УДК 004.722 **DOI** 10.15622/ia.21.4.5

М.В. Иванов, А.А. Полунин

ПОВЫШЕНИЕ ТОЧНОСТИ IP-ГЕОЛОКАЦИИ НА ОСНОВЕ ДАННЫХ, ПРЕДОСТАВЛЯЕМЫХ ОТКРЫТЫМИ IP-ГЕОСЕРВИСАМИ

Иванов М.В., Полунин А.А. **Повышение точности IP-геолокации на основе данных, предоставляемых открытыми IP-геосервисами.**

Аннотация. IP-геолокация – это процесс определения реального географического положения электронного устройства, подключенного к сети Интернет, по его глобальному сетевому адресу [1]. В настоящее время она нашла широкое применение в интернет-торговле, маркетинге и рекламе, информационной безопасности [2] и других направлениях человеческой деятельности. Применяются различные подходы к определению местоположения удаленного сетевого устройства, различающиеся как по типу анализируемой информации (задержка передачи пакетов, ресурсные записи DNS-серверов, контент веб-страниц), так и по выдаваемому результату (название страны или города, почтовый адрес, вероятная зона расположения или точные координаты) [3, 4]. Ошибка IP-геолокации зависит от страны расположения устройства, плотности населения, типа сетевого устройства и лежит в пределах от нескольких десятков метров до сотен километров. При этом для одних и тех же входных данных результаты разных IP-геосервисов могут различаться значительно. Объектом данного исследования выступают общедоступные IP-геосервисы, предоставляющие услуги по IP-геопривязке узлов глобальной сети на основе их IP-адресов, а именно – их точность и полнота. Выборка IP-геосервисов для тестирования были сформирована из числа наиболее популярных [5]. При проведении исследования результаты IP-геолокации сравнивались с достоверными сведениями о расположении некоторых IP-адресов, в качестве показателей точности использовались страна, город и географические координаты. На основе сравнительного анализа результатов тестирования были сделаны выводы о точности IP-геосервисов по выбранным показателям, их существенных свойствах, а также о зависимости ошибки геолокации от размера населенного пункта. Для повышения точности IP-геопривязки авторами предложен ансамблевый метод усреднения координат, полученных от нескольких IP-геосервисов.

Ключевые слова: сеть Интернет, IP-геолокация, IP-геосервисы, Atlas, IpAPI, Shodan.

1. Введение. В приложении к IP-адресам задача геолокации может быть рассмотрена как процесс сопоставления глобального сетевого адреса устройства с его географическим местоположением. В настоящее время для адресации в глобальной сети используется два адресных пространства: IPv4 и IPv6. Количество пользователей и, соответственно, сетевых устройств постоянно растет и стандарт IPv6 был разработан специально для решения данной проблемы. Однако сложности перехода к новой версии протокола маршрутизации пока что позволяют IPv4 оставаться, по факту, основным [6]. Сетевое пространство в данный момент целиком регламентируется и определяется Администрацией адресного пространства Интернет

(IANA), которая распределяет весь набор IP-адресов между 5 региональными интернет-регистраторами (RIR). В свою очередь каждый из них предоставляет свои подсети интернет-провайдерам различного уровня. Таким образом, адресное пространство в административном плане представляет собой четкую иерархическую структуру.

В то же время стандарты IP [7] не привязывают значение адреса к географическому положению. Верно и обратное — не существует общепринятого стандарта по распределению IP-адресов на основе геоположения узла. В результате два хоста, расположенные в одном доме, могут иметь IP-адреса из разнесенных областей адресного пространства (в качестве меры разноса допустимо применить взвешенное евклидово расстояние [8] — чем старше бит IP-адреса, тем он «важнее»). Противоположное утверждение также истинно – соседние IP-префиксы могут принадлежать операторам связи на разных континентах.

Сложившаяся система распределения адресных ресурсов привела к тому, что пространство IP-адресов, будучи спроецировано на поверхность Земли, не является гладким и представляет собой «лоскутное одеяло» с множеством точек разрыва.

Усложняют ситуацию и современные «облачные» подходы к назначению IP-адресов сетевым интерфейсам. Крупные сетевые операторы используют технологии распределения своей сети, такие как IP Anycast [9] и Content Delivery Network [10]. Зачастую они размещают свое сетевое оборудование, имеющее один и тот же IP-адрес, на разных континентах (рисунок 1). Такой подход позволяет учитывать региональные особенности и условия предоставления услуг, а также ускорить доступ к ресурсам. Примерами являются распределенные по всему миру серверы Google, Netflix, Amazon и других известных корпораций. С точки зрения IP-геолокации такой подход к построению сетей вносит некоторые сложности – один и тот же IP-адрес, используемый разными серверами, будет соотносится различными IP-геосервисами с различными местами его расположения и ошибка может достигать нескольких тысяч километров.

IP-геолокация предполагает определение географического расположения, привязку к странам, населенным пунктам или другим объектам. Основное применение полученных при геопривязке данных – сфера услуг и маркетинг. Зачастую веб-страницы предлагают рекламу товаров и автоматически устанавливают язык, а службы доставки товаров автоматически заполняют поля в формируемом заказе [11]. Как правило, эти данные получены на основании IP-

геолокации устройства, а не с помощью спутниковых или мобильных систем определения местоположения.

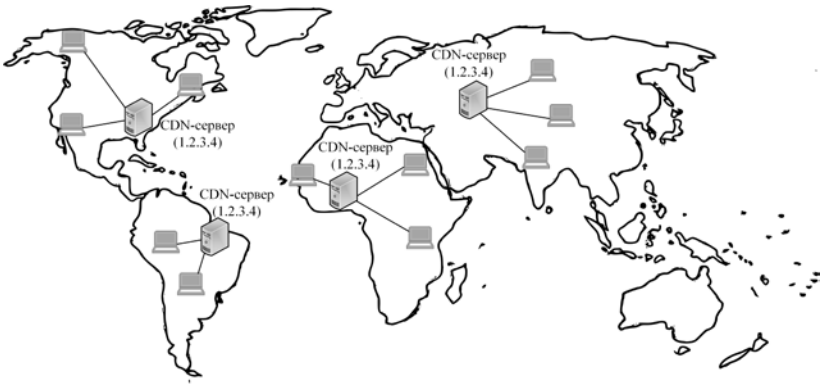

Рис. 1. Зоны обслуживания серверов при использовании технологии Content Delivery Network

IP-геопривязка используется и в технологии GeoDNS [12]. Она, также как и IP Anycast, позволяет распределить сетевую нагрузку, влияя на путь прохождения трафика. Однако в случае GeoDNS маршрутизация трафика осуществляется в зависимости от географической, а не логической удаленности адресата. Сервер GeoDNS содержит несколько A-записей для домена, каждая из которых соответствует одному IP-адресу согласно RFC 1035. Какая именно из A-записей попадет в ответ GeoDNS-сервера зависит от того, в какой географической локации расположен адресат – будет выбран ближайший сервер домена. Такой подход позволяет предлагать пользователям разных стран различный контент в зависимости от их предполагаемых потребностей и интересов, а также уменьшить сетевую задержку между клиентом и сервером.

Другой сферой применения IP-геолокации является информационная безопасность. Поскольку расположение абонентского устройства является одним из свойств его владельца [13, 14], оно используется многими приложениями в процессе многофакторной аутентификации и при значительных отличиях между текущими и обычными значениями координат может служить причиной отказа системы безопасности в доступе к аккаунту [15].

Современные методы IP-геолокации в основном используют метод измерения времени прохождения пакетов до исследуемого узла из известных точек и дальнейшему вычислению координат на их

основе [16]. Точность вычисленных значений может значительно отличаться в зависимости от многих факторов: количество узлов, с которых проводится исследование; их удаленность от определяемого устройства; вид линии связи, по которой проходит пакет. В ходе проведения исследования была обнаружена закономерность: в крупных городах местоположение определяется с точностью до района или улицы, в то время как в малонаселенных районах погрешность может составлять несколько сотен километров.

Однако существуют и другие способы IP-геолокации, основанные на анализе информации о провайдере целевого IP-адреса, а также содержимого веб-страниц. Также местоположение может быть определено по информации из ресурсных записей DNS сервера, которые часто содержат названия стран, городов или других известных мест, например, коды аэропортов [17, 18].

Важным источником данных о геопривязке IP-адресов служат личные данные пользователей онлайн-сервисов доставки. При заказе товара указывается точное местоположение покупателя, которое может быть сопоставлено с его сетевым адресом. Данный факт объясняет более точную геопривязку IP-адресов, находящихся в пользовании частных лиц, а не организаций.

IP-геопривязка конечных подсетей пользователей значительно отличается от той же задачи для подсетей интернет-провайдеров [19]. Это связано с большей географической сосредоточенностью устройств, используемых одной организацией (в пределах одного или нескольких зданий) или физическим лицом, с одной стороны, и распределением подсетей интернет-провайдеров по значительной территории – с другой. Как следствие – определение местоположения IP-адреса только на основании сведений о провайдере, которому он принадлежит, не представляется возможным.

Отдельные IP-геосервисы в зависимости от условий запроса могут давать результаты с высокой точностью, но показатели точности разных IP-геосервисов могут значительно отличаться на различных континентах, в разных странах и регионах. Таким образом, взаимное дополнение результатов IP-геолокации, полученных от различных источников, по мажоритарному принципу с учетом их сильных и слабых сторон позволит скомпенсировать ошибку и получить более точные результаты в среднем.

В соответствии с обозначенными проблемами, целью данного исследования является поиск направлений повышения точности IP-геолокации на основе данных, предоставляемых открытыми IP-геосервисами. IP-геолокация не является новым направлением

исследования глобальной сети Интернет, однако должное внимание стало уделяться данной проблеме сравнительно недавно. Большая часть работ в области IP-геолокации носит эмпирический характер, а количество теоретических материалов ограничено. Одним из перспективных направлений исследования является разработка ансамблевых методов уточнения координат, полученных от разных (в том числе и внешних в широком смысле) IP-геосервисов. Такие методы должны учитывать сильные и слабые стороны каждого из IP-геосервисов.

Задачи исследования: оценка полноты баз данных IP-геосервисов и точности IP-геолокации на уровне страны, на уровне города и на уровне географических координат; определение зависимости точности IP-геопривязки от размера населенного пункта, в котором расположен сетевой узел; разработка предложений по повышению точности IP-геолокации.

Для проведения исследования были отобраны несколько наиболее используемых [5] IP-геосервисов, предоставляющих услуги IP-геолокации. Анализ характеристик каждого из них проводился путем сравнения с достоверной информацией. В качестве априорных сведений были выбраны данные, полученных из системы измерения Интернет пространства Atlas [20]. Зонды данного сервиса имеют привязку к местоположению с точностью не хуже нескольких сотен метров, а их свойства доступны в открытом виде для зарегистрированных пользователей системы. Информация о географическом расположении задается и гарантируется обладателем зонда. Данные геопривязки, предоставленные различными IP-геосервисами, сравнивалась с информацией о зондах данного сервиса. На основе результатов тестирования была подсчитана и проанализирована статистика ошибок, позволяющая сформулировать выводы о точности рассмотренных IP-геосервисах.

2. Анализ области исследования. Основная часть работ по геолокации IP-адресов связана с рассмотрением различных методов определения местоположения. Как правило, они связаны с выработкой комплексных подходов, учитывающих как семантические данные (ресурсные записи серверов DNS, адреса и почтовые индексы, указанные на web-страницах), так и результаты измерений временных и пространственных метрик [21-23]. Полученные результаты сравниваются, в основном, с данными из коммерческой базы данных GeoIP2 [24]. Такой подход значительно снижает репрезентативность исследований за счет возможного наличия недостоверных сведений в самой GeoIP2, в то время как сравнение с достоверно известными

местоположениями выглядит более убедительно [25, 26]. Стоит отметить, что оценке существующих IP-геосервисов уделено недостаточно внимания, публикации по данной теме содержат устаревшие данные [26].

Известны работы [25-28], в которых проводится подробный анализ IP-геосервисов по методике, аналогичной использованной в данной работе, в них за основу сравнения также берутся данные о местоположении IP-адресов на уровне страны, города и координаты. Эти показатели позволяют описывать местоположение как с точки зрения крупных абстракций (страна и город), так и с точки зрения математических вычислений (координаты). Большинство IP-геосервисов приводят информацию именно в таком формате (использование данных о административном делении стран или почтовых индексах, которые также встречаются в ответах – затруднено по техническим причинам). Отличия данного исследования заключаются в (а) использовании информации о расположении зондов системы Atlas для оценивания точности IP-геосервисов, (б) исследовании зависимости качества IP-геолокации от размера населенного пункта и (в) разработке способа улучшения результатов IP-геолокации путем вычисления координат местоположения сетевого устройства с использованием информации от нескольких IP-геосервисов.

3. Используемые данные

3.1. Выбор IP-геосервисов. В качестве IP-геосервисов были выбраны популярные web-приложения, имеющие собственный API, что значительно упростило процесс сбора данных (Таблица 1).

Таблица 1. Сравнение характеристик популярных сервисов IP-геолокации

Название	Регистрация	Ограничения по количеству запросов (в бесплатной версии)	Стоимость платной подписки в месяц	Задержка [5, 29-31]
IpApi	Не требуется	45 запросов в минуту	€13.3/ месяц (количество запросов не ограничено)	Около 50 мс
IpWhois	Не требуется	10 тысяч запросов в месяц	$10 .99/ месяц (250 тысяч запросов в месяц)	Около 600 мс

Продолжение Таблицы 1

BigDataCloud	Требуется	10 тысяч запросов в месяц	$3 за каждые 10 тысяч запросов	Около 20 мс
BGPView	Не требуется	Не ограничено	-	Не указана
Shodan	Требуется		$49 - единоразово	Не указана
IpGeoLocation	Требуется	300 запросов в день	£4.99 / месяц (50 тысяч запросов в день)	Около 45 мс
Spott	Требуется	10 тысяч запросов в месяц	$10/ месяц (10 тысяч запросов)	Около 600 мс
IpSquads	Требуется	1000 запросов в месяц	$5.99/ месяц (50 тысяч запросов в месяц)	Около 350 мс
IpLocation	Требуется	1000 запросов в месяц	$5/ месяц (50 тысяч запросов в месяц)	Около 1630 мс
IpGeolocation And ThreatDetection	Требуется	1500 запросов в день	$10/ месяц (2500 тысячи запросов в день)	Около 70 мс

3.2. Данные об узлах с известной геопозицией. Для того чтобы иметь возможность оценивать точность того или иного IP-геосервиса, необходимо протестировать его на выборке с заранее известными истинными параметрами.

Требования к выборке:

1. Сетевые узлы должны быть рассредоточены по всей территории исследуемого региона;
2. С каждым IP-адресом должна быть сопоставлена достоверная информация о географическом расположении (страна, населенный пункт и координаты);
3. Актуальность информации на момент проведения исследования.

Самым точным способом является создание своей сети либо использование доверенных сетевых устройств, расположение которых можно проверить лично. Другой способ получить проверенные данные – использовать сторонний авторитетный источник. К таковым относятся системы PlanetLab [32] и Atlas. Первый источник – сеть

распределенных по всему миру устройств, использующихся для различных измерений в сети. Данная система поддерживается крупными организациями и учебными заведениями, однако не предоставляет общего доступа для всех желающих, поэтому ее использование в данном исследовании невозможно. Сервис для измерения интернет пространства Atlas – это сеть размещенных по всему миру исследовательских зондов, заявку на установку которых может подать каждый желающий (рисунок 2).

Рис. 2. Расположение зондов системы Atlas на территории России [33]

Архитектура системы Atlas включает:

– слой измерения – распределенная сеть зондов (probes) и анкеров (anchors), отвечающих за выполнение *встроенных* измерений и *пользовательских* измерений (user-defined measurements, UDM);

– слой управления – сервера, отвечающие за работоспособность слоя измерения, хранение и анализ данных, аутентификацию пользователей и т.д.

Работа системы основана на взаимной выгоде ее владельцев и интернет-пользователей: пользователь может подать заявку и, при выполнении определенных условий, безвозмездно получить специализированное сетевое устройство – зонд. Подключив зонд к сети Интернет, пользователь дает возможность системе Atlas проводить встроенные измерения. В свою очередь, Atlas дает возможность пользователю проводить UDM с использованием любых зондов, зарегистрированных в системе.

Atlas позволяет осуществлять следующие виды UDM: проверка сетевой доступности узлов (ping), трассировка маршрутов (traceroute), разрешение доменных имен (DNS), установление шифрованных соединений (SSL/TLS), получение точного времени (NTP), проверка работоспособности web-серверов (HTTP). Создание UDM и получение результатов возможно двумя способами: вручную через web-интерфейс или через web-API. Второй способ является основным при построении автоматизированных систем на основе Atlas.

На рисунке 3 представлен процесс создания и выполнения UDM типа traceroute. Результат выполнения UDM система Atlas передает пользователю в формате JSON.

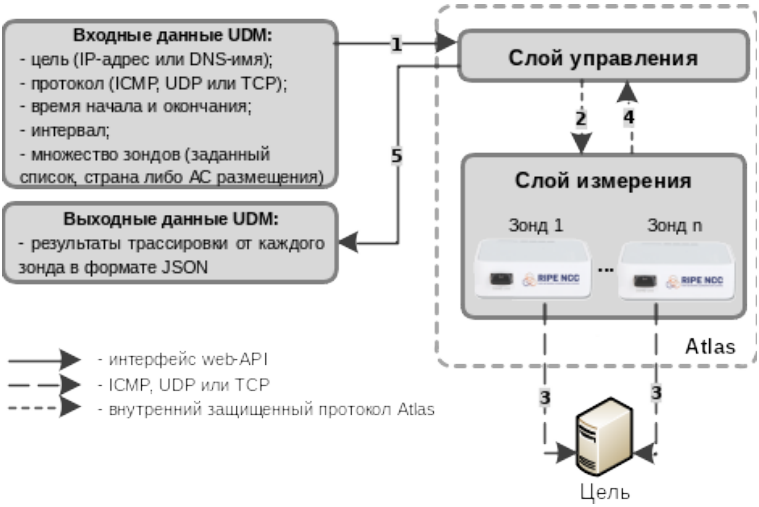

Рис. 3. Процесс создания и выполнения UDM-traceroute

Он является бесплатным и предоставляет API для проведения измерений с помощью зондов. Необходимым условием использования зонда является указание его точной геолокации. Таким образом, система Atlas удовлетворяет предъявляемым требованиям и устройства, входящие в нее, были выбраны в качестве тестовой выборки.

3.3. Данные о населенных пунктах. Для получения сведений о зондах системы Atlas использовались API запросы, параметрами которых явились указывается широта, долгота и радиус зоны поиска. Сервис возвращает список зондов, удовлетворяющих заданным параметрам (рисунок 4).

Для каждого зонда указаны только географические координаты, но не указан город, в котором он находится. Для сопоставления координат зонда и населенного пункта необходимо определить, находятся ли они в его пределах. При данной постановке вопроса целесообразно определить размер города и отправлять запросы с координатой, соответствующей центру города и его радиусом. Известны источники, в которых указана численность населения городов [34], однако источники, содержащие площади малых и средних городов установить не удалось, поэтому формирование значений параметров запросов к системе Atlas осуществлялось на основе математического моделирования.

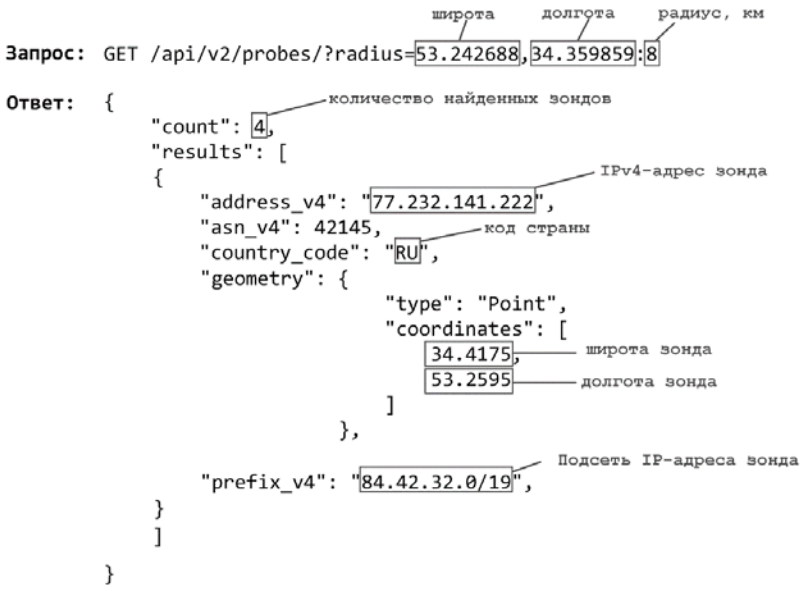

Рис. 4. Пример запроса к системе Atlas и ответа на него

Для создания предсказательной модели была сформирована обучающая выборка, в которую вошли 189 российских городов. Ввиду сравнительно небольшого объема выборки для повышения адекватности модели обучение проводилось с применением кроссвалидации и соотношением training/test равным 80/20. В результате сравнения точности нескольких моделей (линейная регрессия, полиномиальная зависимость, дерево решений, случайный лес) определена лучшая по критерию MSE (minimal square error) –

квадратичная функция (рисунок 5). Предложенная модель не учитывает такие факторы, как уровень технологического развития населенного пункта, его географическое расположение, характер застройки, однако отражает общий характер указанной зависимости.

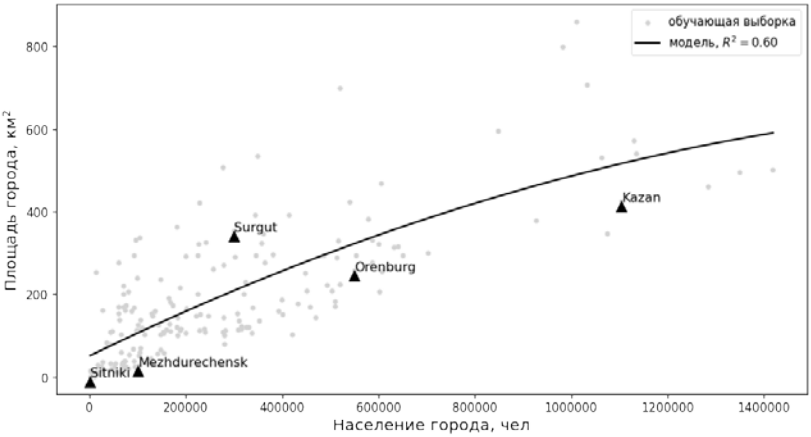

Рис. 5. ML-модель зависимости площади города от численности его населения

4. Проведение исследования

4.1. Сбор данных о геопозиции зондов. Для получения данных о зондах системы Atlas необходимо использовать API запрос, параметрами которого являются географическая координата и радиус, в пределах которого они расположены. Большая часть зондов расположена в городах, поэтому в качестве координат были выбраны центры городов с населением более 500 человек из базы данных Geonames [34]. Радиус города определялся исходя из численности его населения по созданной ML-модели. В результате натурного эксперимента на территории России был обнаружен 691 зонд, установлены их координаты и привязка к населенному пункту.

Априорно известно, что на территории России функционирует 530 зондов [35], поэтому следующим этапом была фильтрация данных, так как один зонд мог входить одновременно в несколько зон различных городов. В этом случае предпочтение отдавалось тому населенному пункту, ближе к центру которого расположен зонд (рисунок 6). Каждому зонду сопоставлялся город, а следовательно, и страна, в которых он расположен. В итоге было найдено 412 зондов, расположенных в 181 городе.

Для получения статистики на основе выборки IP-адресов, соответствующих зондам, были сделаны запросы ко всем IP-геосервисам, указанным в таблице №1.

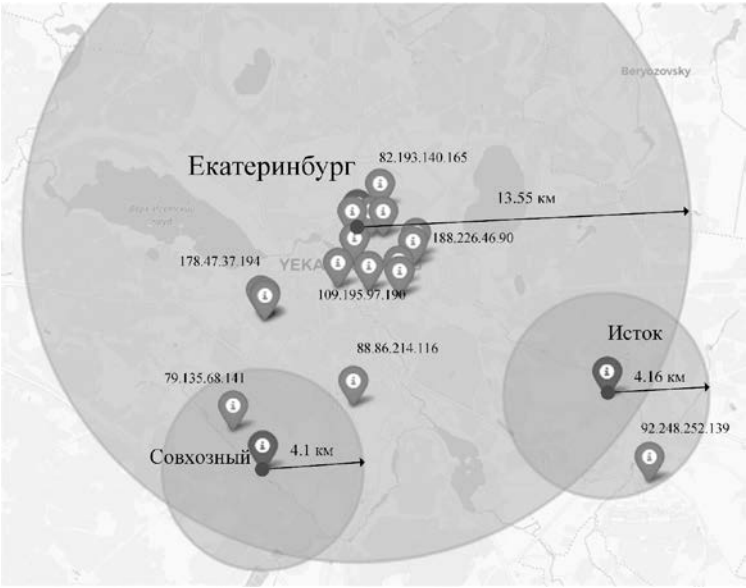

Рис. 6. Расположение зондов на территории населенных пунктов Екатеринбург, Совхозный и Исток. Большинство зондов соотносятся с населенными пунктами. Но, например, зонд с IP-адресом 79.135.68.141 одновременно принадлежит и зоне Екатеринбурга и Совхозного. В данном случае он будет соотнесен с последним, так как находится ближе к его центру

4.2. Анализ полноты и точности IP-геосервисов. По результатам анализа измерений были сделаны выводы о характеристиках IP-геосервисов и тенденциях развития IP-геолокации.

4.2.1. На уровне страны. Анализ IP-геосервисов по точности и полноте на уровне страны позволяет сделать вывод о том, что данная задача решается с высокой точностью (рисунок 7). Практически безошибочно определить государство по IP-адресу могут все без исключения источники. Однако такие IP-геосервисы как Spott, Shodan, IpGeoLocation и IpLocation явно не являются лидерами по полноте охвата сетевого адресного пространства. Что касается других IP-геосервисов, их базы содержат необходимую информацию о государственной принадлежности почти обо всех тестируемых IP-адресах.

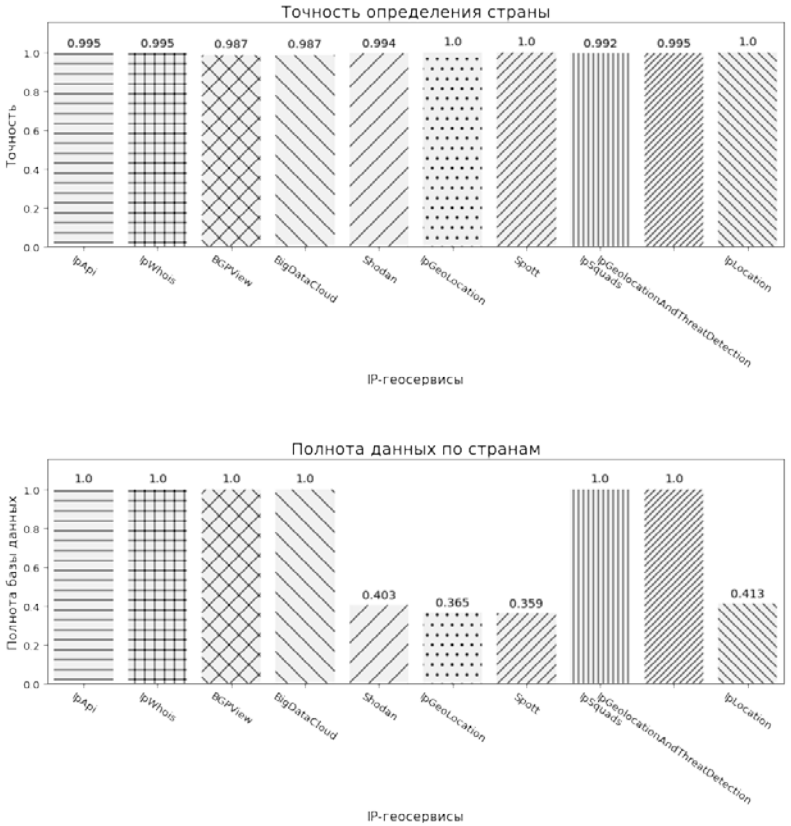

Рис. 7. IP-геолокация на уровне страны

4.2.2. На уровне города. Точность определения города для различных IP-геосервисов варьируется в пределах от 39% до 59% за исключением BGPView, который не предоставляет такой информации вовсе. Полнота баз данных почти не отличается от ситуации с определением страны (рисунок 8). Погрешность в измерения вносит написание названия населенного пункта, поскольку оно может быть написано по-разному в том числе и из-за отсутствия единых правил транслитерации. Собранные от разных IP-геосервисов имена городов имеют отличия и при сравнении дают неверный результат. Так, например, город Санкт-Петербург в ряде источников называется 'st-petersburg', в то время как в других значится как 'saint-petersburg'.

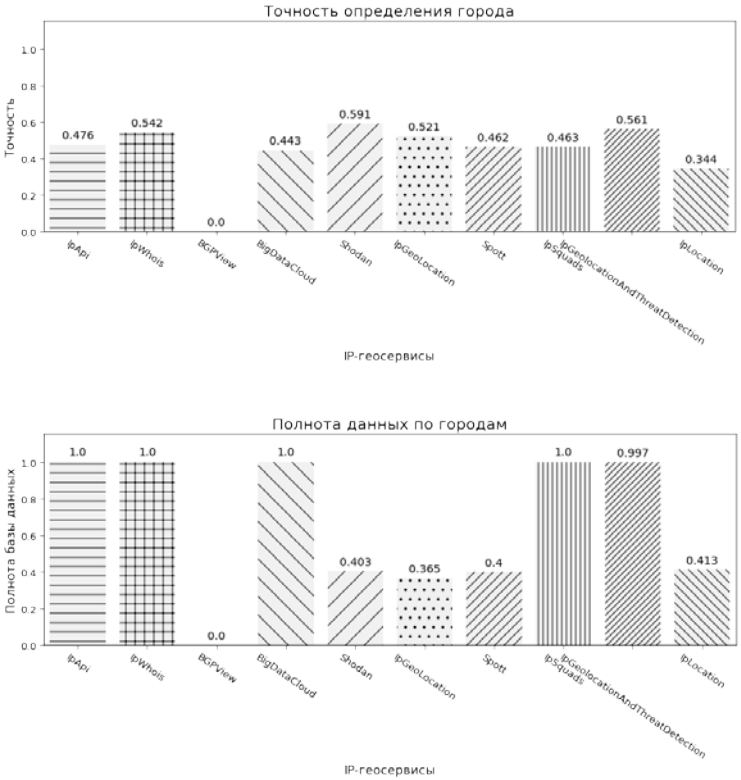

Рис. 8. IP-геолокация на уровне города

По полученным данным можно сделать вывод о том, что полнота базы данных не говорит о ее качестве: Shodan, имея низкие показатели по количеству данных об IP-адресах, показывает довольно высокие результаты по определению города, тогда как у IpSquads ситуация противоположная. Данное качество использовано для назначения уровня «доверия» IP-геосервисам, что позволило увеличить точность IP-геопривязки, используя сильные стороны того или иного источника при обобщении результатов, как показано в пункте 6.

Также частым явлением становится притяжение к более крупному городу. Так, небольшие города вокруг Москвы определялись как Москва. Это может быть связано с отсутствием у IP-геосервисов возможностей для более точного определения местоположения или с отсутствием необходимости иметь настолько точную IP-геопривязку.

Точность определения города по IP-адресу не превышает 60%, однако с учетом различного написания названий городов и притяжения к близко расположенным крупным городам можно обоснованно утверждать, что фактическая IP-геопривязка на уровне городов превышает указанное значение.

4.2.3. На уровне географических координат. Анализ точности определения координат проводился на основе данных о разнице между ответом IP-геосервиса и истинными координатами зонда, выраженными в метрической системе. На основании данных вычислений на всей выборке IP-адресов определялся такой критерий как медианное расстояние (рисунок 9). Он соответствуют значению, при котором равновероятно появление как больших, так и меньших расстояний.

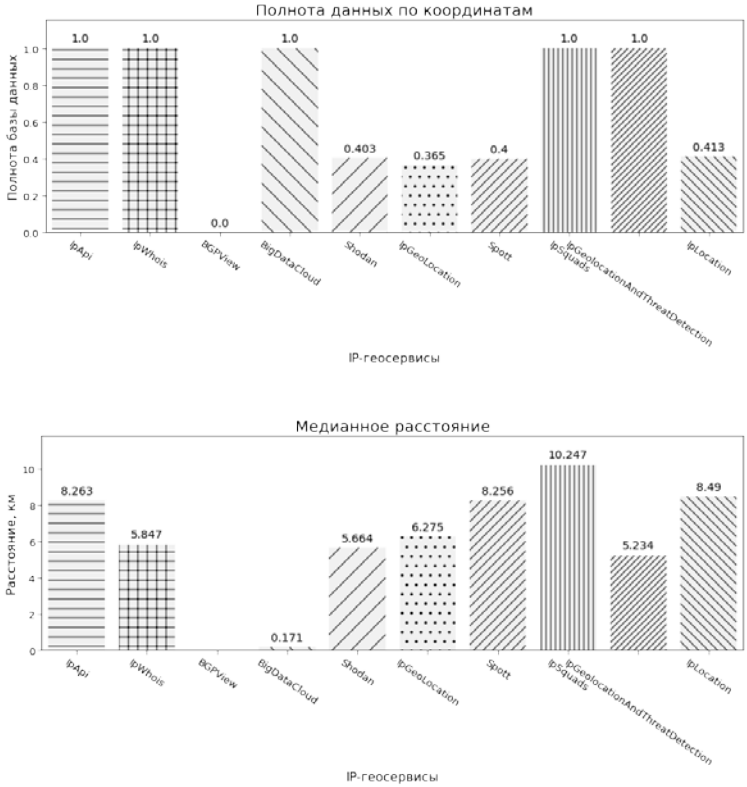

Рис. 9. Статистика определения координаты по IP-адресу

Точность определения координат зонда оценивалась по расстоянию между двумя точками: (а) точкой с истинными координатами зонда и (б) точкой с координатами, полученными от IP-геосервиса. Расстояние измерялось в километрах, было получено 10 выборок (для каждого источника) по 412 значений. Анализ выборок показал, что для них характерно наличие выбросов и распределение значений расстояний не является нормальным. Вследствие этого показателем точности определения координат выбрана медианная ошибка IP-геолокации (MGE – median geolocation error).

Согласно полученным данным наилучший результат показывает BigDataCloud, MGE которого составляет всего около 170 метров. Такая высокая точность может объяснятся наличием у разработчиков BigDataCloud информации о расположении зондов системы Atlas, что отчасти подтверждается статьей на веб-странице сервиса [7]. Точность определения координаты рассмотренных IP-геосервисов различна, но у всех из них MGE сопоставимо с размерами среднего города.

Анализ распределения MGE показывает, что координаты большей части IP-адресов из тестовой выборки определяется с ошибкой не более 200-300 км. Однако при более детальном рассмотрении были обнаружены выбросы при значениях 3-4 тыс. км. Это может быть связано с определением координаты IP-адреса как центра страны, в которой он находится. На практике такая ситуация может возникнуть при применении технологии IP Anycast, когда один и тот же IP-адрес принадлежит сразу нескольким устройствам, расположенным на большом удалении друг от друга. В этом случае каждый IP-геосервис будет выбирать только одно из истинных местоположений, внося большую погрешность в результаты измерений.

Анализ рисунка 10 позволяет сделать вывод о том, что большинство IP-геосервисов имеют распределение «с тяжелым хвостом» со смещением влево. Распределения для Spott, IpSquads и IpLocation являются бимодальными со второй модой в области значения MGE 300-1500 километров. Предположительно это связано с отсутствием у данных IP-геосервисов информации о местоположении IP-адреса, кроме его страны нахождения. Распределение BigDataCloud значительно отличается ото всех остальных. Ошибка в определении координат у данного IP-геосервиса не превышает 500 метров практически в 100% случаев. Данный факт может говорить о наличии у BigDataCloud априорных сведений о размещении зондов системы Atlas.

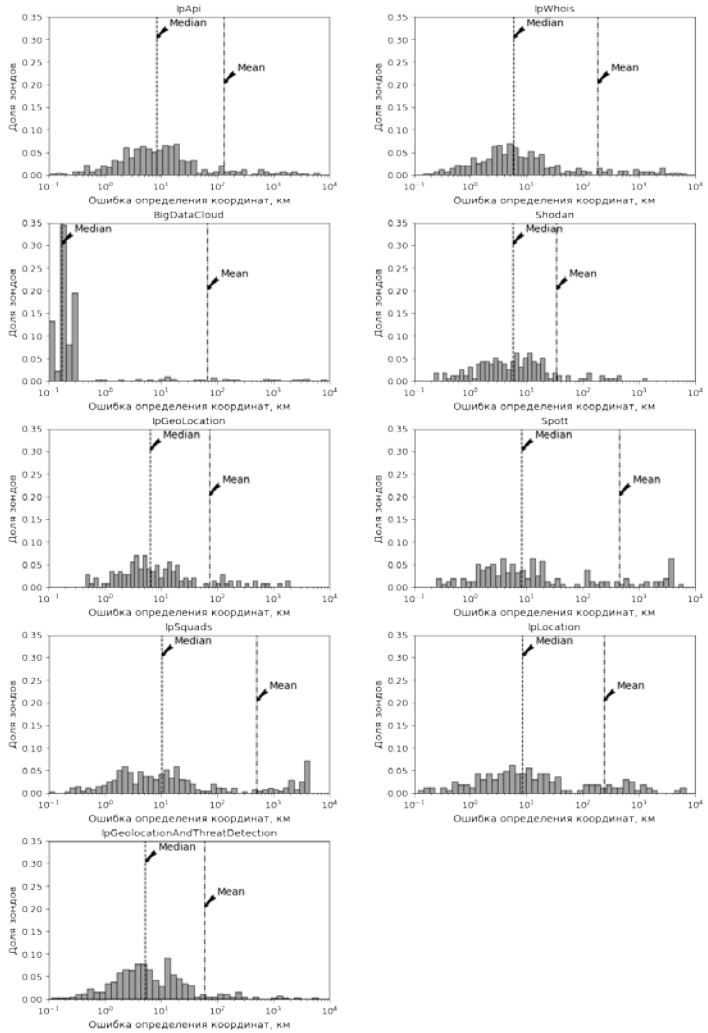

Рис. 10. Гистограммы распределения ошибки определения координат

Для проверки предположения о том, что в больших городах IP-геопривязка точнее, была исследована зависимость расстояния ошибки определения координат от радиуса города. Анализ рисунка 11 позволяет утверждать, что расстояние ошибки обратно пропорциональна размеру города. Так, почти все зонды, расположенные в Москве и Санкт-Петербурге были точно отнесены к

данному месту, в то время как города с радиусом менее 5 км имеют достаточно низкую точность и, вероятно, соответствуют региональному центру.

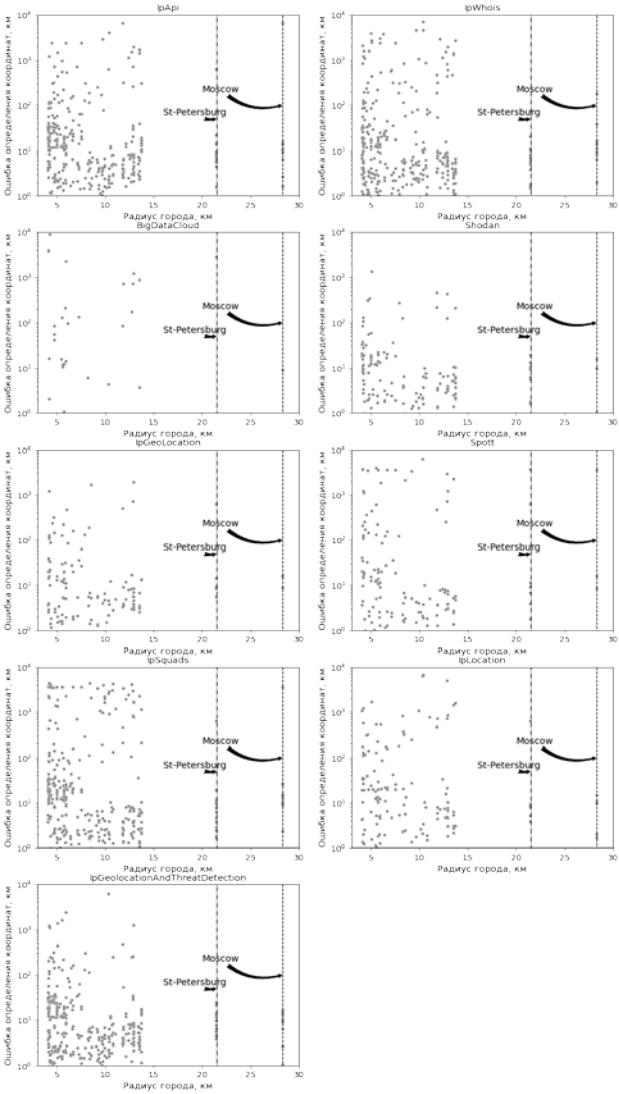

Рис. 11. Зависимость расстояния ошибки определения координат от размеров населенного пункта

Данный факт может быть объяснен двумя взаимосвязанными феноменами развития сетевой инфраструктуры. Во-первых, большой концентрацией в крупных городах тех сетевых узлов, которые могут быть применены для решения задач IP-геолокации (хостинги, дата-центры, университеты, сетевые энтузиасты). Большее количество измерений, проведенное с «городских» узлов, позволяет снизить влияние отдельных «выбросов» и скомпенсировать систематическую погрешность измерений, вносимую различными типами сетевой инфраструктуры за пределами городов. Во-вторых, опережающим развитием информационного сектора в больших городах и меньшим рассредоточением конечных узлов, что также положительно отражается на качестве IP-геолокации.

5. Ограничения. Доступ к ресурсам IP-геосервисов в большинстве случаев ограничен количеством запросов, которые могут быть выполнены в течение месяца. Данный факт накладывает ограничения на объем выборок исследования.

Страна и регион, в котором производятся исследования, оказывают значительное влияние на точность IP-геопривязки. В настоящей работе рассматриваются результаты IP-геолокации только на территории России. Существенные различия в точности определения местоположения наблюдаются как между IP-геосервисами, так и среди данных одного источника в разных регионах. Причины различий в объеме и актуальности баз данных могут заключаться как в возможностях IP-геосервиса по сбору данных на выбранной территории, так и в приоритетности того или иного региона.

Система Atlas позволяет определять геолокацию IP-адресов, присвоенных зондам. Однако информация о местонахождении устройства заполняется его владельцем и в общем случае не может быть достоверной. Погрешность в введении данных, а также умысел владельца могут позволить определить лишь зону, в которой находится зонд. Кроме того, для сохранения конфиденциальности геолокации пользователей сам сервис незначительно изменяет координаты зонда, предоставляемые по запросу, что также негативно сказывается на точности исследования. По доступным данным инструментальная погрешность измерений выдаваемой координаты от истинного местоположения составляет около 100-500 метров, при исключении факта умышленного искажения данных владельцем зонда.

Для формирования значений параметров запросов к системе Atlas была построена ML-модель зависимости между численностью населения города и его площадью. В качестве обучающей выборки

использовались только российские города с населением более 1000 человек, так как количество сетевых устройств в меньших населенных пунктах незначительно, а их площадь отличается несущественно. Радиус города определялся исходя из предположения о том, что он имеет радиально-кольцевую структуру. ML-модель достаточно точно определяет размеры маленьких и средних городов, однако при возрастании численности населения качество результатов деградирует значительно. Это связано с географическим положением и уровнем развития города, преобладающим типом застройки. Так, созданная модель прогнозирует с наибольшей ошибкой площади Санкт-Петербурга и Москвы, в связи с чем в рамках исследования площади данных городов рассчитаны вручную. Таким образом, при необходимости определения площади городов с населением более 1.5 миллионов человек следует использовать детерминированные источники.

6. Предложения по повышению точности IP-геолокации. Для повышения точности IP-геолокации авторами предложен ансамблевый метод усреднения координат, полученных от разных IP-геосервисов. При этом каждый ресурс характеризуется коэффициентом, отражающим его точность на данной территории, что позволяет учитывать достоинства и недостатки всех IP-геосервисов.

Уточнение IP-геопривязки на уровне страны и города

Из всех ответов IP-геосервисов по мажоритарному принципу выбирается наиболее вероятная страна местонахождения устройства. Источники, ответы которых не совпали с решением большинства, в последующих действиях не учитываются. Аналогично уточняется населенный пункт. В результате остаются только те IP-геосервисы, у которых совпали страна и город в местоположении целевого IP-адреса.

Уточнение IP-геопривязки на уровне географических координат

Пусть r_i – медианная ошибка IP-геолокации i -го IP-геосервиса. Тогда уровень доверия c_i к IP-геосервису может быть выражен как:

$$c_i = 1 - \frac{r_i - \min R}{\max R}, \quad (1)$$

где R – множество значений медианных ошибок определения местоположения для всех IP-геосервисов.

Вычисление уточненных координат сводится к усреднению координат, полученных от разных IP-геосервисов, и использованию корректирующих слагаемых, рассчитанных на основании уровней

доверия (рисунок 12). Поскольку $0 < c_i \leq 1$, каждый IP-геосервис оказывает влияние на значение уточненных координат.

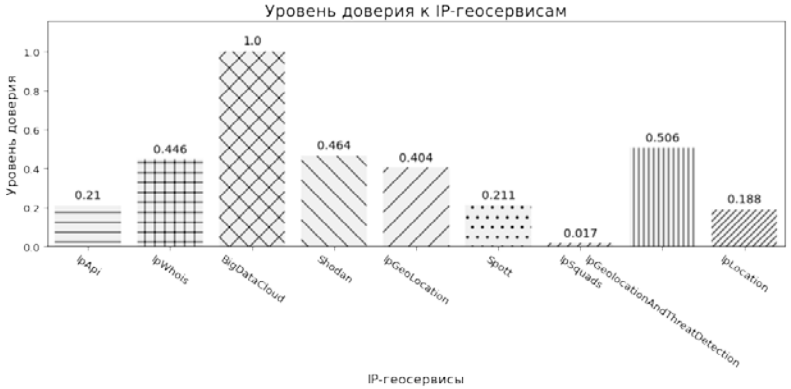

Рис. 12. Уровень доверия к IP-геосервисам

Усредненная широта рассчитывается по формуле (2), а уточненная широта – по формуле (3):

$$lat_{cp} = \frac{1}{R}\sum_{i=1}^{|R|} lat_i\,, \tag{2}$$

$$lat_{ym} = lat_{cp} + \sum_{i=1}^{|R|} c_i\,(lat_i - lat_{cp})\,. \tag{3}$$

Аналогичные формулы используются для расчета долготы. Применение формул (1)-(3) проиллюстрировано на рисунке 13. При вычислении усредненных значений широты и долготы не используются преимущества каждого из IP-геосервисов. В то же время уточненные значения координат лишены этого недостатка за счет использования уровня доверия к IP-геосервисам в качестве коэффициентов, линейно приближающих результирующее значение к показаниям наиболее точных из них.

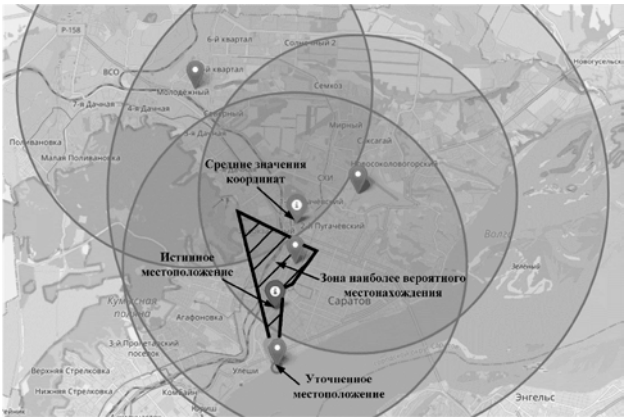

Рис. 13. Иллюстрация предлагаемого метода уточнения IP-геопривязки

Таким образом, на основании данных от нескольких IP-геосервисов уточняется местоположение сетевого устройства на уровнях страны, города и географических координат.

Предложенный метод усреднения координат был проверен экспериментально на зондах системы Atlas на территории России, MGE составила 5.2 км. Данный показатель незначительно превосходит некоторые из IP-геосервисов, при этом позволяет получить более стабильный результат в большем количестве случаев, о чем свидетельствует малое количество выбросов в области более 70 км (рисунок 14). Полученные результаты подтверждают целесообразность применения ансамблевых методов уточнения координат, полученных от разных IP-геосервисов.

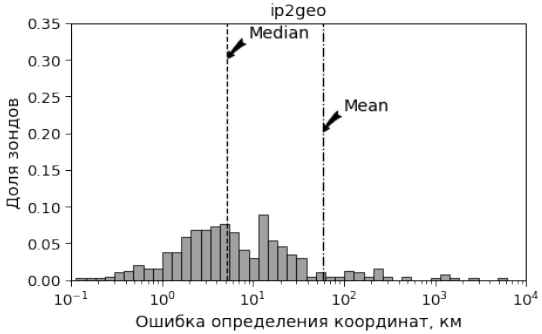

Рис. 14. Гистограммы распределения ошибки определения координат для предлагаемого алгоритма

7. Заключение. Информация о местоположении устройства с известным IP-адресом используется при обеспечении информационной безопасности, для оптимизации сетевого трафика, в интернет-маркетинге и других сферах человеческой деятельности, что определяет актуальность задачи IP-геолокации.

Учитывая динамику изменений глобальной сети на уровне интернет-провайдеров, развитие «облачных» сервисов, расширение интернет-аудитории задача IP-геолокации требует как совершенствования измерительной сети, так и развития методик анализа результатов измерений.

В результате исследования получены данные о точности и полноте баз данных наиболее распространенных IP-геосервисов на уровне определения страны, города и координат сетевого устройства. Сформулированы предположения о причинах различных аномалий в результатах IP-геопривязки.

Установлено, что точность IP-геолокации зависит от размера населенного пункта, в котором размещен сетевой узел — при увеличении размера города расстояние ошибки определения координат уменьшается.

Авторами предложен метод повышения точности IP-геолокации, заключающийся в определении страны и города методом простого большинства и вычислении координат с учетом уровня доверия к каждому из рассмотренных IP-геосервисов. Уровень доверия устанавливается в результате сравнительного анализ IP-геосервисов по точности.

Необходимо отметить, что для получения качественного результата IP-геолокации требуется периодическое оценивание основных показателей работы IP-геосервисов, а определение уровня «доверия» должно стать неотъемлемой частью процесса эксплуатации IP-геосервисов.

Литература

1. Wang, Zhihao, et al. “Towards IP Geolocation with Intermediate Routers Based on Topology Discovery.” Cybersecurity, vol. 2, no. 1, Apr. 2019.
2. Williams J. Identification of IP address using fraudulent geolocation data, Imperial College London, 15 June 2020
3. Wang, Z., Li, H., Li, Q.: Towards IP geolocation with intermediate routers based on topology discovery. Cybersecurity 2(1), 1–13 (2019) 5
4. Zhao, Fan & Luo, Xiangyang & Gan, Yong & Zu, Shuodi & Cheng, Qingfeng & Liu, Fenlin. (2018). IP Geolocation based on identification routers and local delay distribution similarity. Concurrency and Computation: Practice and Experience. 31. 10.1002/cpe.4722.
5. Top 10 Best IP Geolocation APIs (in 2022) [Электронный ресурс] - Режим доступа: URL: https://rapidapi.com/blog/ip-geolocation-api/ (21.02.2022)

6. Adebayo, Semiu. Migration of IPv4 to IPv6; Translation Method, 2018.
7. P. Nisenblat, IP Geolocation Demystified [Электронный ресурс] - Режим доступа: URL: https://www.bigdatacloud.com/blog/ip-geolocation-demystified (10.12.2021)
8. Measures of distance between samples: Euclidean [Электронный ресурс] - Режим доступа: URL: http://www.econ.upf.edu/~michael/stanford/maeb4.pdf (26.11.2021)
9. Zhihao Li, Dave Levin, Neil Spring, and Bobby Bhattacharjee. 2018. Internet anycast: performance, problems, & potential. In Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '18). Association for Computing Machinery, New York, NY, USA, 59–73.
10. Aljumaily, Mustafa. (2016). Content Delivery Networks Architecture, Features, and Benefits. 10.13140/RG.2.1.1762.0722
11. Mohammed Jubaer Arif, Shanika Karunasekera, Santosh Kulkarni, Ajit Gunatilaka, and Branko Ristic. 2010. Internet Host Geolocation Using Maximum Likelihood Estimation Technique. In 24th IEEE International Conference on Advanced Information Networking and Applications. IEEE, Perth, Australia, 422-429
12. J. Hawley, "GeoDNS -Geographically-aware, protocol-agnostic load balancing at the DNS level," in Proceedings of the Linux Symposium, pp. 123-130, Linux Symposium Inc., heinäkuu 2009.
13. Сухаревская Е.В. ИССЛЕДОВАНИЕ СИСТЕМ АУТЕНТИФИКАЦИИ // Международный студенческий научный вестник. – 2018. – № 1
14. Аутентификация, основанная на местоположении выхода в интернет [Электронный ресурс] - Режим доступа: URL: https://studfile.net/preview/16435809/page:4/#8 (02.03.2022)
15. Skrill: инструкция по верификации аккаунта 2022 [Электронный ресурс] - Режим доступа: URL: https://baxity.com/ru/skrill-instruktsiya-po-verifikatsii-akkaunta-2022 (25.03.2022)
16. Taylor, J., Devlin, J., Curran, K. (2012) Bringing location to IP Addresses with IP Geolocation. The Journal of Emerging Technologies in Web Intelligence, Vol. 4, No. 3, August 2012
17. V.N. Padmanabhan and L. Subramanian. An investigation of geographic mapping techniques for internet hosts. In ACM SIGCOMM, pages 173–185, 2001
18. Ovidiu Dan, Vaibhav Parikh, and Brian D. Davison. 2018. IP Geolocation through Reverse DNS. CoRR abs/1811.04288(2018), 1-10.
19. Luckie, Matthew & Dhamdhere, Amogh & Huffaker, Bradley & Clark, David & claffy, kc. (2016). bdrmap: Inference of Borders Between IP Networks. 381-396. 10.1145/2987443.2987467.
20. RIPE Atlas [Электронный ресурс] - Режим доступа: URL: https://atlas.ripe.net/ (15.12.2021)
21. Dan, Ovidiu & Parikh, Vaibhav & Davison, Brian. (2018). IP Geolocation through Reverse DNS.
22. Scheitle, Quirin & Gasser, Oliver & Sattler, Patrick & Carle, Georg. (2017). HLOC: Hints-Based Geolocation Leveraging Multiple Measurement Frameworks.
23. Spring, Neil & Mahajan, Ratul & Wetherall, David. (2002). Measuring ISP Topologies with Rocketfuel. ACM SIGCOMM Computer Communication Review. 32. 133-145. 10.1145/633025.633039.
24. GeoIP Databases & Services: Industry Leading IP Intelligence [Электронный ресурс] - Режим доступа: URL: https://www.maxmind.com/en/geoip2-services-and-databases (25.01.2022)
25. Gharaibeh, Manaf & Shah, Anant & Huffaker, Bradley & Zhang, Han & Ensafi, Roya & Papadopoulos, Christos. (2017). A look at router geolocation in public and commercial databases. 463-469. 10.1145/3131365.3131380.

26. B. Hufaker, M. Fomenkov, and kc claffy. Geocompare: a comparison of public and commercial geolocation databases. In Proceedings of the Network Mapping and Measurement Conference (NMC), 2011.
27. M. Gouel, K. Vermeulen, O. Fourmaux, T. Friedman, R. Beverly. IP Geolocation Database Stability and Implications for Network Research. Network Traffic Measurement and Analysis Conference, Sep 2021, Online, United States.
28. Du, Ben & Candela, Massimo & Huffaker, Bradley & Snoeren, Alex & claffy, kc. (2020). RIPE IPmap active geolocation: mechanism and performance evaluation. ACM SIGCOMM Computer Communication Review. 50. 3-10. 10.1145/3402413.3402415.
29. IP Geolocation API [Электронный ресурс] - Режим доступа: URL: https://ip-api.com/ (12.01.2022)
30. Документация по API IP Geolocation API's [Электронный ресурс] - Режим доступа: URL: https://rapidapi.com/ru/IPSquads/api/ip-geolocation-api-s (12.01.2022)
31. Документация по API IP Geolocation and Threat Detection [Электронный ресурс] - Режим доступа: URL: https://rapidapi.com/ru/ipregistry3-ipregistry/api/ip-geolocation-and-threat-detection/ (12.01.2022)
32. PlanetLab Europe [Электронный ресурс] - Режим доступа: URL: https://www.planet-lab.eu/Home (19.02.2022)
33. Global RIPE Atlas Network Coverage [Электронный ресурс] - Режим доступа: URL: https://atlas.ripe.net/results/maps/network-coverage (22.01.2022)
34. GeoNames [Электронный ресурс] - Режим доступа: URL: https://www.geonames.org (26.01.2022)
35. RIPE Atlas Probes [Электронный ресурс] - Режим доступа: URL: https://atlas.ripe.net/probes/ (14.02.2022)

Иванов Максим Владимирович — канд. техн. наук, сотрудник, Академия Федеральной службы охраны Российской Федерации (Академия ФСО России). Область научных интересов: представление и обработка данных в виде графов, методы описания иерархических сетей, применение методов дискретной оптимизации, технологии разработки распределенных программных комплексов. Число научных публикаций — 22. maximivanov@mail.ru; улица Приборостроительная, 35, 302015, Орел, Россия; р.т.: +7(4862)549-615.

Полунин Александр Александрович — сотрудник, Академия Федеральной службы охраны Российской Федерации (Академия ФСО России). Область научных интересов: архитектуры компьютерных сетей, машинное обучение, разработка приложений, информационная безопасность. Число научных публикаций — 1. polunin2002@mail.ru; улица Приборостроительная, 35, 302015, Орел, Россия; р.т.: +7(4862)549-615.

DOI 10.15622/ia.21.4.5

M. IVANOV, A. POLUNIN

IMPROVING THE ACCURACY OF IP GEOLOCATION BASED ON PUBLIC IP GEOSERVICES DATA

Ivanov M., Polunin A. **Improving the Accuracy of IP Geolocation Based on Public IP Geoservices Data.**

Abstract. IP geolocation is the process of determining the real geographic location of an electronic device connected to the Internet, by its global network address [1]. Currently, it has found wide application in Internet commerce, marketing and advertising, information security [2], and other areas of human activity. There are different methods for determining the location of a remote network device, which differ both in type of analyzed information (delay packet transmission, resource records DNS-servers, the content of Web pages), and the result (country or city name, mail address, probable area of location or exact coordinates) [3, 4]. IP geolocating error depends on the country, population density, type of network device and ranges from several tens of meters to hundreds of kilometers. For the same input data, the results of different IP-geoservices can vary significantly. The object of this study is the public IP-geoservices that provide geolocating services for nodes in the global network based on their IP addresses, and specifically, their accuracy and completeness. The sample of IP-geoservices for testing was formed from the most popular ones [5]. During the study, the results of IP-geolocation were compared with reliable information about the location of some IP addresses, as indicators of accuracy country, city and geographic coordinates were used. Based on the comparative analysis of the test results, conclusions about the accuracy of IP-geolocation services according to the selected indicators, their essential properties, as well as the dependence of geolocation error on the size of the settlement were made. To improve the accuracy of IP georeferencing, the authors proposed an ensemble method for averaging coordinates obtained from several IP geoservices.

Keywords: Internet, IP-geolocation, IP-geoservice, Atlas, IpAPI, Shodan.

References

1. Wang, Zhihao, et al. “Towards IP Geolocation with Intermediate Routers Based on Topology Discovery.” Cybersecurity, vol. 2, no. 1, Apr. 2019.
2. Williams J. Identification of IP address using fraudulent geolocation data, Imperial College London, 15 June 2020
3. Wang, Z., Li, H., Li, Q.: Towards IP geolocation with intermediate routers based on topology discovery. Cybersecurity 2(1), 1–13 (2019) 5
4. Zhao, Fan & Luo, Xiangyang & Gan, Yong & Zu, Shuodi & Cheng, Qingfeng & Liu, Fenlin. (2018). IP Geolocation based on identification routers and local delay distribution similarity. Concurrency and Computation: Practice and Experience. 31. 10.1002/cpe.4722.
5. Top 10 Best IP Geolocation APIs (in 2022) - Available at: https://rapidapi.com/blog/ip-geolocation-api/ (21.02.2022)
6. Adebayo, Semiu. Migration of IPv4 to IPv6; Translation Method, 2018.
7. P. Nisenblat, IP Geolocation Demystified - Available at: https://www.bigdatacloud.com/blog/ip-geolocation-demystified (10.12.2021)
8. Measures of distance between samples: Euclidean - Available at: http://www.econ.upf.edu/~michael/stanford/maeb4.pdf (26.11.2021)
9. Zhihao Li, Dave Levin, Neil Spring, and Bobby Bhattacharjee. 2018. Internet anycast: performance, problems, & potential. In Proceedings of the 2018 Conference of the

ACM Special Interest Group on Data Communication (SIGCOMM '18). Association for Computing Machinery, New York, NY, USA, 59–73.
10. Aljumaily, Mustafa. (2016). Content Delivery Networks Architecture, Features, and Benefits. 10.13140/RG.2.1.1762.0722
11. Mohammed Jubaer Arif, Shanika Karunasekera, Santosh Kulkarni, Ajit Gunatilaka, and Branko Ristic. 2010. Internet Host Geolocation Using Maximum Likelihood Estimation Technique. In 24th IEEE International Conference on Advanced Information Networking and Applications. IEEE, Perth, Australia, 422-429
12. J. Hawley, "GeoDNS -Geographically-aware, protocol-agnostic load balancing at the DNS level," in Proceedings of the Linux Symposium, pp. 123-130, Linux Symposium Inc., heinäkuu 2009.
13. Suharevskaja E.V. ISSLEDOVANIE SISTEM AUTENTIFIKACII [Authentication systems research]// Mezhdunarodnyj studencheskij nauchnyj vestnik [International student scientific journal] – 2018. – № 1 (In Russ.)
14. Autentifikacija, osnovannaja na mestopolozhenii vyhoda v internet [Authentication based on the location of the Internet connection] - Available at: https://studfile.net/preview/16435809/page:4/#8 (02.03.2022) (In Russ.)
15. Skrill: instrukcija po verifikacii akkaunta 2022 [Skrill: account verification instructions 2022] - Available at: https://baxity.com/ru/skrill-instruktsiya-po-verifikatsii-akkaunta-2022 (25.03.2022) (In Russ.)
16. Taylor, J., Devlin, J., Curran, K. (2012) Bringing location to IP Addresses with IP Geolocation. The Journal of Emerging Technologies in Web Intelligence, Vol. 4, No. 3, August 2012
17. V.N. Padmanabhan and L. Subramanian. An investigation of geographic mapping techniques for internet hosts. In ACM SIGCOMM, pages 173–185, 2001
18. Ovidiu Dan, Vaibhav Parikh, and Brian D. Davison. 2018. IP Geolocation through Reverse DNS. CoRR abs/1811.04288(2018), 1-10.
19. Luckie, Matthew & Dhamdhere, Amogh & Huffaker, Bradley & Clark, David & claffy, kc. (2016). bdrmap: Inference of Borders Between IP Networks. 381-396. 10.1145/2987443.2987467.
20. RIPE Atlas, Available at: https://atlas.ripe.net/ (15.12.2021)
21. Dan, Ovidiu & Parikh, Vaibhav & Davison, Brian. (2018). IP Geolocation through Reverse DNS.
22. Scheitle, Quirin & Gasser, Oliver & Sattler, Patrick & Carle, Georg. (2017). HLOC: Hints-Based Geolocation Leveraging Multiple Measurement Frameworks.
23. Spring, Neil & Mahajan, Ratul & Wetherall, David. (2002). Measuring ISP Topologies with Rocketfuel. ACM SIGCOMM Computer Communication Review. 32. 133-145. 10.1145/633025.633039.
24. GeoIP Databases & Services: Industry Leading IP Intelligence - Available at: https://www.maxmind.com/en/geoip2-services-and-databases (25.01.2022)
25. Gharaibeh, Manaf & Shah, Anant & Huffaker, Bradley & Zhang, Han & Ensafi, Roya & Papadopoulos, Christos. (2017). A look at router geolocation in public and commercial databases. 463-469. 10.1145/3131365.3131380.
26. B. Hufaker, M. Fomenkov, and kc claffy. Geocompare: a comparison of public and commercial geolocation databases. In Proceedings of the Network Mapping and Measurement Conference (NMC), 2011.
27. M. Gouel, K. Vermeulen, O. Fourmaux, T. Friedman, R. Beverly. IP Geolocation Database Stability and Implications for Network Research. Network Traffic Measurement and Analysis Conference, Sep 2021, Online, United States.
28. Du, Ben & Candela, Massimo & Huffaker, Bradley & Snoeren, Alex & claffy, kc. (2020). RIPE IPmap active geolocation: mechanism and performance evaluation.

ACM SIGCOMM Computer Communication Review. 50. 3-10. 10.1145/3402413.3402415.
29. IP Geolocation API - Available at: https://ip-api.com/ (12.01.2022)
30. Dokumentacija po API IP Geolocation API's [IP Geolocation API Documentation] - Available at: https://rapidapi.com/ru/IPSquads/api/ip-geolocation-api-s (12.01.2022) (In Russ.)
31. Dokumentacija po API IP Geolocation and Threat Detection [IP Geolocation and Threat Detection API Documentation] - Available at: https://rapidapi.com/ru/ipregistry3-ipregistry/api/ip-geolocation-and-threat-detection/ (12.01.2022) (In Russ.)
32. PlanetLab Europe - Available at: https://www.planet-lab.eu/Home (19.02.2022)
33. Global RIPE Atlas Network Coverage - Available at: https://atlas.ripe.net/results/maps/network-coverage (22.01.2022)
34. GeoNames - Available at: https://www.geonames.org (26.01.2022)
35. RIPE Atlas Probes - Available at: https://atlas.ripe.net/probes/ (14.02.2022)

Ivanov Maxim — Ph.D., Researcher, Academy of Federal Security Guard Service of the Russian Federation. Research interests: graph models, hierarchical networks, discrete optimization, distributed software. The number of publications — 22. maximivanov@mail.ru; 35, Priborostroitelnaya St., 302015, Orel, Russia; office phone: +7(4862)549-615.

Polunin Alexander — Researcher, Academy of Federal Security Guard Service of the Russian Federation. Research interests: models of the computer networks, machine learning, application development, information security. The number of publications — 1. polunin2002@mail.ru; 35, Priborostroitelnaya St., 302015, Orel, Russia; office phone: +7(4862)549-615.