Использование данных системы мобильной связи для оценки мобильности населения. Оценка мобильности городского населения и потоков между городами с использованием больших данных в рамках интегрированного подхода.

Барбара Фурлетти (Barbara Furletti), Лоренцо Габриэлли (Lorenzo Gabrielli), Джузеппе Гарофало (Giuseppe Garofalo), Фоска Джанотти (Fosca Giannotti), Летиция Милли (Letizia Milli), Мирко Нанни (Mirco Nanni), Дино Педрески (Dino Pedreschi), Роберта Вивио (Roberta Vivio).

Перевод: Статкомитет СНГ

Аннотация. Большие данные, извлекаемые из цифровых следов человеческой деятельности, которые воспринимались как побочный продукт технологий, используемых нами в повседневной жизни, позволяют вести наблюдение за коллективным и индивидуальным поведением людей на беспрецедентном уровне детализации. У многих сторон нашей социальной жизни существуют «замещающие переменные» из области больших данных, например можно использовать вызовы с мобильных телефонов для оценки мобильности населения. В данной работе мы исследуем то, до какой степени такие «большие данные» в сочетании с административными данными могут использоваться для получения надежных и своевременных оценок передвижений людей между городами. Это исследование было проведено совместно Институтом статистики Италии (ИСТАТ), Национальным исследовательским советом (CNR) и Университетом Пизы в рамках работы Исследовательской комиссии по развитию работы ИСТАТа в области больших данных (*"Commssione di studio avente il compito di orientare le scelte dell Istat sul tema dei Big Data"*). В рамках текущего проекта ИСТАТа, который называется «Люди и места» (Persons and Places), была составлена первая версия матрицы «Начало – Пункт назначения поездок» (O/D matrix) на муниципальном уровне. При этом были использованы административные данные из разных источников, и предполагалось, что места проживания и места работы (или учебы)

являются конечными пунктами обычной индивидуальной миграции, связанной с работой или учебой. Если для лиц город проживания и город работы (или учебы) совпадают, то считается, что отсутствует миграция таких лиц между городами (мы называем их статичными резидентами). В противоположном случае считается, что наблюдается мобильность (лицо является динамичным резидентом, маятниковым мигрантом или включенным лицом). Административные данные не содержат информации о частоте передвижений, идея состоит в том, чтобы определить и оценить метод, где в качестве поддержки используются данные о телефонных вызовах, для определения в каждом муниципальном образовании численности статичных резидентов, включенных городских пользователей и ежедневных городских пользователей (маятниковых мигрантов).

Ключевые слова: большие данные, городское население, мобильность между городами, извлечение данных

Дино Педрески (Dino Pedreschi), Университет Пизы, Пиза, Италия e-mail: pedre@di.unipi.it

Джузеппе Гарофало (Giuseppe Garofalo), Роберта Вивио (Roberta Vivio)
ИСТАТ, Рим, Италия e-mail: surname@istat.it

Барбара Фурлетти (Barbara Furletti), Лоренцо Габриэлли (Lorenzo Gabrielli), Фоска Джанотти (Fosca Giannotti), Летиция Милли (Letizia Milli), Мирко Нанни (Mirco Nanni)
KDDLAB ISTI CNR, Пиза, Италия e-mail: name.surname@isti.cnr.it

1. Вступление

Мобильные телефоны представляют собой важный источник информации для изучения поведения людей, мониторинга окружающей среды, изучения транспортных потоков, социальных сетей и бизнеса. Интерес к использованию данных, получаемых от мобильных телефонов, растет довольно быстро, в частности, благодаря развитию и распространению телефонов с большим количеством сложных функций.

Наличие таких данных стимулировало проведение исследований для разработки алгоритмов извлечения данных (Data mining) о действиях для изучения привычек людей, схем мобильности, мониторинга окружающей среды и определения и предсказания событий. Некоторыми примерами являются: выявление социальных отношений, изучаемых в работе [1], где было отмечено существование корреляции между сходством передвижений индивидов и их близостью в социальных сетях; построение таблиц «начало – пункт назначения» поездок для использования в транспортных моделях [2]; и анализ того, как посетители крупной туристической зоны используют территорию, на основе информации GSM роуминга (пользователи, прибывающие из других стран) с особым акцентом на посещения достопримечательностей [3]. Для целей извлечения данных, GSM данные оказались существенными с точки зрения размера и репрезентативности выборки. Вообще говоря, наличие информации о локализации или поведении людей или движущихся единиц позволяет создать инструменты для поддержки исследований в нескольких областях, таких как здравоохранение, координация социальных групп, транспорт и туризм.

В этой работе мы предлагаем и опробуем процесс анализа, построенный с использованием, так называемого *Социометра* (*Sociometer)* – инструмента для извлечения данных (Data mining) для классификации пользователей на основе их привычек в осуществлении телефонных звонков. Первый прототип *Социометра* был разработан в рамках проекта «Наблюдение за потоками туристов – Пиза», целью которого была оценка присутствия разных категорий людей в городе [5]. Проект, который выполнялся в сотрудничестве с муниципалитетом г. Пиза, был нацелен на изучение потоков туристов, посещающих город, для получения оценки общего качества системы приема туристов на данной территории и для установки системы постоянного мониторинга. *Социометр* был

опробован с положительными результатами для оценки реальной ситуации в Пизе и в Козенце [6].

В данной работе мы расширяем использование основного метода для работы с большей территорией и для включения потоков между разными территориальными единицами (в данном исследовании – муниципальными образованиями), тогда как *Социометр* используется для оценки присутствия в одной области. Цель состоит в том, чтобы сформировать статистические данные, сопоставимые с получаемыми в рамках текущего проекта в ИСТАТе, который называется «Люди и места», где места проживания и потоки людей изучаются с использованием административных источников данных. Если мы добьемся успеха по данному направлению, то это будет означать возможность безопасного объединения существующей статистики численности и миграционных потоков с постоянно обновляющимися оценками, полученными на основе GSM данных, и таким образом будет сделан первый шаг в направлении использования *больших данных* в официальной статистике.

2 Цели и условия эксперимента

Цель этой работы состоит в разработке больших объемов постоянно обновляемой информации, содержащейся в данных о вызовах с мобильных телефонов, для получения оценок статистики населения, касающихся места проживания и мобильности. В данном разделе мы сначала опишем, какую информацию содержат данные о вызовах, и подробно остановимся на массиве данных, использованном в экспериментах. Затем мы введем категории пользователей и показатели мобильности, которые мы намереваемся получить на основе подробных записей о вызовах (call detail records – CDR).

2.1 Подробные записи о вызовах *(CDR)*

Сеть GSM (Глобальная система мобильной связи) - это сеть, которая обеспечивает связь между мобильными устройствами. Протокол GSM основан на так называемой сотовой сетевой архитектуре, где географический район покрывается некоторым количеством антенн, излучающих сигналы, принимаемые мобильными устройствами. Каждая антенна покрывает район, называемый сотой. Таким образом, покрываемая область разделена на

некоторое количество, возможно, пересекающихся сот, однозначно определяемых антенной. Горизонтальные радиусы сот могут быть различны в зависимости от высоты антенны, коэффициента усиления, плотности населения и условий распространения сигнала и могут составлять от пары сотен метров до нескольких десятков километров.

Подробные записи о вызовах (CDR) это данные журнала, документирующего каждый вызов с мобильного телефона, которые операторы используют для целей биллинга. В данной работе эти данные используются в следующем формате: <*Timestamp,Caller_id,d,Cell_*1, *Cell_*2>, где *Caller_id* - анонимный идентификатор пользователя, который сделал вызов, *Timestamp* – начальное время вызова, *d* - продолжительность, *Cell_1* и *Cell_2* – идентификаторы сот, где начался и закончился вызов.

Массив данных, использованных в данной работе, состоит из 7,8 миллионов подробных записей о вызовах, полученных с 9 января по 8 февраля 2012 года. Массив данных содержит записи о вызовах, осуществленных примерно 232 тысячами пользователей - держателей национальных контрактов мобильной связи (не включены пользователи услуг роуминга).

Следует отметить, что основное ограничение в использовании данных CDR состоит в том, что локализация происходит только во время телефонного вызова, что может давать неполное представление о мобильности пользователей. Мы обсуждаем этот вопрос в разделе 3, где приводим сложную методологию для обработки данных и частично преодолеваем проблему неполноты.

2.2 *Категории пользователей и ИК-матрица* (*O/D matrix)*

В данной работе мы изучаем территориальные единицы на уровне муниципальных образований. Мы рассматриваем 39 муниципальных образований в провинции Пиза, в Тоскане. Количество жителей в этих муниципальных образованиях сильно различается: от менее одной тысячи человек в мелких образованиях до примерно 86 тысяч человек в центральном муниципальном образовании Пизы; при этом средняя численность жителей муниципального образования составляет примерно 10 000 чел. Каждое муниципальное образование покрыто в среднем тремя-четырьмя GSM антеннами.

Первая задача данной работы состоит в том, чтобы для каждого муниципального образования корректно оценить население, принадлежащее одной из указанных ниже категорий, которые уже рассчитываются ИСТАТом на основе административных данных в рамках текущего проекта «Люди и места»:

- **Стационарный резидент в пункте А:** резиденты, которые имеют формальное место жительства и место работы (учебы) в муниципальном образовании А или не работают (учатся).
- **Включенный городской пользователь в пункте А:** люди, которые проводят долгие периоды времени работая (учась) в муниципальном образовании А (например, большую часть недели), хотя формально являются резидентами другого муниципального образования.
- **Ежедневные городские пользователи в пункте А:** люди, которые ежедневно приезжают в муниципальное образование А, при этом имея формальное место жительства в другом муниципальном образовании**.**

Каждая категория характеризуется разным образом жизни на определенной территории и, соответственно, разным использованием ее ресурсов.

В рамках проекта ИСТАТа была построена первая версия матрицы O/D (Origin Destination matrix – ИК-матрица, «исходный-конечный пункты поездки») на муниципальном уровне в предположении, что место проживания и место работы (или учебы) являются конечными пунктами обычных индивидуальных передвижений на работу или учебу. Совпадение города проживания и города работы (или учебы) считается замещающей переменной для отсутствия меж городской мобильности лица (мы определяем такое лицо как «статичного резидента»). Противоположный случай является замещающей переменной для наличия мобильности (такое лицо определяется, как динамичный резидент, маятниковый мигрант или «включенное лицо»). Между маятниковыми мигрантами и «включенными» лицами различий не проводится.

Как мы покажем ниже, процесс анализа, разработанный нами для данных GSM, позволяет делать выводы о немного различающихся категориях пользователей. В частности,

Стационарный резидент и *Включенный городской пользователь* пока не различаются, и статистика предоставляет агрегированные данные по ним.

3 Методология

Основное положение, которого мы придерживаемся в данной работе, состоит в том, что категория лица в рамках конкретного муниципального образования может быть определена на основании временного распределения его присутствия в данной области. Например, люди, совершающие ежедневные поездки в муниципальное образование на работу, обычно бывают там только в рабочие часы и в будние дни, и ожидается, что исключения являются редкими. С другой стороны, как уже отмечалось, данные CDR могут описать местоположение пользователей только частично, поэтому распределения присутствия, которые мы можем построить на основе этих данных, обычно являются заниженными оценками реальных распределений. По этой причине после рассмотрения профилей индивидуальных вызовов (пункт 3.1), представляющих такие неполные распределения присутствия, и ожидаемых типичных распределений присутствия для наших основных категорий пользователей (пункт 3.2), мы рассмотрим метод полуавтоматической классификации профилей вызовов (пункт 3.3). Эта процедура состоит в том, что эксперта просят вручную промаркировать небольшое количество представительных профилей вызовов, которые затем используются для автоматической маркировки всех остальных профилей вызовов.

3.1 Индивидуальные профили вызовов (*Individual call profiles – ICP*)

Индивидуальные профили вызовов представляют собой агрегированные пространственно-временные профили пользователя, рассчитанные с применением правил пространственного и временного анализа в отношении первичных данных – подробных записей о вызовах (CDR). Структура представляет собой матрицу такого типа, как показано на рис. 1. Временное агрегирование проводится по неделям с разбивкой на рабочие и

выходные дни. Например, если мы рассмотрим период, равный 28 дням (4 недели), то матрица будет иметь 8 столбцов (два столбца на каждую неделю – будние и выходные дни).

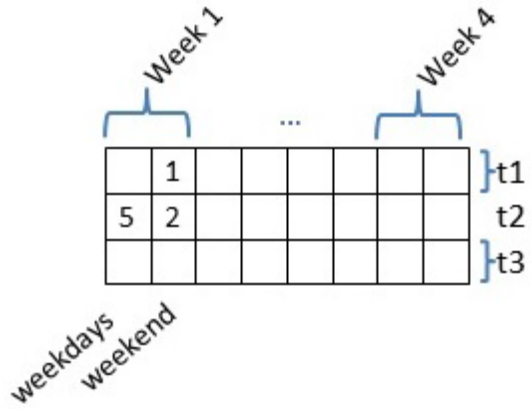


Рис. 1 Индивидуальный профиль вызовов

Дальнейшая временная разбивка проводится по часам, что добавляет новые строки в матрицу. У нас три временных отрезка (t1, t2, t3), так что количество строк в матрице равно трем. Числа в матрице представляют собой количество действий (в данном случае присутствие пользователя), выполненных пользователем в конкретный период и временной отрезок. Например, число 5 на рис. 1 означает, что пользователь присутствовал в интересующей нас области в течение 5 будних дней в неделе 1 и только в отрезок времени t2.

3.2 Категории профилей и шаблоны осуществления вызовов

Индивидуальные профили вызовов (ICP), рассчитанные выше, обеспечивают синтез присутствия пользователей, что позволяет довольно легко охарактеризовать некоторые конкретные категории. В частности, в данной работе мы рассматриваем четыре следующие категории:

- **Резиденты (**или **статичные резиденты)** – это те лица, которые живут и работают в одном районе и их присутствие является значимым для всех дней и всех временных отрезков для конкретного муниципального образования.

- **Динамичные резиденты** – люди, проживающие в муниципальном образовании (A), но работающие в другом месте (B). Их присутствие в пункте A ожидается значимым всегда, за исключением рабочих дней и рабочих часов (интервал t2).

- **Маятниковые мигранты** – люди, проживающие в некотором муниципальном образовании (B) (динамичные резиденты), чье место работы или учебы находится в пункте A. Ожидается, что их присутствие в пункте A почти полностью приходится на рабочие дни и рабочие часы (отрезок времени t2).

- **Посетители:** люди, которые посещают муниципальное образование только однажды или небольшое количество раз.

Если сравнить эти определения с категориями, представленными в пункте 2.2 и принятыми в проекте «Люди и места» ИСТАТа, то следует отметить, что отсутствие административных данных о пользователях GSM не позволяет провести различия между такими категориями как *Стационарный резидент* и *Включенный городской пользователь*, поскольку на практике их физическое присутствие в месте проживания/включенности, как правило, идентично. С другой стороны, физическое присутствие пользователей позволяет легко отличить (по крайней мере, в принципе) динамичных резидентов от статичных, поскольку первые обычно не присутствуют в муниципальном образовании своего проживания в рабочее время. Эти небольшие нестыковки между двумя классификациями будут обсуждены позже, когда мы будем сравнивать оценки населения, полученные с использованием двух методов: на базе официальных данных и на базе данных GSM.

Используя четыре приведенные здесь категории можно сформулировать правила классификации, которые при наличии данных об индивидуальном профиле вызовов (ICP) автоматически помещают пользователя в соответствующую категорию. К сожалению, данные об индивидуальных профилях вызовов подвержены искажениям, происходящим из двух источников: (i) ожидаемое распределение присутствий для какой-то категории не всегда точно соответствует реальным присутствиям, хотя отклонения должны быть относительно малы; (ii) индивидуальные профили вызовов представляют только выборку фактических присутствий, поскольку выявляются только те присутствия, которые соответствуют хотя бы одному телефонному вызову. Последнее обстоятельство приводит к значительным

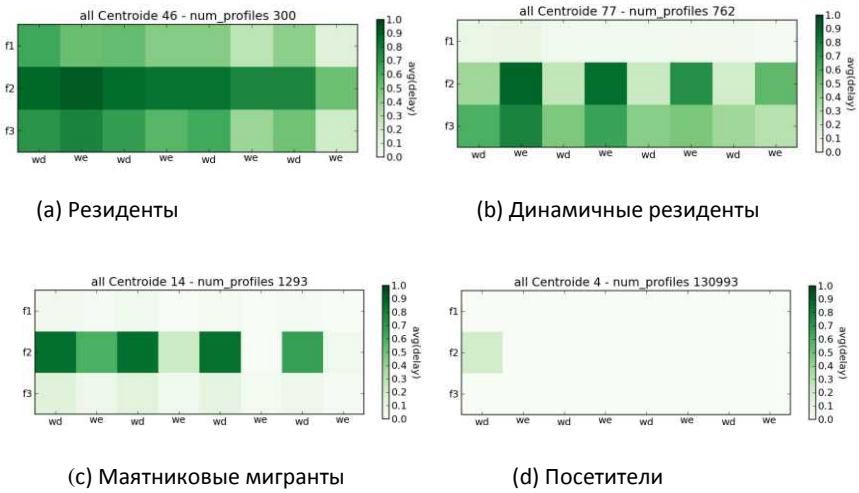
возмущениям в распределении присутствий, которые не могут быть легко устранены при помощи каких-либо процедур фильтрации или масштабирования. Эмпирическая оценка показала, что если пытаться привести индивидуальные профили вызовов в соответствие с идеальным распределением присутствий, то результаты будут очень плохими. Для того чтобы справиться с изменчивостью, присутствующей в данных об индивидуальных профилях вызовов (ICP), мы разработали полуавтоматическую процедуру, описанную в следующем пункте.

3.3 Классификация профилей

Предлагаемый нами метод классификации состоит из двух частей. Сначала мы извлекаем репрезентативные профили вызовов (Representative call profiles - RCP), то есть относительное небольшое множество синтетических профилей, где каждый является агрегатом однородного множества реальных индивидуальных профилей вызовов (ICP). При помощи этого шага сокращается множество выборок для классификации, которая затем может быть проведена вручную экспертом на основе приведенных выше определений категорий и собственного опыта и суждений эксперта. Затем метки, присвоенные репрезентативным профилям, распространяют на все множество ICP.

На первом этапе использовался стандартный метод k-средних для разбивки *n* профилей индивидуальных вызовов на *k* гомогенных кластеров, где среднее значение ICP, принадлежащих каждому кластеру, служит прототипом/представителем кластера. Алгоритм выполняется итеративно. Вначале создаются *k* случайных разбиений, затем рассчитываются центроиды каждой группы и создается новое разбиение, где каждый объект (ICP) попадает в тот кластер, чей центроид ближе всего к нему.

Наконец центроиды заново вычисляются для нового кластера и процедура повторяется итеративно, пока алгоритм не достигает стабильного состояния (сходится). Близость двух ICP, что является основной операцией метода k-средних, рассчитывается здесь при помощи обычного евклидового расстояния, то есть путем сравнения каждой пары соответствующих отрезков времени для двух сравниваемых ICP.



(a) Резиденты (b) Динамичные резиденты

(c) Маятниковые мигранты (d) Посетители

Рис. 2 Примеры маркированных RCP

Центроид кластера получают, рассчитывая для каждого отрезка времени среднее значение соответствующих величин ICP для данного кластера. Выбор параметра *K* был осуществлен на основе большого количества экспериментов, где производятся попытки минимизировать расстояние внутри кластера и максимизировать расстояние между кластерами. Наиболее подходящим значением оказалось *K*=100. Когда представительные профили вызовов (RCP) были извлечены, они были отмечены экспертами в данной области при согласовании с теми экспертами, о которых говорилось в пункте 3.2. На рис. 2 приведены некоторые примеры RCP, полученные на основе реальных данных – одно значение для каждой категории пользователей. Более темные (и соответственно более светлые) цвета представляют более высокие (и соответственно, низкие) частоты.

На втором шаге, то есть при распространении вручную поставленных RCP меток, использовался стандартный шаг классификации с использованием только одного ближайшего соседа (алгоритм 1-NN). Это соответствует присвоению каждому ICP метки ближайшего RCP. Расширения для решения могут быть легко получены, если принять K-NN классификацию, где, K>1 и выбирать метку большинства.

4 Оценка

В данном разделе мы обобщаем экспериментальные результаты, полученные при расчете некоторых статистических показателей численности и потоков населения в провинции Пиза в Италии и при сравнении их с аналогичными оценками, полученными на основе официальных данных в рамках проекта ИСТАТа «Люди и места». GSM данные для

нашего исследования были получены от одного оператора мобильной связи, и поэтому, они не охватывают всех пользователей мобильных телефонов на данной территории. По этой причине для того, чтобы добиться сопоставимости наборов данных GSM и данных, используемых ИСТАТом, мы провели масштабирование наших результатов для всех показателей, учитывая рыночную долю оператора в каждом рассматриваемом муниципальном образовании.

Ниже для каждого из 39 муниципальных образований провинции Пиза мы приводим оценки численности резидентов, динамичных резидентов и систематических потоков (маятниковой миграции) и проверяем корреляцию с административными данными. Категория «Посетители» здесь не рассматривалась, поскольку было невозможно получить соответствующую статистику из официальных источников данных, то есть на данном этапе здесь невозможно провести сравнение.

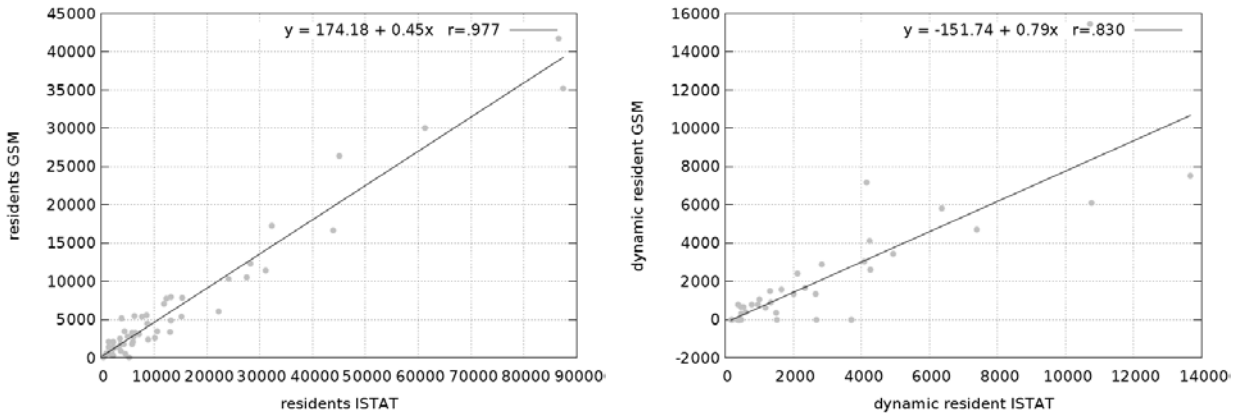
4.1 Резиденты

Категория резидентов при работе с данными GSM – это статичные резиденты, то есть те лица, которые постоянно проживают или живут в связи с работой или учебой в данном муниципальном образовании в период наблюдения. Мы сравниваем эти данные с теми, кто имеет зарегистрированное место жительства в данном муниципальном образовании (ИСТАТ).

На рис. 3a для каждого муниципального образования представлены соответствующие величины, полученные на основе данных ИСТАТа (горизонтальная ось) и данных GSM (вертикальная ось). На графике визуально видна ясная корреляция между двумя переменными, что подтверждается высоким значением *R* статистики – $R = 0.977$ ($R^2 = 0.955$). Это подтверждает то, что наш метод позволяет получить хорошую оценку численности *Резидентов* на этой территории.

На рис. 3b приведены подобные оценки для *Динамичных резидентов*. Эти значения сравниваются с оценками численности тех, кто не работает или учится в том же муниципальном образовании, где проживает. Как мы видим, наш метод, основанный на данных GSM, напрямую указывает на наличие статуса *Динамичного резидента* у пользователя. График и величина R-статистики для этого случая ($R = 0.830$, $R^2 = 0.689$) показывают, что результаты все еще достаточно хороши, хотя и менее точные, чем в

предыдущем случае. Среди возможных причин этого мы выделяем фактическое отсутствие данных для малых муниципальных образований, которые можно будет добавить в следующий раз. Другая причина состоит не в полной сопоставимости двух классификаций пользователей в городах, полученных на основе административных данных и данных GSM.



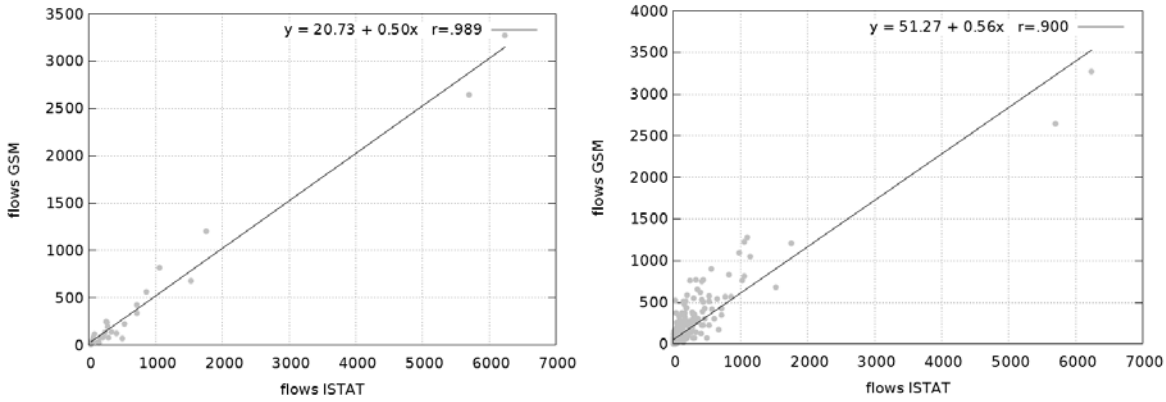
(a) Резиденты (b) Динамичные резиденты

Рис. 3 Корреляция между Резидентами и Динамичными резидентами по данным GSM и ИСТАТ

4.2 Потоки маятниковой миграции

При проведении переписи ИСТАТ спрашивает, в каких муниципальных образованиях проживают и работают респонденты. В нашем случае мы оцениваем потоки из дома на работу, выбирая для каждого лица пару муниципальных образований, в которых это лицо считается, соответственно, резидентом и маятниковым мигрантом. На рис. 4 (a) показано, что, несмотря на достаточно хорошее значение R-статистики ($R = 0.900$, $R^2 = 0.810$), существует значительное количество пар, для которых оценки потока не очень точны. Скорее всего, это опять-таки происходит из-за того, что для многих мелких муниципальных образований в нашем массиве данных отсутствует информация. Однако если рассматривать только потоки в сторону Пизы, показанные на рис. 4 (b), мы видим, что оценки улучшаются (теперь $R = 0.989$ и $R^2 = 0.978$). Это можно объяснить тем, что муниципальное образование Пиза является центром притяжения в регионе, и поэтому набор соответствующих выборочных потоков значительно больше. В целом, можно отметить, что оценки потоков,

полученные нашим методом, являются более точными для более крупных городов, поскольку те обычно привлекают более систематические потоки.



(a) Все потоки (b) Пиза

Рис. 4 Корреляция между систематическими потоками, измеренными ИСТАТ и Социометром

5 Выводы

В данной работе мы рассмотрели метод оценки численности и потоков населения на основе *больших данных*, извлеченных из системы мобильной связи, которые используются здесь в качестве замещающих переменных для присутствия и мобильности индивидов. Полученные результаты в целом являются обнадеживающими, а для некоторых конкретных показателей очень точными по сравнению с аналогичными статистическими оценками, полученными на основе официальных данных.

Описанная здесь забота была проведена в рамках продолжающегося проекта, и мы планируем произвести некоторые усовершенствования по следующим направлениям: (i) принять более эффективные методы кластеризации для извлечения репрезентативных протоколов вызовов (RCP); (ii) построить классификацию вокруг пользователей, а не вокруг индивидуальных профилей вызовов (ICP), то есть классифицировать все индивидуальные профили вызовов пользователя вместе, используя отношения и зависимости, которые существуют между ними, например каждый пользователь должен иметь только один район проживания (кроме явных исключений); (iii) расширить территорию эксперимента, чтобы

увеличить выборку охваченного населения и избежать *пограничных эффектов*, связанных с входящими/исходящими потоками за пределами нашей зоны исследования.

Литература

1. Wang, D., Pedreschi, D., Song, C., Giannotti, F., and Barabasi, A.-L. *Human mobility, social ties, and link prediction.* In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD 11. ACM, New York, NY. 2011.
2. Nanni, M., Trasarti, R., Furletti, B., Gabrielli, L., Mede, P. V. D., Bruijn, J. D., Romph, E. D., and Bruil, G. MP4-A project: Mobility planning for Africa. In D4D Challenge @ 3rd Conf. on the Analysis of Mobile Phone datasets (NetMob 2013). 2013.
3. Oltenau, A.-M., Trasarti, R., Couronne, T., Giannotti, F., Nanni, M., Smoreda, Z., and Ziem-licki, C. GSM data analysis for tourism application In Proceedings of 7th International Symposium on Spatial Data Quality (ISSDQ). 2011.
4. F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, R. Trasarti Unveiling the complexity of human mobility by querying and mining massive trajectory data. The VLDB Journal, 2011
5. B. Furletti, L. Gabrielli, C. Renso, S. Rinzivillo Turism fluxes observatory: deriving mobility indicators from GSM calls habits In the Book of Abstracts of NetMob 2013
6. B. Furletti, L. Gabrielli, C. Renso, S. Rinzivillo. Analysis of GSM calls data for understanding user mobility behavior In the Proceedings of Big Data 2013