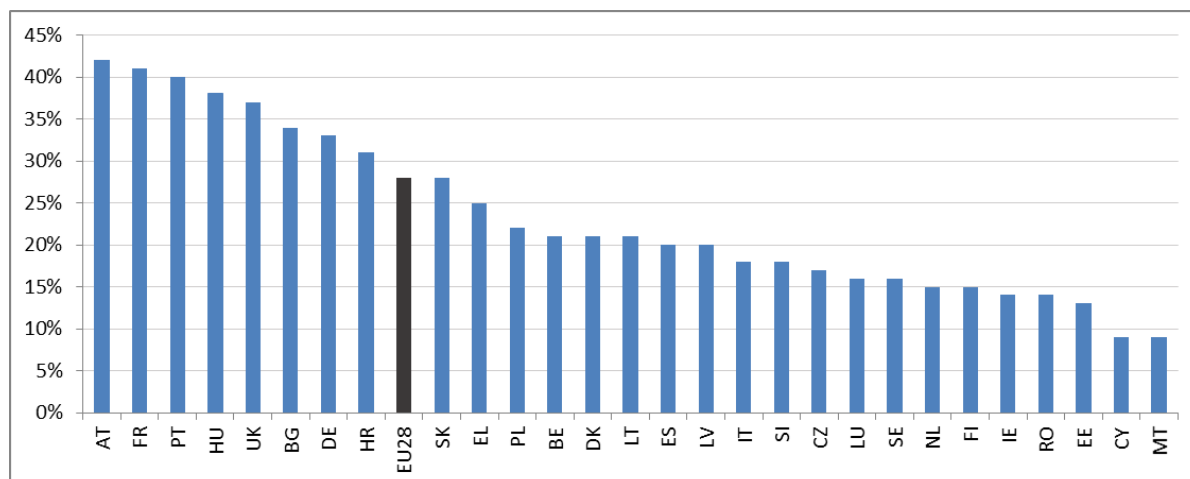


Рисунок 4.5 Доля лиц (в возрасте 15 лет и старше, путешествующих и имеющих личный мобильный телефон), которые обычно выключают свой мобильный телефон и никогда не используют его во время посещения другой страны ЕС ³⁰

²⁹ Специальный выпуск Евробарометра 414 (2014). Результаты обследования домашних хозяйств в области электронных коммуникаций и роуминга на едином рынке телекоммуникаций. Исследование сети TNS Opinion & Social по запросу Европейской комиссии <https://ec.europa.eu/digital-agenda/en/news/e-communications-household-survey-and-telecom-single-market-survey-roaming-results-special>

³⁰ Там же

Лучшим, но и более дорогостоящим способом оценки смещения было бы проведение выборочных обследований с целью оценки качества по каждому аспекту охвата, обеспечивающих основу для индивидуальных оценок. Эти обследования позволяют получить оценки доли людей, не имеющих или не использующих мобильные телефоны, доли людей, использующих два или более устройства и т. д. Эта информация позволит вычислить общую степень смещения при определенных предположениях. Связывание данных обследования с данными мобильного позиционирования на микроуровне (которое выявило бы проблемы с качеством не только на агрегированном уровне, но и на уровне записей) позволило бы проводить более точные оценки смещения. Однако это решение идеально с теоретической точки зрения, но осуществить его на практике не представляется реалистичным.

Менее затратный способ для уменьшения смещения состоит в использовании информации из других источников (например, результатов опросов для выявления взаимосвязи между количеством однодневных поездок и многодневных поездок в страну для въездного туризма, использование мобильных телефонов в разных странах для выездного туризма). Литература по выборочным обследованиям показала, что с ошибкой охвата можно справиться с помощью таких методов, как калибровка, если имеется подходящая вспомогательная информация. Кроме того, могут быть построены различные оценки на основе моделей с использованием агрегированных данных из других источников. Оценки, полученные при помощи моделей и на основе моделей, хорошо освещены в литературе, в которой также представлены инструменты для выполнения валидации моделей и методы вычисления показателей точности.

Если бы не было проблем с охватом и данные от операторов мобильной связи представляли бы идеальную основу для интересующей совокупности, в оценке не было бы необходимости. Однако этого никогда не бывает, и поэтому использование данных от всех операторов мобильной связи из рассматриваемой страны снижает общую систематическую ошибку охвата, но получение оценок по-прежнему необходимо.

Оценка степени всех перечисленных проблем охвата должна выполняться для каждой страны отдельно, так как многие перечисленные здесь проблемы зависят от среды (например, цены на звонки) и культурных факторов (например, у детей есть мобильные телефоны, у абонентов имеется много мобильных телефонов).

Приложение 1 – Тематическое исследование: Франция

Введение

В июле 2015 года Банк Франции и департамент статистики туризма Министерства экономики Франции³¹ начали эксперимент по изучению возможности использования статистики на основе данных позиционирования мобильных телефонов для разработки статистики туризма. Цель этого эксперимента состояла в том, чтобы оценить, может ли статистика на основе данных мобильного позиционирования постепенно заменять данные о передвижениях, которые в настоящее время используются для оценки потоков посетителей-нерезидентов (туристов и однодневных посетителей) во Франции. Этот эксперимент позволил французским статистикам заниматься вопросами, связанными с большими данными: доступом к данным, нормативными ограничения, методологией и качеством оценок.

Ситуация в стране

Франция – одна из ведущих стран в мире по развитию туризма, с огромным разнообразием происхождения посетителей и быстро меняющимися моделями. Поступления на счет текущих операций платежного баланса превышают 40 млрд долларов США в год и составляют значительную часть сальдо счета текущих операций. На этом фоне одной из задач для французских статистиков является поддержание долговременного качества статистики туризма (показатели количества посетителей и поступлений от туризма). В настоящее время эта статистика основана на имеющихся данных о передвижениях по видам транспорта в сочетании с обследованием на границе. Обследование на границе включает ручной подсчет путешественников в аэропортах и в выбранных пунктах выезда на дорогах с целью оценки доли путешественников-нерезидентов в общем потоке выезжающих лиц. Затем разбивка по странам оценивается с помощью анкет, собранных в ходе пограничного обследования. Метод получения статистики затруднен внешними факторами, такими как важность транзитного трафика в узловых аэропортах и препятствия для измерения движения по дорогам в зоне открытых границ, такой как страны Шенгенского соглашения. В этой ситуации данные мобильного позиционирования могут предоставить возможность оценить количество посетителей-нерезидентов по странам происхождения. В настоящее время данные мобильного позиционирования уже используются для статистики туризма, в основном, в Эстонии и в

³¹ Банк Франции отвечает за составление платежного баланса (включая статью «путешествия»), а Управление предприятий (DGE) Министерства экономики Франции отвечает за оценку статистики туризма.

некоторых местных администрациях во Франции.

Потребность государственных органов во Франции в данных о туризме

Статистические данные, необходимые Банку Франции и Министерству экономики Франции, включают количество прибывших посетителей и количество туристов, останавливающихся с ночевкой во Франции, в месяц, с разбивкой по стране происхождения и, если возможно, по виду транспорта. Обследование на границе в любом случае должно быть необходимо для получения данных для оценки расходов посетителей.

Правовая основа и сотрудничество между НСУ и операторами мобильной связи

Во Франции операторам мобильной связи не разрешается сохранять и продавать индивидуальные данные, полученные в результате мобильного позиционирования. Однако, операторам мобильной связи разрешается собирать и обрабатывать личные данные, если они соблюдают законодательство о защите конфиденциальности данных. Вот почему операторы мобильной связи начали предлагать статистические услуги частным компаниям и администрациям. Поэтому Банк Франции и Министерство экономики Франции в июне 2015 года объявили конкурс на государственные закупки для приобретения необходимых оценок у одного оператора мобильной связи (МО). Задачей МО было предоставление необходимой информации ежемесячно с задержкой в один месяц в течение одного года. Два МО представили предложения, и был выбран один кандидат, принимая во внимание, среди прочего, способность скорректировать свои собственные данные с учетом доли рынка. После первого года было решено продлить эксперимент еще на один год, потому что первого года было недостаточно для завершения проекта.

Что касается интеллектуальной собственности, то алгоритмы, разработанные МО и связанные с преобразованием сигнала мобильного телефона в статистические данные, не раскрываются и остаются интеллектуальной собственностью МО, в то время как статистика, полученная на основе этих необработанных данных, является собственностью НСУ.

Дизайн проекта

Дизайн проекта состоит в определении поведенческих критериев, соответствующих определению туризма: ночевка определяется как присутствие между полуночью и 6 утра; прибытие туриста определяется как ночевка без пребывания накануне; наконец, отъезд туриста определяется как ночевка с последующим отсутствием следующей ночью.

Используется прокси, и предполагается, что страной проживания посетителя является страна его SIM-карты.

Источники данных и методология

МО создает алгоритм, который в качестве входящих данных использует индивидуальные данные о соединениях между мобильными телефонами и антеннами (SIM-карта, дата и время, антенна, активность мобильного телефона³²). Индивидуальные данные не сохраняются: алгоритм работает в режиме реального времени и рассчитывает количество людей, соответствующих различным критериям (прибытие, ночевка...). Главный недостаток этой методологии заключается в том, что изменения критериев не могут быть включены в алгоритм ретроспективно.

Алгоритм включает в себя различные процедуры распространения выборочных данных на генеральную совокупность, которые необходимы для оценки количества посетителей на основе наблюдения за мобильными телефонами (рис. 0.1). Некоторые из этих процедур распространения для роуминга, связанные с рыночной долей оператора мобильной связи, не могут быть изолированными и являются частью процесса анонимизации. Другие факторы распространения выборочных данных на генеральную совокупность связаны с использованием мобильных телефонов и владением ими. Их можно модифицировать. Это допущение оказалось очень важным на практике, поскольку операторы мобильной связи не имеют точных сведений о привычках использования мобильных телефонов иностранными посетителями. Оснащенность оборудованием, наблюдаемая в стране происхождения посетителя, в основном используется в качестве прокси для оценки использования оборудования за границей. Легко представить, насколько сильным может быть разрыв для некоторых стран между уровнем оснащенности всего населения и оснащенностью туристов, выезжающих за границу.

³² Эти индивидуальные данные включают как пассивную сигнализацию, так и активность мобильных телефонов (вызовы, сообщения и пр.)

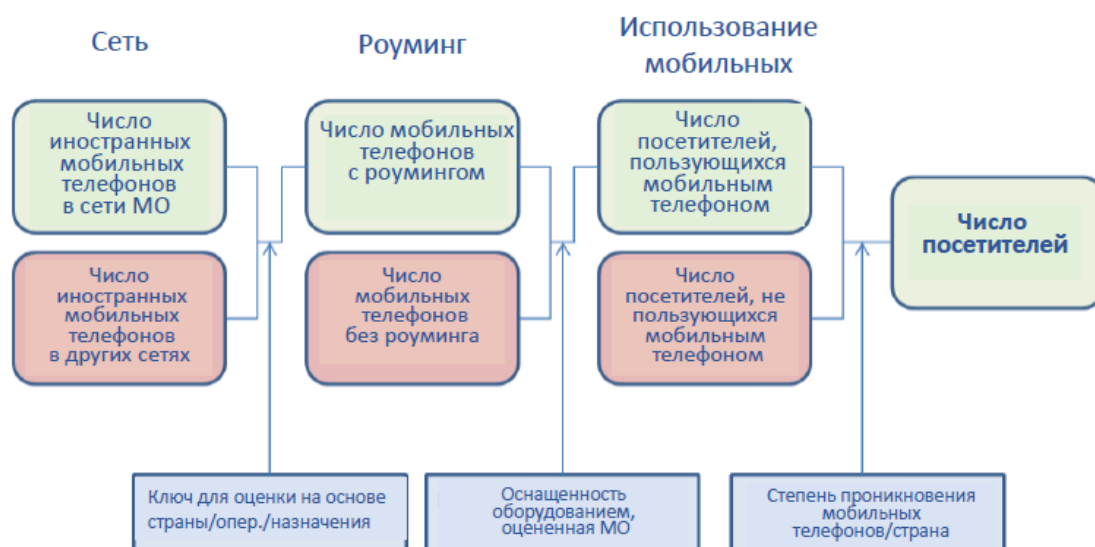


Рисунок 0.1 Распространение выборочных данных: от числа мобильных телефонов к числу посетителей

Качество

Проблемы и вопросы улучшения качества

В ходе эксперимента мы сначала обнаружили большие расхождения между оценками на основе мобильного позиционирования и оценками на основе нашего обследования. Поэтому мы регулярно стали работать с операторами мобильной связи, чтобы выявлять потенциальные проблемы и повышать качество. Ниже кратко излагаются основные проблемы, которые повлияли на качество оценок мобильного позиционирования, и принятые меры:

- **Резиденты Франции, могут иметь иностранные SIM-карты** (например, приграничные работники) и могут быть идентифицированы как иностранцы (основные примеры – Люксембург и Швейцария)
- **Произвольное переключение сети:** несмотря на привилегированные соглашения о роуминге между иностранными и местными операторами мобильной связи, иностранные мобильные телефоны могут менять сеть, что вносит погрешность в оценки из-за использования алгоритма досчета до генеральной совокупности с учетом доли рынка.
- **Использование мобильных телефонов:** использование мобильных телефонов туристами из дальних стран оценено недостаточно. Следовательно, распространение данных на генеральную совокупность приводило к ненадежным оценкам для таких стран, как США, Канада и Китай. На самом деле, высокая плата за роуминг, взимаемая с посетителей из дальних стран, может способствовать использованию местных SIM-карт

во время путешествий, и нет точных данных, чтобы это учесть.

- **Неправильное прерывание пребывания:** было замечено, что значительная часть прибытий относилась к посетителям, которые уже были в стране несколько дней назад. Это явление может быть связано с фактическим пребыванием (например, водители грузовиков), но также может быть связано с посетителями, которые на самом деле не выезжали из страны (смена сети, отключение мобильной связи в течение определенного периода...).

Были приняты различные изменения в методологии, чтобы учесть эти явления:

- **Резиденты Франции с иностранными SIM-картами:** был введен новый критерий, в соответствии с которым лица, которые провели 30 ночей во Франции из последних 60 ночей, считаются резидентами. Из-за процесса обучения результаты были доступны только в конце февраля 2016 г.
- **Произвольное переключение сети:** идея заключалась в том, чтобы ввести различные критерии, чтобы гарантировать, что учтенные мобильные телефоны принадлежат реальным пользователям сети МО. В ноябре 2015 года был добавлен критерий, связанный с кумулятивным подключением к сети. Наконец, в феврале 2016 года был добавлен критерий, связанный с использованием сети (звонки, сообщения ...).
- Текущий рабочий подход — это выбор мобильных телефонов, чьи иностранные операторы мобильной связи имеют привилегированное соглашение о роуминге с французскими операторами мобильной связи.
- **Прерванные посещения:** проводились эксперименты с разными условиями отсутствия (от одного до шести дней) до приезда; оптимальный критерий пока не найден.
- **Использование мобильных телефонов:** для того, чтобы улучшить факторы досчета до генеральной совокупности, связанные с использованием мобильных телефонов, в обследование на границе были добавлены некоторые вопросы по этому поводу. Первые результаты будут доступны только в конце 2017 года (Рисунок 0.2).

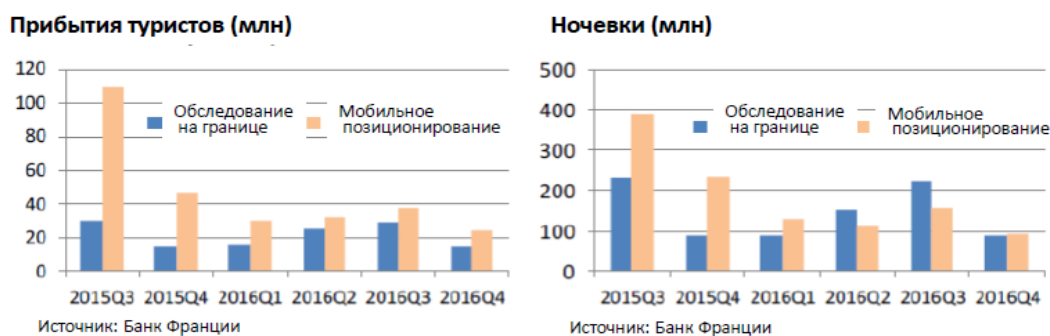


Рисунок 0.2

Сравнение оценок на основе мобильного позиционирования и обследования на границе

Не предполагается, что оценки на основе мобильного позиционирования будут в точности соответствовать нашим оценкам, полученным на основе обследования на границе, особенно потому, что ожидалось, что мобильное позиционирование улучшит качество наших оценок (таблица 0.1). Однако сравнение с оценками нашего обследования на границе — это способ измерения надежности оценок мобильного позиционирования. Как было сказано, мы сначала наблюдали огромные расхождения, которые затем были сокращены благодаря методологическим усовершенствованиям, описанным выше. Эти методологические улучшения, хотя и вызванные несоответствиями с данными обследования, не зависят от них. Количество ночевок оказалось более надежным показателем, чем количество прибытий, потому что последнее более чувствительно к прерванным пребываниям. В изучаемый период мы наблюдаем конвергенцию между оценками мобильного позиционирования и оценками на основе наших обследований на границе для агрегированных показателей. Однако, когда речь идет о цифрах на национальном уровне, все еще имеются очень большие различия. В основном, в настоящее время при помощи мобильного позиционирования недооцениваются дальние страны из-за плохого знания поведения в роуминге для этих стран. Мы надеемся, что наши будущие результаты обследования помогут нам решить эту проблему.

Таблица 0.1

Прибытие туристов и ночевки (млн)

		2015Q3	2015Q4	2016Q1	2016Q2	2016Q3	2016Q4
Прибытия	Обследование на границе	29,5	14,5	15,3	24,6	28,5	14,3
	Мобильное позиционирование	109,6	46,0	29,6	31,4	37,2	24,5
	разница	272%	217%	93%	28%	30%	71%
Ночевки	Обследование на границе	231,2	88,1	87,0	154,7	223,6	86,6
	Мобильное позиционирование	387,0	234,1	128,6	114,4	155,2	93,4
	разница	67%	166%	48%	-26%	-31%	8%

Источник: Банк Франции

Усвоенные уроки

Среди преимуществ мобильного позиционирования можно назвать то, что этот источник данных в значительной степени представляет интерес потому, что он позволяет измерять краткосрочные изменения в связи с наличием ежедневных данных. Эти данные доступны с небольшой задержкой и относительно эффективны с точки зрения затрат по сравнению с

используемыми в настоящее время квартальными данными о передвижении.

Что касается этих свойств (измерение краткосрочных изменений, своевременность, эффективность с точки зрения затрат), то еще предстоит оценить, в какой степени данные мобильных телефонов привносят нечто большее, чем имеющиеся данные кредитных карт, уже используемые для отслеживания краткосрочных тенденций (хотя ежемесячно, а не ежедневно). Однако эти два источника данных на самом деле не заменяют друг друга, так как данные кредитных карт напрямую касаются расходов, в то время как данные мобильных телефонов предназначены для оценки количества людей и поэтому кажутся более подходящими для статистики туризма.

Сейчас мы считаем, что данные мобильных телефонов обеспечат конкретное и неоспоримое преимущество, если можно будет получить надежные оценки «количества ночевок в разбивке по стране происхождения туристов». Сопоставление этих данных с информацией, полученной в результате обследования, и с данными кредитных карт может улучшить процесс определения численности посетителей и позволить более точно оценить относительные показатели «расходов на одного посетителя» для каждой страны.

В то же время эксперимент дал нам возможность понять, что получение надежных оценок требует времени. Итак, по прошествии 18 месяцев Банк Франции и его партнеры по-прежнему считают, что оценки на основе мобильного позиционирования не готовы заменить оценки, получаемые на основе имеющейся системы. Если продление экспериментального периода не увенчается успехом, мы должны решить, стоит ли использовать эти данные в качестве дополнительных отправных точек.

Важно: Данные и выводы, представленные в этом тематическом исследовании, предназначены для работы Целевой группы по использованию данных мобильных телефонов. Их не следует распространять за пределами этой целевой группы, и они могут измениться до окончательной публикации в связи с изменениями в цифрах и необходимыми разрешениями от различных организаций, затронутых этим тематическим исследованием.

Приложение 2 – Тематическое исследование: Индонезия

Введение

Национальное статистическое управление Индонезии (BPS-Statistics Indonesia) — это учреждение, которое производит и публикует данные о туризме в Индонезии. До использования данных мобильного позиционирования основным источником данных о прибывающих туристах были административные данные Главного иммиграционного управления, которые охватывали все основные аэропорты и пункты въезда. Другим источником данных было обследование при пересечении границы (туда и обратно), проводимое НСУ с целью получения данных о туристах / посетителях в определенных пунктах въезда/выезда на границе, где административные данные отсутствуют, или обследование на границе не может быть проведено из-за географических условий.

В октябре 2016 года Статистическое управление Индонезии приступило к внедрению данных мобильного позиционирования для нескольких приграничных территорий. Целью является расширение охвата и возможность точного сбора данных о прибывающих туристах. Таким образом, данные мобильного позиционирования используются в качестве дополнения к другим источникам, которыми являются административные данные и обследование.

Цель

Объем туризма в Индонезии из года в год растет. Об этом свидетельствует рост числа иностранных туристов, время от времени посещающих Индонезию. Как упоминалось ранее, до октября 2016 года статистика въездного туризма основывалась на административных данных и обследованиях. Обследования включали в себя обследование при пересечении границы и обследование при выезде пассажиров (PES). Обследование при пересечении границы проводится на пограничных въездах, где нет иммиграционных контрольно-пропускных пунктов, и пункт въезда/ворота охраняются только военными или главой деревни. Однако Индонезия имеет довольно обширную территорию, и многие границы труднодоступны, поэтому проведения обследования на границе требовало высоких затрат, особенно на транспортировку.

Поэтому НСУ Индонезии начало использовать данные мобильного позиционирования для производства статистики, опубликованной в октябре 2016 года, чтобы иметь возможность точно фиксировать и увеличивать охват прибытий международных посетителей. Данные были получены от одного оператора мобильной связи (МО), который имеет самую

высокую долю рынка в приграничной зоне (около 92 процентов на границе и 70 процентов на национальном уровне).

Мобильное позиционирование используется для корректировки данных пограничных постов в 19 округах, в которых иммиграционный контрольно-пропускной пункт отсутствует, а обследование при пересечении границы трудно провести из-за географических трудностей, или оно требует больших затрат. Используя технологию мобильного позиционирования, мы можем эффективно и точно регистрировать и контролировать прибытия иностранных туристов на границу.

Для НСУ Индонезии использование данных мобильного позиционирования в статистике туризма соответствует стратегическому плану на 2015-2019 гг. (Оптимизация использования ИКТ) и рекомендациям ООН, изложенным в докладе консультативной группы экспертов по революции в использовании данных в целях устойчивого развития, озаглавленном “A World That Counts” («Мир, который имеет значение»). Революция в области данных определяется как взрывной рост объема данных, скорости их производства, количества производителей данных, распространения данных и диапазона вещей, по которым имеются данные, полученные с помощью новых технологий, таких как мобильные телефоны и Интернет вещей, а также из других источников, таких как качественные данные, данные, полученные от граждан, и данные о восприятии. Революция данных в целях устойчивого развития означает интеграцию этих новых данных с традиционными данными для производства высококачественной информации, которая является более подробной, своевременной и актуальной для многих целей и пользователей, особенно для стимулирования и мониторинга устойчивого развития.

Методология

Как упоминалось выше, данные мобильного позиционирования получают (только) от одного из крупнейших операторов мобильной связи в Индонезии. В настоящее время НСУ получает данные мобильного позиционирования от Министерства туризма, поскольку существует Меморандум о взаимопонимании между НСУ Индонезии и Министерством туризма. Соглашение таково: Оператор мобильной связи обрабатывает данные мобильного позиционирования, превращая их в статистические данные, и передает в Министерство туризма, которое передает их в НСУ. Правила и фильтры были установлены НСУ, Министерство туризма подготовило бюджет для обработки данных мобильного позиционирования. Таким образом, необработанные индивидуальные данные остаются у операторов мобильной связи, который предоставляет статистику в НСУ Индонезии каждый месяц с лагом $n-1$.

Используемый фильтр заключается в том, что каждый мобильный телефон, находящийся в регионе (районе) не менее семи дней подряд или в совокупности менее 20 дней, считается принадлежащим туристу. Если два или более номера (IMSI) всегда находятся в одной местности и близко друг к другу, то они считаются принадлежащими одному туристу, а не двум или трем туристам. Для прибывающих туристов / посетителей используемые данные мобильного позиционирования — это данные сигнализации, а не данные записи вызовов (CDR). Однако для выездного туризма используемые данные мобильного позиционирования — это CDR, поскольку нет данных сигнализации.

Статистика, предоставляемая оператором мобильной связи, сформирована в виде таблицы, которая содержит название района (прокси места въезда), страну (с SIM-карты / MCC) и количество туристов/посетителей. Затем НСУ Индонезии сравнивает статистику с данными иммиграционного и пограничного обследования (если доступно), и разница добавляется к данным, или данные мобильного позиционирования используются всякий раз, когда нет данных иммиграционной службы или нет обследования на границе.

Сроки

Таблица 0.1 Сроки выполнения проекта в 2016-2017 гг.

Мероприятие	Время
Встреча по использованию данных мобильного позиционирования для туристических данных на границе	Июль 2016
Встреча по принятию решения о сотрудничестве с оператором мобильной связи	Август 2016
Обсуждение методологии и представление калибровки данных	Сентябрь 2016
Принятие решений по поводу методологии и пограничной зоны, где будут использоваться данные мобильного позиционирования	Октябрь 2016
Внедрение данных мобильного позиционирования в разработку данных за октябрь	Ноябрь 2016
Обсуждение и составление проекта Меморандума о взаимопонимании между НСУ- Министерством туризма и министерством связи и информации (на основе закона о статистике № 16/1997)	Декабрь 2016
Меморандум о взаимопонимании между НСУ – Министерством туризма и Министерством связи и информации (на основе закона о статистике №16/1997) подписан	Январь 2017
Представление методологии и внедрения данных мобильного позиционирования на Форуме статистического общества	Февраль 2017
Инициирование и планирование проведения обследования использования мобильных телефонов на границе	Март 2017

Семинар по использованию данных мобильного позиционирования (Участники: статистические бюро провинций и округов, Министерство планирования, Министерство туризма)	Апрель 2017
Подготовка обследования по использованию мобильных телефонов (обучение интервьюеров и пр.)	Май 2017
Проведение обследования по использованию мобильных телефонов (полевая работа)	Июнь 2017

Таблица 0.2 Будущие планы (Что дальше?)

Мероприятие	Время
Обсуждение и составление проекта Соглашения о техническом партнерстве для доступа к данным мобильного позиционирования	2017
Подписание Соглашения о техническом партнерстве для доступа к данным мобильного позиционирования	2017
Технический семинар/консультации по данным мобильного позиционирования	Август 2017
Разработка методологического руководства по использованию данных мобильного позиционирования для целей статистики туризма	2017
Меморандум о взаимопонимании с одним оператором мобильной связи	2017
Технический семинар/консультации по обработке данных	Октябрь 2017
Подготовка сервера, сети и пр. для обмена данными	2017
Разработка системы обеспечения качества (QAF) для данных мобильного позиционирования	2017

Преимущества и недостатки данных мобильного позиционирования

Данные мобильного позиционирования считаются одним из самых многообещающих источников данных ИКТ (информационно-коммуникационных технологий) для измерения передвижений людей, включая передвижения туристов. Почти у каждого человека на планете теперь есть мобильный телефон, который он использует для общения, доступа в Интернет, повседневных банковских операций и развлечений. Цифровой след, оставленный пользователями, очень чувствителен, но также очень ценен, поскольку предоставляет новые возможности для измерения и мониторинга деятельности населения в пространственно-временном аспекте. Для статистических целей данные мобильного позиционирования предоставляют новые возможности с точки зрения качества данных, а также новые перспективы. Сбор статистики, основанной на больших данных, то есть на данных мобильного позиционирования, может осуществляться автоматически, в некоторых случаях почти в режиме реального времени, и требует меньше ручного труда. Безусловно, задача анализа и интерпретации статистических показателей остается за статистиками и исследователями, но новая концепция быстрого и широкого сбора данных должна улучшить качество процесса принятия решений и результатов в государственном и частном секторах.

Чтобы получить оптимальные преимущества от технологии мобильного позиционирования, нам также необходимо понимать, что у этой технологии также есть свои сильные и слабые стороны. Например, с одной стороны, данные мобильного позиционирования имеют довольно хорошую согласованность во времени в отношении количества поездок и ночевок по сравнению с данными, основанными на «традиционных» методах. Их использование также улучшает своевременность статистики (почти до реального времени) и возможность использовать мобильные данные в качестве неподтвержденных быстрых индикаторов. Они также более точные, поскольку обследование проводится только в течение одной недели для оценки данных за месяц. Однако при пиковом или низком уровне количества въездных туристов возникнет проблема с оценкой. Кроме того, мы также можем столкнуться с трудностями при оценке качества статистики на основе данных мобильного позиционирования, потому что использование мобильных телефонов во время путешествий в значительной степени неизвестно, а также в том случае, если методология не является устойчивой. Также может быть относительный недостаток информации о цели поездки, расходах, типе размещения и используемых транспортных средствах.

НСУ Индонезии должно принять во внимание множество вопросов, чтобы иметь

возможность производить ценные и высококачественные статистические данные на основе нового источника данных. Основная проблема — это доступ к данным, где необходимо преодолеть правовые, административные и коммерческие барьеры.

Что касается юридической основы, то НСУ Индонезии имеет Закон о статистике (Закон о статистике № 16/1997), в котором говорится о сотрудничестве НСУ и отраслевых министерств или ведомств. Поэтому, чтобы обеспечить непрерывность данных и доступ к данным мобильного позиционирования, НСУ Индонезии подписывает Меморандум о взаимопонимании с Министерством туризма и Министерством связи и информации. Затем будут подписаны соглашения о техническом партнерстве для доступа к данным.

Технические вопросы

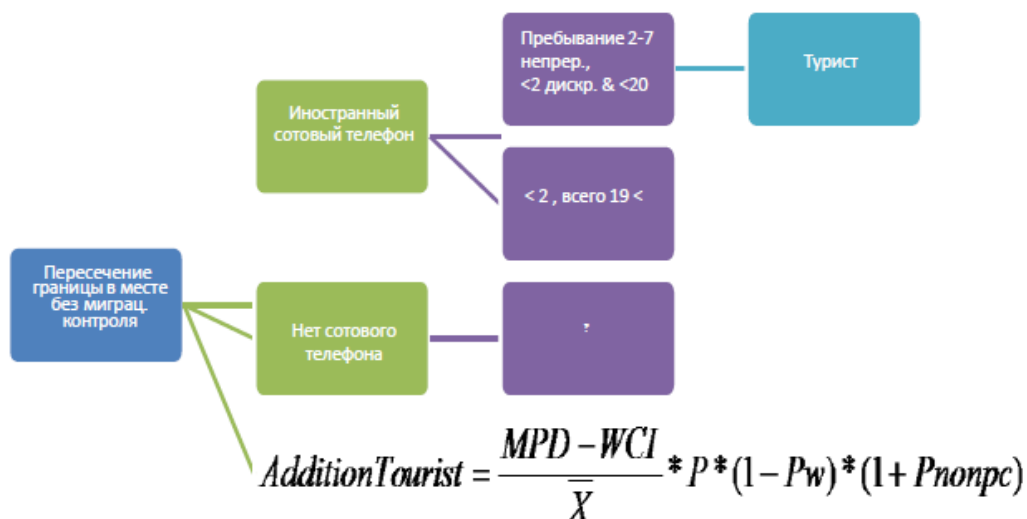
Несмотря на то, что данные мобильного позиционирования были внедрены, по-прежнему существуют некоторые проблемы с качеством данных. Некоторые основные технические проблемы таковы:

- **Жители Индонезии** могут использовать иностранные SIM-карты при пересечении границы и могут быть идентифицированы как иностранные туристы (основной пример — Малайзия). И наоборот, жители Малайзии могут использовать индонезийскую SIM-карту. У пограничных пунктов есть продавцы, которые продают сим-карты обеих стран (Индонезия и Малайзия).
- **Белое пятно:** существует вероятность того, что посетитель, который уже был в стране за несколько дней до въезда через иммиграционный пункт, исчезает (мобильный телефон выключен). Затем сигнал улавливается антеннами на границе (возможен двойной учет).
- **MCC/MNC:** неправильные коды страны из-за инвестиций (один из примеров: в Восточном Тиморе очень много вьетнамских номеров, хотя возможно, что ими пользуются не вьетнамцы, а восточные тиморцы).

Улучшения

Проведенные улучшения:

- **Обследование использования мобильных телефонов:** чтобы исправить данные об использовании мобильных телефонов и SIM-карт, некоторые вопросы по этому поводу были добавлены в обследование при пересечении границы. Это обследование позволит получить результат по следующей формуле:



Где:

Additional Tourist – дополнительный турист

MPD = количество сигналов иностранных сотовых телефонов

WCI = количество туристов, проходящих через иммиграционный пост

$\bar{X}$ = среднее количество сотовых телефонов, активно используемых туристом

P = доля иностранных путешественников

P_w = доля иностранных путешественников с целью работы или учебы

P_{nonpc} = доля иностранных путешественников без мобильного телефона

- **Подготовка методологического справочника:** следует разработать методологический справочник по использованию данных мобильного позиционирования для статистики туризма, а затем – по использованию данных мобильного позиционирования для других областей официальной статистики.
- **Подготовка отчета об обеспечении качества:** следует подготовить отчет об обеспечении качества использования данных мобильного позиционирования для целей статистики туризма.
- **Меморандум о взаимопонимании с оператором мобильной связи:** Меморандум о взаимопонимании между НСУ и оператором мобильной связи с самой большой долей рынка.
- **Соглашение о техническом партнерстве** между НСУ Индонезии, министерством туризма и министерством связи и информации для получения данных мобильного позиционирования от других операторов мобильной связи.