International Journal of
Geo-Information

Article

Evaluation of Device-Independent Internet Spatial Location

Dan Komosny [1],*, Paul Pang [2], Miralem Mehic [3] and Miroslav Voznak [3]

[1] Department of Telecommunications, Brno University of Technology, Brno 61600, Czech Republic
[2] Department of Computing, Unitec Institute of Technology, Auckland 1025, New Zealand; ppang@unitec.ac.nz
[3] Department of Telecommunications, Technical University of Ostrava, Ostrava 70833, Czech Republic; miralem.mehic.st@vsb.cz (M.M.); miroslav.voznak@vsb.cz (M.V.)
* Correspondence: komosny@feec.vutbr.cz; Tel.: +420-54114-6973

Academic Editors: Norbert Bartelme and Wolfgang Kainz
Received: 27 February 2017; Accepted: 24 May 2017; Published: 27 May 2017

Abstract: Device-independent Internet spatial location is needed for many purposes, such as data personalisation and social behaviour analysis. Internet spatial databases provide such locations based the IP address of a device. The free to use databases are natively included into many UNIX and Linux operating systems. These systems are predominantly used for e-shops, social networks, and cloud data storage. Using a constructed ground truth dataset, we comprehensively evaluate these databases for null responses, returned country/region/city, and distance error. The created ground truth dataset differs from others by covering cities with both low and high populations and maintaining only devices that follow the rule of one IP address per ISP (Internet Service Provider) and per city. We define two new performance metrics that show the effect of city population and trustworthiness of the results. We also evaluate the databases against an alternative measurement-based approach. We study the reasons behind the results. The data evaluated comes from Europe. The results may be of use for engineers, developers and researchers that use the knowledge of geographical location for related data processing and analysis, such as marketing.

Keywords: spatial location; device independent; IP address; geolocation; database; location-aware; city population; internet

1. Introduction

This paper deals with the spatial location of Internet devices. It focuses on *device-independent* location where the location of a device is estimated remotely by its known IP address and not locally by the device itself, such as using a GPS module. Therefore, the paper focuses on an Internet device-independent spatial locations that

- does not use GPS or other global positioning systems,
- does not use terrestrial radio-based location systems, such as triangulation in WiFi or mobile cellular networks,
- does not use location information entered by users.

The device-independent service providers of do not communicate with the device located. The location is estimated based on the knowledge of device IP address. Such IP-based location is used for a great number of location-aware services and applications, including web content and social network personalization [1]; user behaviour analysis [2] (including visitor maps for websites); load balancing by redirecting users to geographically close data/resource replicas [3]; geographical-based data collection from a large number of users [4]; spam filtering by sender location;

detection of ID and password sharing; detection of credit card online fraud [5]; and law enforcement on media distribution by delivery restrictions.

Compared with Internet *device-dependent* location, the service providers directly communicate with the device located to obtain its location. The shared location is typically obtained from GPS or WiFi triangulation. A user typically has to agree with sharing their location. This may happen when an application/service is being installed and a user accepts the legal agreements. Applications of Internet device-dependent location vary a lot. Typical examples are finding nearby points of interests, finding nearest social media contacts, traffic reports, municipal transport schedules, etc.

In this paper, we evaluate the performance of Internet device-independent spatial location. We focus on free to use location databases as they are commonly used for running Internet services and applications. The free location databases are natively included in many open-source UNIX and Linux operating systems. These systems are predominantly used for hosting e-shops, social networks, and cloud data storage. The web analytic and market share company W3Cook monitors the top million servers monthly and shows that around 97% of the web servers are hosted by open-source Linux operating systems [6]. The free location databases are also available through software packaging systems (rpm, deb) for such operating systems.

We use a filtered ground truth dataset that we constructed. It includes devices coming from Europe with a known correct location. These devices are mobile phones with WiFi. Our ground truth dataset differs from those used in the related research by covering different types of country areas, especially including less inhabited places. The datasets used in the related work typically come from large and major cities. An example of such a commonly used dataset is PlanetLab [7]. The devices from large cities are generally located with a better accuracy. However, for general use, the devices from lesser populated places should also be considered. We provide a detailed description of the dataset created, including the city population numbers.

An uneven distribution of the ground truth devices among the ISPs influences location accuracy. The reason is that devices with an IP address assigned via the same ISP provider are mapped to the same or a small set of locations. We therefore process our ground truth dataset to maintain only devices that follow the rule of one IP address per ISP and per city. This way, the results should be error-free as they are not influenced by a multiplication of the same results for the devices with an IP address obtained through the same ISP.

We particularly evaluate these metrics: null responses versus returned locations, correctly resolved country/region/city, and distance error. We additionally define two new metrics that show the particular performance differences. These metrics are 'accuracy dependence on city population' and 'trustworthiness of the locations'. We also compare the free databases with an alternative measurement-based location.

The paper is structured as follows: Section 2 describes the background on device-independent spatial location based on IP addresses. Section 3 gives a list of the related research and surveys in IP geolocation. We summarize the presented results and discuss their evaluation. Section 4 presents in detail the principle of IP geolocation. A description of the free to use databases used in this research is also included. In Section 5, we describe our method for creating the ground truth dataset. The used methodology for location performance evaluation is presented in Section 6. The results and comments on the findings are described in Section 7. Finally, we conclude the paper.

2. Background on Device-Independent Spatial Location

There are several methods used for device-independent location based on IP addresses. We summarize them in Table 1. The first two are used by the general public. These methods can be accessed freely or through a paid subscription. The last method is typically dedicated for legally authorized entities only.

The first method—database lookup—deals with searching the global location databases. A geolocation database maintains blocks of continuous IP addresses. Each such block has a location assigned. The assigned locations are obtained through several sources that we describe in Section 4. A device location is estimated by searching the database for the corresponding block of IPs and if a match is found, the geographical position stored for the block is returned [8–10].

The second method—Internet measurement—is based on measurement and analysis of data communication. Typically, communication latency is measured from a set of servers called landmarks with known locations to the device located. The latencies are converted to a geographical location of the device using several techniques, such as distance multilateration [11,12].

The last method—ISP private data—is based on private records of Internet service providers (ISP). An ISP leases IP addresses to the devices. Country-specific laws specify what information has to be recorded and how long it has to be maintained by ISPs. ISP subscribers provide their details including postal addresses in the billing contracts. The location of a device can be tracked down by linking this information. The legal details may vary across countries. The police and justice services are usually authorised to be given the location upon a formal request.

The accuracy of device-independent location is generally low, usually in a range of tens or hundreds of kilometres. The accuracy of the ISP internal location information varies according to the records kept.

Table 1. Overview of device-independent location based on IP addresses.

How	Accuracy	Typically Used by
Database lookup	Low	Non-device user (everybody)
Internet measurement	Low	Non-device user (everybody)
ISP private data	High	Non-device user (authorized)

There are other minor methods of device-independent location, such as data-mining of web pages and other Internet resources for spatial information [13]. Another method is based on an enhancement of the DNS (Domain Name System) service called DNS LOC [14]. This service provides the geographical location for a domain name. The disadvantage of this solution is a poor coverage of Internet address space, and it is not widely used.

Different methods are used for *device-dependent* location. These methods are usually based on global positioning systems (GPS), measuring radio signal strength (RSSI), time of arrival (TOA), and angle of arrival (AOA) [15,16]. The accuracy of device-dependent location is higher than device-independent. Some principles are shared with device-independent location, such as distance multilateration.

3. Related Work

In this section, we review the related work and summarize the results with a focus on these points of interest: location accuracy, location efficiency (null responses versus returned locations), and the up-to-dateness problem. The problem dealing with the currentness of the stored location data is specific to any database-based IP geolocation and is caused by the non-static IP address space since new assignments and reallocations of IP addresses continually happen.

Ref. [17] studied the accuracy of the common location databases—MaxMind, IPligence, Geobytes, HostIP, Digital Envoy/NetAcuity, and IP2Location. The study indicated that the location performance varied a lot—from 20 to 80% of the correct city estimations. The authors of this study used a 100 km range for the city-level definition, whereas other researchers used a 40 km range, such as [8,9]. The results within the country-level were around 80% and better. The study also involved a statistic of the successful location queries. Some databases did not locate about 30% of the IP addresses requested (we left the HostIP database as it reported a much higher percentage of the unsuccessful location attempts than the others). The authors of the study emphasised that this test showed whether an entry for an IP address exists and it does not say anything about the accuracy of the locations obtained. The study also compared the location accuracy reported by the database vendors. The reported country-level accuracy was over 97% and the city-level accuracy was over 80%. These values were of a greater accuracy compared to the data collected in the study.

Ref. [8] evaluated the accuracy of several location databases including Digital Envoy/NetAcuity, MaxMind/GeoIP, MaxMind/GeoLite, IPligence and HostIP. They found that over 95% of the locations were correct at the country-level. Of the correct locations, 70–90% were within the city-level (40 km).

Ref. [9] considered five location databases—MaxMind, IP2Location, InfoDB, HostIP, and Software77. The authors found that the location accuracy of the country-level was above 96%. They also noticed that the location accuracy was quite low for the city-level (40 km), approximately 20%.

The CAIDA (Center for Applied Internet Data Analysis) technical report [10] studied IP geolocation accuracy at the country-level granularity. The report focused on the RIR (Regional Internet Registers) database and compared it against the MaxMind/GeoLite database. Both databases performed approximately the same—95% of the IP address space was located identically at the country-level.

The continual evolution of the IP address space (i.e., new IP addresses assignments or reallocations) is a significant problem of any database-based IP geolocation. The vendors claim to lose accuracy of a couple of percent in each month without reflecting these changes. For example, IP2Location reports up to 5% of the IP records being updated each month [18]. Ref. [17] showed that approximately 1–2% of the locations was changed each month for the databases IPligence, HostIP, IP2Location, and NetAcuity.

The related work reported quite different results, mainly dealing with location accuracy. We explain this by the use of the different ground truth datasets without filtering the redundant or wrong entries. This consequently misinterprets the results. Some studies also used 'virtual' ground truth data by creating groups of close IP address called PoPs (Points of Presence) [17]. The PoPs were determined by an analysis of the network topology structure and communication latency measurements. Each PoP was assigned a location derived as a compromise of the different location results from the location databases. Such evaluation resulted in a better location accuracy for the city-level in contrast with accuracy being evaluated using other ground truth datasets.

4. Device-Independent Location Using IP Geolocation Databases

IP geolocation databases are commonly used for device-independent location because of the convenient access and the short time of the location process when compared to alternative measurement-based location. The location databases group continuous IP addresses into blocks, usually storing the first and last address of the block (or the block size given by a network mask). The defined blocks have location information assigned. This location information is derived from various sources, such as from Internet registries managed by the Internet Assigned Numbers Authority (IANA), Internet traffic measurement and analysis, mining of public online data (typically web pages), information from GPS devices, and volunteers providing their locations. An example of community-based IP geolocation database is Hostip [19]. More sources are usually combined to improve the location information validity.

4.1. Construction of IP Geolocation Databases

An example construction of a geolocation database is shown in Listings 1 and 2. The listings are based on the free to use database GeoLiteCity by MaxMind. The database uses the CSV format. Both listings were simplified for a better understanding. The first listing shows a sample content of the first file, where the blocks of the continuous IP addresses are defined. The block is delimited by the first IP address and its size is given by the network mask. The next items in the list are the block ID and the bound location information in the form of latitude and longitude.

```
77.87.232.0,119,3077706,50.2391,12.7486
88.101.213.0,120,3073254,49.5985,18.1452
```

Listing 1: Blocks of continuous IP addresses with assigned location information.

```
3077706,EU,Europe,CZ,Czechia,KA,Karlovarsky region,Chodov
3073254,EU,Europe,CZ,Czechia,MO,Moravskoslezsky region,Koprivnice
```

Listing 2: Location information for blocks of continuous IP addresses.

The second listing shows other location information assigned to the block of the continuous IP addresses. It starts with the block ID followed by the continent code and name, country code and name, region code and name, and city name.

4.2. Location Information Provided

The geolocation databases are known for a large diversity of the location accuracy as we show in Section 3. An example of such location diversity for an Internet device is shown in Figure 1. It shows locations obtained from seven free to use location databases:

- GeoLite2 City by MaxMind [20], in the paper referredas GeoLite2,
- IP address to city (low resolution) by DB-IP [21], in the paper referred as IPtoCity,
- DB11.LITE by IP2Location [22],
- Lite Free by IPligence [23],
- hostip [19],
- freegeoip [24],
- software77 [25].

The map in the figure shows that the estimated locations are sometimes far apart. Some estimated locations pointed to the capital city of the country (Prague)—IPtoCity, and DB11.LITE. Other databases estimated only the correct country—Lite Free and software77. In this case, the geographical centre of the country was used. The other two databases estimated the correct region, but the wrong city—GeoLite2, and freegeoip. Finally, the hostip database returned a location outside the correct country.

The location errors are summarized in Table 2. The correct location was at latitude 48.97 and longitude 16.61. The location errors varied from 7 km (GeoLite2) to 201 km (DB11.LITE). Some databases returned directly the coordinates of the estimated location. In the cases when the coordinates were not returned directly, we derived them as the geographical centre of the returned city/region/country, respectively. In the table, we marked these cases as '*'.

Table 2. Accuracy of locations obtained from seven free to use IP geolocation databases.

Vendor/Database	Lat, Lon	Error [km]	Note
MaxMind/GeoLite2	48.98, 16.52	7	Correct region
DB-IP/IPtoCity	50.08, 14.47 *	190	Capital city
IP2Loc./DB11.LITE	50.09, 14.42	201	Capital city
IPligence/Lite Free	49.75, 15.5 *	118	Centre of country
hostip	50.0, 18.47	177	Wrong country
freegeoip	49.0, 16.86	19	Correct region
software77	49.75, 15.5 *	118	Centre of country

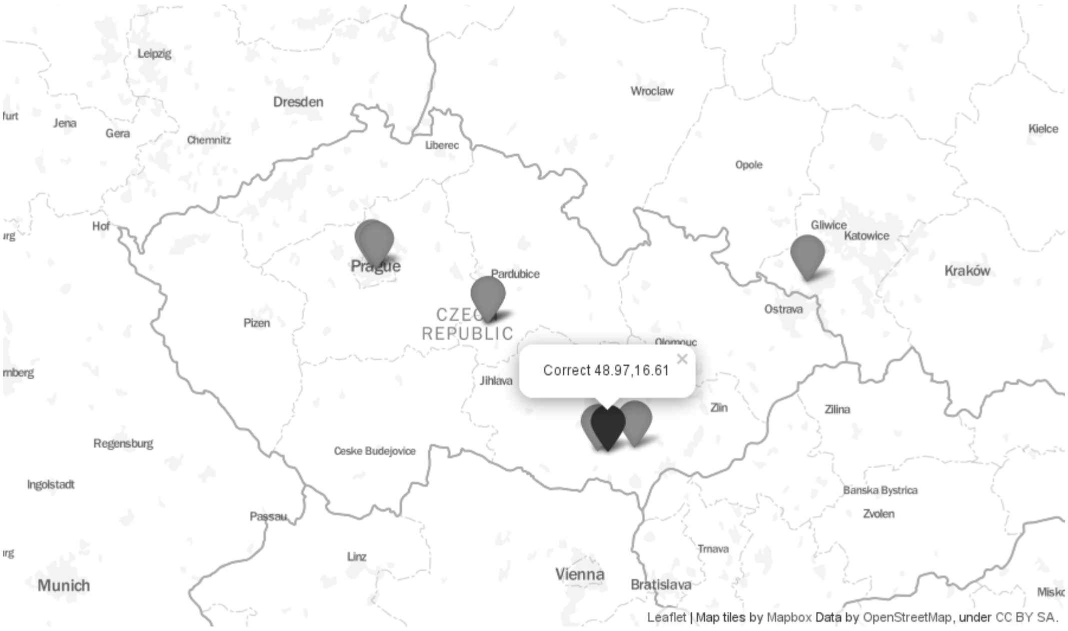


Figure 1. Locations obtained from seven free to use IP geolocation databases (map shows regions).

The databases vary in what information they provide. Table 3 gives an overview of the location information provided by each database. GeoLite2, DB11.LITE, and freegeoip provide a full list of geographical information, i.e., country, region, city, and coordinates. The database IPtoCity does not provide the coordinates and the database hostip does not provide the region. The databases Lite Free and software77 provide only country.

Table 3. Location information provided by IP geolocation databases.

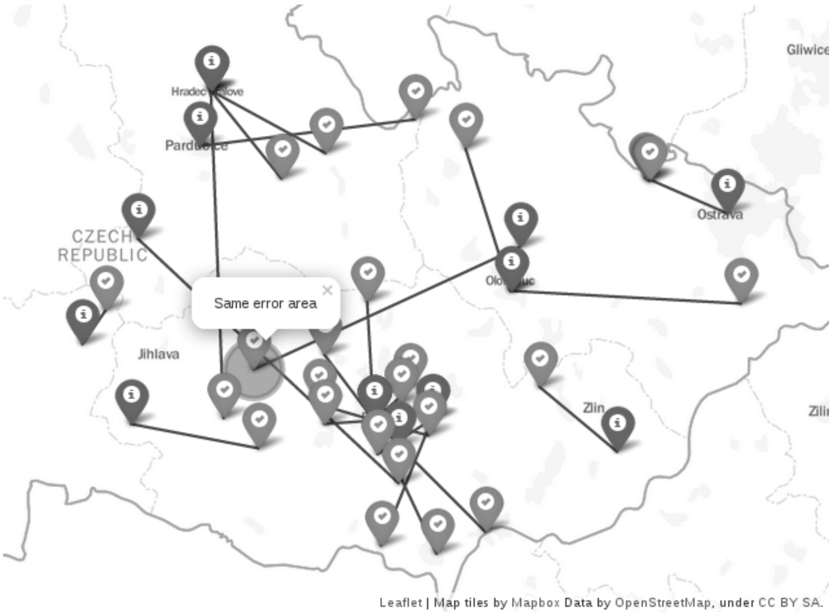
Vendor/Database	Country	Region	City	Coordinates
MaxMind/GeoLite2	x	x	x	x
DB-IP/IPtoCity	x	x	x	-
IP2Location/DB11.LITE	x	x	x	x
IPligence/Lite Free	x	-	-	-
hostip	x	-	x	x
freegeoip	x	x	x	x
software77	x	-	-	-

Geolocation databases can be accessed remotely or locally. Remote access is preferred for a low number of location lookups as each lookup generates traffic and is delayed. In this case, the location service provider maintains the database to be updated. Usually, the number of location queries is recorded and the service provider charges the customers based on these numbers.

5. Construction of Ground Truth Dataset

We constructed the ground truth dataset by a long-term collection of the correct locations of mobile WiFi devices. For each device, we stored its public IP address, ISP, country, region, city, city population, and the geographical coordinates. The coordinates were obtained through in-built GPS. The network-related information was obtained using a developed application run on the collection devices. The devices were typically connected through WiFi access points with NAT (Network Address Translation). The application therefore resolved the public IP address of the WiFi access point. Considering that device-independent IP location works with a resolution of kilometres and the WiFi typical coverage is of a hundred of meters, we linked the trusted GPS location of the device with the public IP address of the WiFi access point. The application run on the devices reported the data to our collection server. The reported data was processed to resolve city, region, and country. We obtained this information using reverse geocoding of the coordinates. The ISP of the devices was taken from the RIPE NCC (Reseaux IP Europeens Network Coordination Centre) WHOIS database based on the public IP address. Finally, the demographic information for cities was obtained through Natural Earth [26] using developed scripts in Python.

The databases may return the same location for a set of devices belonging to the same ISP. This has a detrimental effect on location accuracy evaluation as the results are shifted towards the values for such devices. We therefore particularly focused on the ISP geographical distribution. We inspected the ISPs in each city involved in the ground truth dataset. If we found more devices within a city and belonging to the same ISP, we filtered out the additional devices in the same city to leave only one ISP/city device in the dataset. For the devices found in the country and belonging to the same ISP, we used a minimal distance of 10 km between the devices. If the distance was smaller, we again filtered out the additional devices. We set this distance empirically after inspecting the relation of the ground truth devices (correct locations) and the estimated locations. An example of such an area given by this distance is shown in Figure 2. The figure shows the correct locations (OK-sign icon) and the estimated locations (info-sign icon). The particular pairs—correct and estimated locations—are connected with a line. The highlighted area labelled as 'same error area' shows the place at which the additional unfiltered nodes were estimated to the same location (connected with the line).



Leaflet | Map tiles by Mapbox Data by OpenStreetMap, under CC BY SA.

Figure 2. Same error area.

We collected the ground truth devices for three years and recorded more than 3000 locations in total. From the result of ISP geographical distribution processing, we obtained data for approx. 600 devices in 234 different cities in Europe. Therefore, the presented results should not be influenced by using multiple devices belonging to the same ISP in a city.

ISPs may reassign the same IP address to a different device which might be (or might not be) far apart from the geographical location of the previous device. As described in the related work, the databases change up to 1–5% of their location records each month. For example, IP2Location reports up to 5% of the IP records being updated each month [18], and Ref. [17] showed that approximately 1–2% of the locations was changed each month.

We approached this problem by locating the nodes on a one-by-one basis when new nodes were added to the ground truth. The particular steps were: (i) a trusted location was reported by a community member who contributed to the ground truth dataset. A mobile GPS device with our developed application was used. The GPS location and the relevant networking data was immediately transferred to our collection server. (ii) On the collection server, once a day another developed application was run. This application located the newly reported IP addresses using the location databases. The application also calculated the location error, city/region/country correctness, and other relevant data. The results were then merged with the previously stored data. This way, we worked with the up-to-date correct locations and eliminated the problem of IP address re-allocations. We also updated the locally-stored databases on a monthly basis to keep them up-to-date.

The databases typically locate the devices from large cities with a better accuracy. We therefore also focused on small cites and, as a result, the ground truth dataset covers a range of cities with a different population. The city population details are shown in Figure 3. The median city population in the dataset is 17,000. The smallest city has 64 inhabitants and the largest city has 1.8 million people.

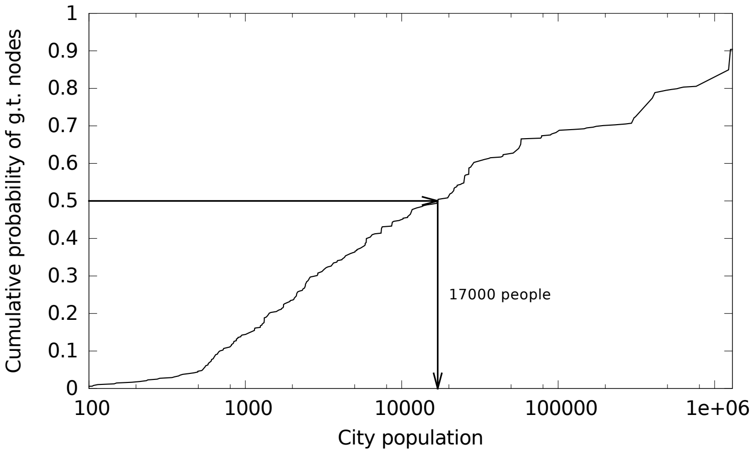


Figure 3. Groundtruth devices in cities with different population.

6. Metrics for Performance Evaluation

We approach the performance evaluation by the use of an extended set of metrics. We also define our additional metrics to particularly study the important properties of the databases. Where applicable, we compare the databases with an alternative measurement-based location.

First, we analyse whether a location for the requested device is returned or not, i.e., location efficiency. The returned location can be country, region, city, or coordinates. The location efficiency is also compared with the measurement-based location. It is not differentiated between the correct and wrong locations returned and we do not measure the distance error in this case.

Next, we deal with location accuracy. We work with the accuracy as being discrete and quantitative. With discrete accuracy, we inspect if the country/region/city returned is correct. The databases differ in the provided discrete location information as shown in Table 3. For quantitative accuracy, we calculate the error distance between the correct and estimated coordinates. We consider these two particular points for obtaining the coordinates:

(i) Some of the databases do not return the coordinates directly. Table 4 shows our approach in obtaining them in such cases. In the table, we marked the directly returned coordinates with the letter 'a'. The letter 'b' denotes that we calculate the coordinates as the geographical centre of a city returned. If a city is not returned, the centre of the returned region is used. If a region is not returned, we consider the coordinates to be unknown. The letter 'c' denotes the situation when the virtual coordinates of the centre of a country are returned.

Table 4. Source of estimated device coordinates.

Vendor/Database	Country	Region	City
MaxMind/GeoLite2	-	-	a
DB-IP/IPtoCity	-	b	b
IP2Location/DB11.LITE	-	-	a
IPligence/Lite Free	c	-	-
hostip	-	-	a
freegeoip	-	-	a
software77	c	-	-

a—directly returned; b—centre of city or region; c—centre of country (not included in evaluation).

(ii) The measurement-based location does not return the coordinates, but it results in a geographical area where the device is located [27]. With this type of location, communication latency from a set of landmarks (with known geographical position) is measured to the device located. As there exists a positive correlation between latency and geographical distance [28], the measured latencies are converted to geographical distances that define the radius of the great-circles around each landmark. The intersection of the circles delimits the area of the device location as shown in Figure 4. The figure is a sample output from our developed measurement location system. Our landmark software is run on PlanetLab [7] servers situated in Europe [29]. We use the centre of gravity of the intersected area to obtain the specific coordinates, as it was introduced in [27].

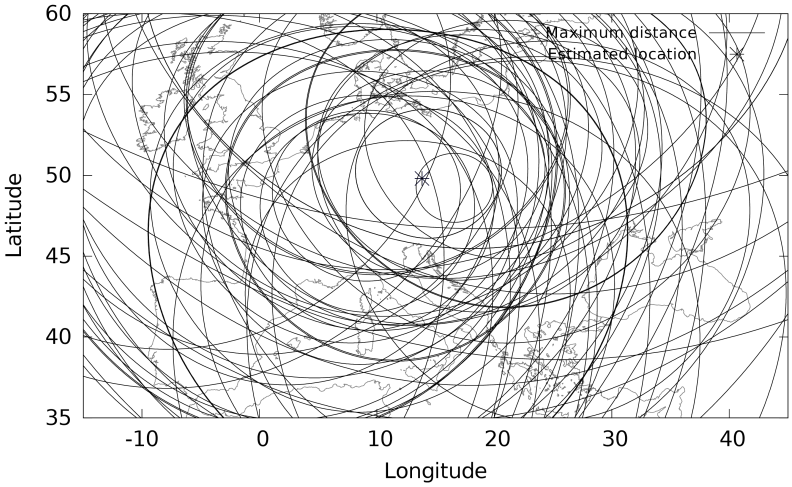


Figure 4. Used latency-based geolocation.

Our first metric is accuracy dependence on city population. As discussed before, the database vendors claim to locate devices from highly populated areas with a better accuracy. However, the low and high population density differences are not given or studied in the related work we found. We divide the cities in the ground truth dataset into two sets—small (less than 17,000 people) and large (more than 17,000 people) based on the median city population shown in Figure 3.

Our second metric is the trustworthiness of the databases. As the results in Section 7 show, some databases return a location for almost any device. It seems therefore that the database performs well. However, some of the returned results are not correct. This might lead to false conclusions about the location performance. The trustworthiness of the databases indicates how much the results can be trusted.

7. Results and Discussion

7.1. Number of Returned Locations

The results on location efficiency are shown in Table 5. All the databases almost always returned a country, except the hostip database that returned a country for only half of the ground truth devices requested. The probable reason for such low efficiency of the returned countries is that hostip is constructed using locations manually submitted on a voluntary basis by Internet users. These locations are submitted for individual IP addresses or for a range of IP addresses.

Table 5. Location efficiency—returned correct or incorrect locations [%].

Vendor/Database	Country	Region	City	Coordinates
MaxMind/GeoLite2	99	38	45	99
DB-IP/IPtoCity	100	100	100	100
IP2Location/DB11.LITE	99	92	99	99
IPligence/Lite Free	99	-	-	-
hostip	53	-	16	49
freegeoip	100	39	43	100
software77	99	-	-	-
latency-based	-	-	-	44

The databases DB-IP/IPtoCity and IP2Location/DB11.LITE performed well for the regions and cities. We observed that many of the returned locations point to the capital city of a country or to the country's geographical centre. If the city is not known for the device requested, these databases approximate the result by these two types of location.

The database MaxMind/GeoLite2 in some cases did not return a region, but, at the same time, a city was returned. The same phenomena happened for the freegeoip database. The reason is that Freegeoip is based on the freely available MaxMind/GeoLite2 database.

Table 5 also shows the percentages for the returned coordinates. The majority of the databases returned coordinates for 99% to 100% of the location cases. However, latency-based geolocation provided coordinates for less than half of the devices located. The reason is that the ICMP traffic for latency measurements from the landmarks to the devices located may be filtered by the networking devices en route (routers, firewalls) or by the end devices. The efficiency of latency-based geolocation was therefore lower than with all the databases except hostip, which gave a similar result.

7.2. Distance Error

Distance errors between the estimated and correct coordinates stored in the ground truth dataset are shown in Table 6. The table shows the median and average errors that are in the 30–150 km and 80–270 km ranges, respectively. Again, the community-filled database hostip performed the worst. A comparison of databases with latency-based geolocation shows the biggest difference in the first quartile of location error. A detailed study of these differences is shown in Figure 5. The cumulative probability function shows that latency-based geolocation gave low location errors (less than 10 km) for only a small percentage of the cases.

The probable reason for this behaviour is given by the actual distance from the measurement landmarks to the ground truth devices. When a landmark is close the device measured, the error is low, as the centre-of-gravity of a single great circle around this landmark is used as the returned location. In other cases, when multiple intersections of the great circles occur, the error distance is bigger. Figure 6 specifically shows such a case where the correct location is close to the borders of the area that was a product of multiple great circle intersections. The error distance in this case was 137 km.

Table 6. Location accuracy—distance error of returned coordinates [km].

Vendor/Database	1st Quartile	Median	3rd Quartile	Average
MaxMind/GeoLite2	7	74	145	86
DB-IP/IPtoCity	6	55	193	133
IP2Location/DB11.LITE	5	35	171	105
hostip	84	149	219	263
freegeoip	7	81	167	95
latency-based	121	182	230	167

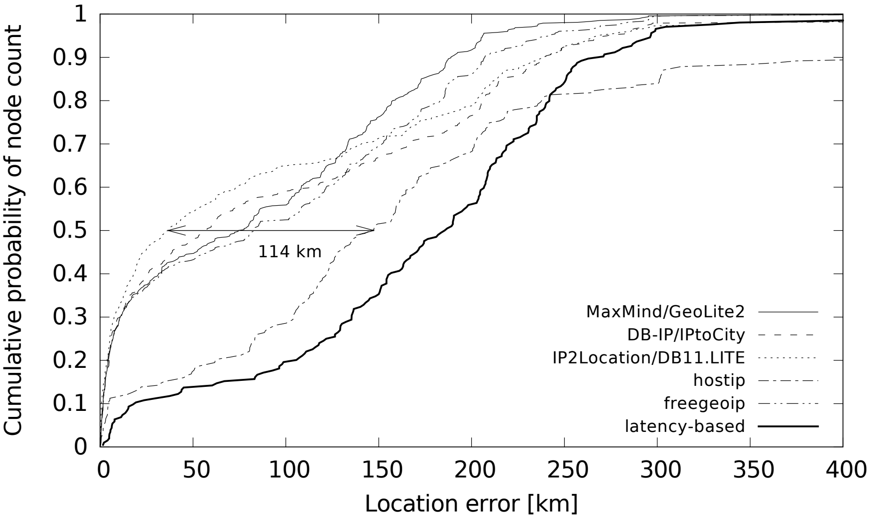


Figure 5. Location accuracy—distance errors.

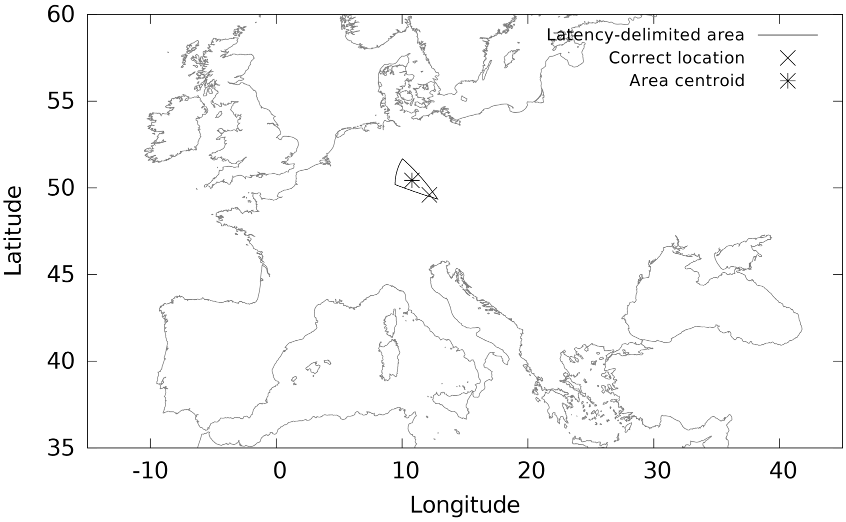


Figure 6. Delimited area as a product of multiple great-circle intersections.

With distance error analysis, we also include a comparison with another ground truth dataset to demonstrate the differences between a filtered dataset covering all sizes of cities and an unfiltered dataset covering only major cities. For this purpose, we use the PlanetLab ground truth dataset [30] available from [7]. The nodes of this dataset come from major cities as they are run by universities and large companies. The dataset used covers nodes in Europe. We used only the PlanetLab nodes with a valid IP address and correct location [31]. We located the ground truth nodes using the location databases that provide coordinates for their estimations. The results are shown in Figure 7.

The cumulative probability function for each database shows a better location accuracy compared to results from the filtered ground truth dataset. The reason is that major cities are included in the PlanetLab dataset. This dataset also covers multiple nodes in some cities assigned to the same ISP. The results are therefore also influenced towards better values by multiplication of low location errors for these same city/ISP nodes.

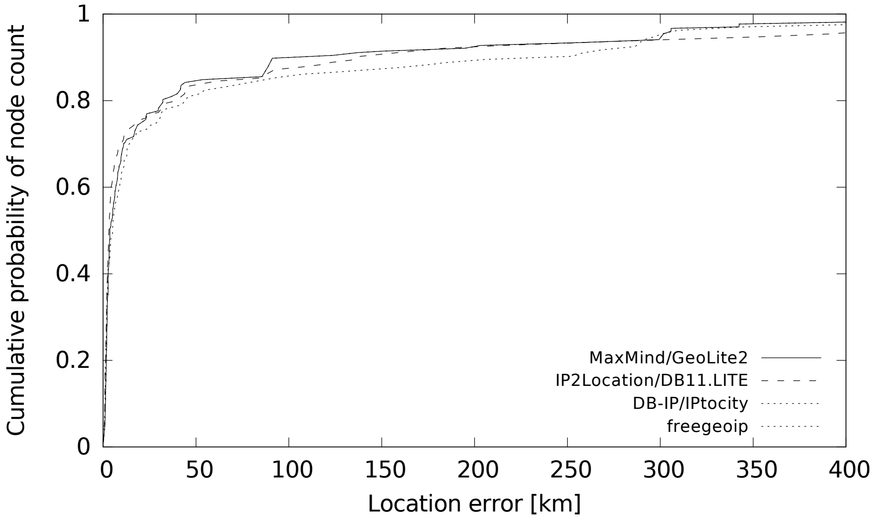


Figure 7. Location accuracy for ground truth dataset with major cities.

Table 7 shows the details of differences between the two ground truth datasets. It shows that the median location error was reduced from tens of kilometers to units of kilometers. The average dropped approximately to one half. This clearly shows that the size of the cities along with the ISP assignation is important to consider as the results can be very different.

Table 7. Differences in location accuracy between ground truth covering all cities and ground truth with major cities.

Vendor/Database	All Cities, ISP Filtered		Major Cities, ISP Unfiltered	
	Median	Average	Median	Average
MaxMind/GeoLite2	74	86	4	44
DB-IP/IPtoCity	55	133	5	53
IP2Location/DB11.LITE	35	105	3	53
freegeoip	81	95	4	45

7.3. Effect of Small and Large Cities on Location Accuracy

We further study the effect of small and large cities on location accuracy and efficiency. We divided the cities to small and large to demonstrate the differences. The returned locations for small and large cities are shown in Table 8. The numbers indicate that the locations for devices in large cities were returned with approximately the same efficiency compared to small cities. An exception is the community-filled hostip database that showed a 10% location efficiency difference between small and large cities. Considering the returned coordinates, the database-based and latency-based location did not show any significant dependence on city population.

Table 8. Location efficiency—returned locations for small/large cities (p < 17,000/p > 17,000) [%].

Vendor/Database	City	Coordinates
MaxMind/GeoLite2	44/46	100/99
DB-IP/IPtoCity	100/100	100/100
IP2Location/DB11.LITE	99/98	100/99
IPligence/Lite Free	-	-
hostip	11/21	42/56
freegeoip	41/46	100/100
software77	-	-
latency-based	-	45/43

On the other hand, there were differences in error distances between large and small cities. Table 9 shows the differences between the median distance errors. The majority of the databases showed a positive trend in returning the coordinates with a lower median distance error for the devices in large cities. DB-IP/IPtoCity and freegeoip showed the greatest change in distance error. Latency-based location showed a low error difference between large and small cities.

Table 9. Location accuracy—median distance error for cities with population p.

Vendor/Database	p < 17,000	p > 17,000	Diff.
MaxMind/GeoLite2	80	39	41
DB-IP/IPtoCity	103	14	89
IP2Location/DB11.LITE	64	10	54
hostip	131	172	-41
freegeoip	113	33	80
latency-based	196	154	42

The distribution of error distances for small cities is plotted in Figure 8 and for large cities in Figure 9. The maximum difference of the median distance error was 67 km for small cities and 162 km for large cities. An interesting point is the sharpness of the function in Figure 9 for low location errors. It shows that the devices from large cities were located with a very low location error below 10 km for around 40% of the cases. This holds for all the databases except hostip. The hostip database and latency-based location gave such a low error only for less than 20% of the location cases. If we compare this with the situation in small cities in Figure 8, the databases (except hostip) gave low location errors below 10 km for around 20% of the cases. Latency-based geolocation did not give results in this error range at all.

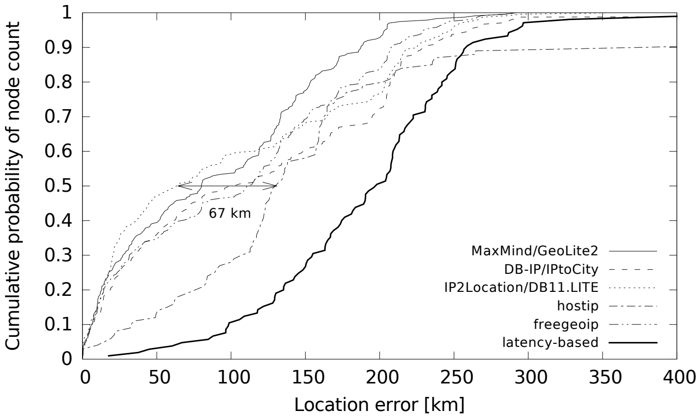


Figure 8. Location accuracy—distance errors for small cities with population p < 17,000.

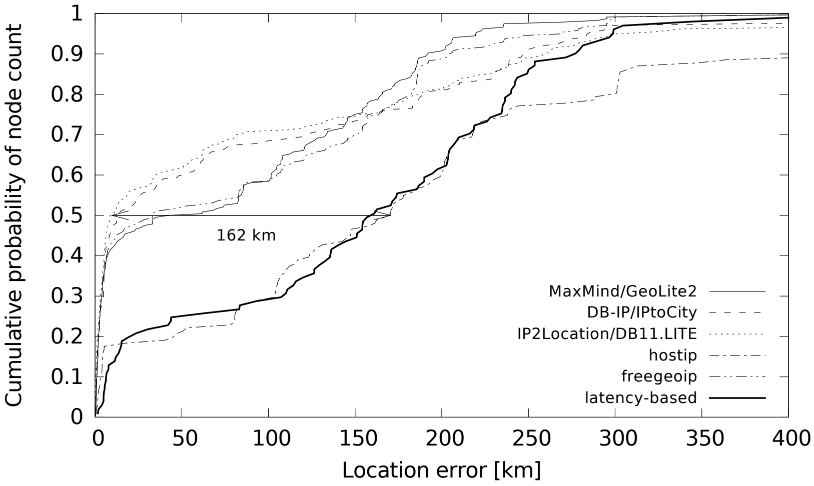


Figure 9. Location accuracy—distance errors for large cities with population $p > 17,000$.

7.4. Trusted Locations?

Figure 10 shows the performance of the databases for discrete location accuracy where we inspected if the country/region/city returned was correct. With discrete accuracy evaluation, we faced a problem of place names (cities and regions) as they are not the same in different languages. Some of the location databases return different names for the same places and we needed to process them to find a correct city/region match. We therefore developed matching replacements for different languages to find proper matches.

All the databases performed well at the country level, except the hostip database that responded with a correct country for only around 40% of the cases. Much worse results happened at the region level. To further discuss the trustworthiness of the returned locations at the region and city level, we link Figure 10 to Table 5. We note the numbers from the table in brackets.

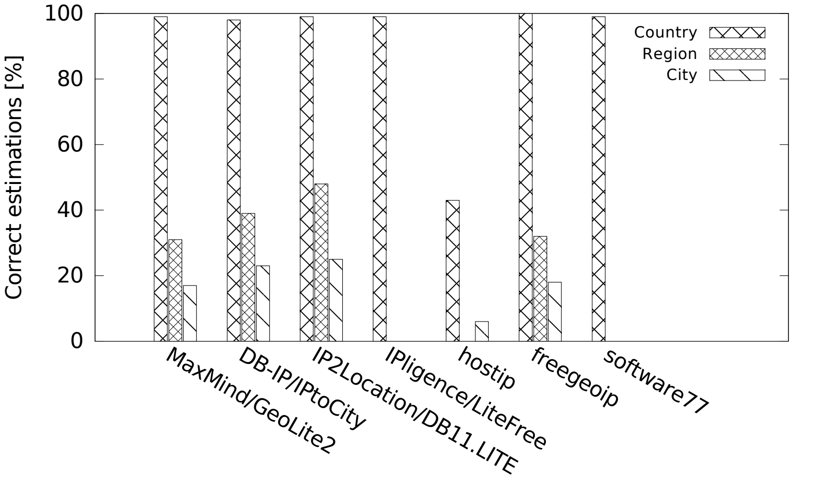


Figure 10. Correct discrete locations.

The database MaxMind/GeoLite2 returned a correct region for more than 30% of the location cases (38% of the regions returned were either correct or incorrect). Freegeoip returned similar results: 32% (39%). These numbers show that the databases MaxMind/GeoLite2 and freegeoip were highly accurate with the returned locations at the region level, and such results can be fairly trusted.

On the other hand, DB-IP/IPtoCity returned a correct region for around 40% of the location cases (100% of the regions returned were correct or incorrect). Similar results were for IP2Location/DB11.LITE—50% (92%). This indicates a low trustworthiness of the regions returned.

The same tendency continued with the city level. The databases DB-IP/IPtoCity and IP2Location/DB11.LITE returned most of the locations incorrect, about 70–80%.

Based on these large differences, we use an additional metric to show the 'trustworthiness' of the returned locations. Figure 11 shows that 'one can trust' the results provided by the databases MaxMind/GeoLite2 and freegeoip as they show the lowest percentages of incorrect results. We note that the hostip database is not included in this evaluation. The reason is that its location efficiency was very low (only 16% of returned cities correct or incorrect). The evaluation of such a low number of returned locations would not give a fair result when compared to the other databases with much higher location efficiency.

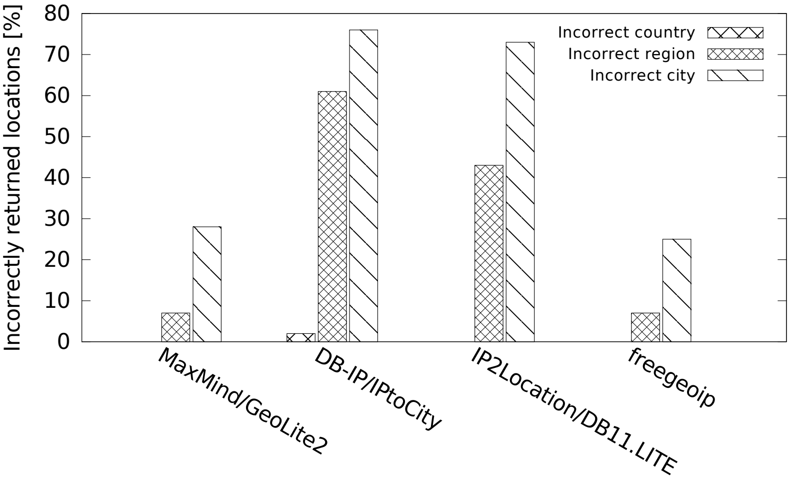


Figure 11. Trustworthiness when providing region or city. Lower percentage means that locations can be trusted more.

8. Conclusions

In the paper, we described the construction of a filtered ground truth dataset. The ground truth dataset differs from those used in the related research by covering cities with both low and high populations. We also processed the ground truth dataset to maintain only devices that follow the rule of one IP address per ISP and per city. This way, the presented results should be error-free as they are not influenced by a multiplication of the same location accuracy values. We obtained the following results:

- The location efficiency (null responses versus returned locations) does not correlate with the location accuracy (correct locations). There are large differences in the correctness of the returned cities and regions. This may lead to confusion when these locations are used without taking this into account. To describe this phenomena, we used a measure showing 'trustworthiness' of the locations returned.
- There are large differences between high and low populated cities. The number of the returned correct regions and cities for devices from low-population cities is low. The maximum percentage of the returned correct regions is around 50%. The maximum percentage of the returned correct cities is less than 30%. The related work gives values from 20 to 80% of the correct city estimations. We observed a decrease of the median error distance from 100 km for low-population cities to 14 km for high-population cities.

- The community-provided location information is not a reliable source for IP geolocation. This information does not cover the whole IP address space, which is indicated by low location efficiency results. In addition, location accuracy is low when this information is used.
- The alternative latency-based location provides worse results compared with the database-based location. In particular, we observed that location efficiency was poor. The reason is that latency measurements are often blocked by networking devices en route or by the end devices. We observed this for over 50% of the devices located.

As can be seen from the results, the current approaches may produce incorrect spatial data. This makes the dependent Internet location-aware services prone to errors. Further research into device-independent location may be motivated by a better use of the available Internet resources for public safety. New widely used Internet-based communication technologies, such as online messaging and VoIP (Voice over IP), may be used in public safety provided that the locations are accurate to a specific level. For example, with VoIP, the location of the caller in emergency is needed to be known in cases when the location cannot be shared due to specific circumstances. These may be housebreaking, kidnapping, and health injuries not allowing the caller to speak, or the caller may be forced to abandon the call. Another issue related to public safety are hoax calls. With better knowledge in the field of device-independent spatial locations, hoax calls could be detected by a comparison of the caller's location to the reported place of trouble [32].

Acknowledgments: Research described in this paper was financed by the National Sustainability Program under grant LO1401. For the research, the infrastructure of the SIX (Sensor, Information and Communication Systems) Center was used.

Author Contributions: Dan Komosny proposed the idea of the paper. He designed the experiments and wrote the programs needed. He also analysed the data. Paul Pang provided the information about the accuracy and efficiency of spatial locations. He also contributed to the topic dealing with location-aware applications. Miralem Mehic and Miroslav Voznak provided data of spatial locations by covering various cities and countries.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chiara, D.; Giovanni, P.; Sebillo, M.; Tortora, G.; Vitiello, G. Geomarketing Policies and Augmented Reality for Advertisement Delivery on Mobile Devices. In Proceedings of the 17th International Conference on Distributed Multimedia Systems. Knowledge Systems Institute, Florence, Italy, 18–20 October 2011; pp. 78–83.
2. Barnes, R.; Lepinski, M.; Cooper, A.; Morris, J.; Tschofenig, H.; Schulzrinne, H. *An Architecture for Location and Location Privacy in Internet Applications*; Request for Comments: 6280 (Best Current Practice); Internet Engineering Task Force: Fremont, CA, USA, 2011.
3. Yen, N.Y.; Huang, R.; Ma, J.; Jin, Q.; Shih, T.K. Intelligent route generation: discovery and search of correlation between shared resources. *Int. J. Commun. Syst.* **2013**, *26*, 732–746.
4. Burget, R.; Komosny, D.; Kathiravelu, G. Topology Aware Feedback Transmission for Real-Time Control Protocol. *J. Netw. Comput. Appl.* **2012**, *35*, 723–730.
5. MaxMind. *Fraud Detection through IP Address Reputation and a Mutual Collaboration Network*; White Paper; MaxMind: Waltham, MA, USA, 2011.
6. W3cook. Available online: www.w3cook.com (accessed on 26 May 2017).
7. PlanetLab. Available online: www.planet-lab.org (accessed on 26 May 2017).
8. Huffaker, B.; Fomenkov, M.; Claffy, K. *Geocompare: A Comparison of Public and Commercial Geolocation Databases*; Technical Report; Cooperative Association for Internet Data Analysis (CAIDA): La Jolla, CA, USA, 2011.
9. Poese, I.; Uhlig, S.; Kaafar, M.; Donnet, B.; Gueye, B. IP geolocation databases: Unreliable? *ACM SIGCOMM Comput. Commun. Rev.* **2011**, *41*, 53–56.
10. Zander, S. *How Accurate Is IP Geolocation Based on IP Allocation Data*? Technical Report 120524A; Swinburne University of Technology: Hawthorn, VIC, Australia, 2012.

11. Moravek, P.; Komosny, D.; Simek, M.; Sveda, J.; Handl, T. Vivaldi and other localization methods. In Proceedings of the 32nd International Conference on Telecommunications and Signal Processing, Gyor-Moson-Sopron, Hungary, 26–27 August 2009; pp. 214–218.
12. Moravek, P.; Komosny, D.; Burget, R.; Sveda, J.; Handl, T.; Jarosova, L. Study and performance of localization methods in IP based networks: Vivaldi algorithm. *J. Netw. Comput. Appl.* **2011**, *34*, 351–367.
13. Guo, C.; Liu, Y.; Shen, W.; Wang, H.; Yu, Q.; Zhang, Y. Mining the Web and the Internet for Accurate IP Address Geolocations. In Proceedings of IEEE International Conference on Computer Communications 2009, Rio de Janeiro, Brazil, 19–25 April 2009; pp. 2841–2845.
14. Fioreze, T.; Heijenk, G. Extending DNS to support geocasting towards VANETs: A proposal. In Proceedings of Vehicular Networking Conference (VNC) 2010, Jersey City, NJ, USA, 13–15 December 2010; pp. 271–277.
15. Moravek, P.; Komosny, D.; Simek, M.; Jelinek, M.; Girbau, D.; Lazaro, A. Investigation of radio channel uncertainty in distance estimation in wireless sensor networks. *Telecommun. Syst.* **2013**, *52*, 1549–1558.
16. Moravek, P.; Komosny, D.; Simek, M.; Girbau, D.; Lazaro, A. Energy Analysis of Received Signal Strength Localization in Wireless Sensor Networks. *Radioengineering* **2011**, *20*, 937–945.
17. Shavitt, Y.; Zilberman, N. A Geolocation Databases Study. *IEEE J. Sel. Areas Commun.* **2011**, *2*, 2044–2056.
18. IP2Location. Available online: www.ip2location.com/data-accuracy (accessed on 26 May 2017).
19. Hostip. Available online: www.hostip.info (accessed on 26 May 2017).
20. GeoLite Legacy Downloadable Databases. Available online: dev.maxmind.com/geoip/legacy/geolite (accessed on 26 May 2017).
21. Database downloads. Available online: db-ip.com/db (accessed on 26 May 2017).
22. Free IP Geolocation Database. Available online: lite.ip2location.com (accessed on 26 May 2017).
23. IPligence Lite Free. Available online: www.ipligence.com/free-ip-database (accessed on 26 May 2017).
24. Freegeoip Net. Available online: www.freegeoip.net (accessed on 26 May 2017).
25. IP to Country Database (IPV4 and IPV6). Available online: software77.net/geo-ip (accessed on 26 May 2017).
26. Natural Earth. Available online: www.naturalearthdata.com (accessed on 26 May 2017).
27. Gueye, B.; Ziviani, A.; Crovella, M.; Fdida, S. Constraint-based geolocation of internet hosts. *IEEE/ACM Trans. Netw.* **2006**, *14*, 1219–1232.
28. Percacci, R.; Vespignani, A. Scale-free behavior of the Internet global performance. *Phys. B Condens. Matter.* **2003**, *32*, 411–414.
29. Komosny, D.; Pruzinsky, J.; Ilko, P.; Polasek, J.; Masek, P.; Kocatepe, O. On Geographic Coordinates of PlanetLab Europe. In Proceedings of 38th International Conference on Telecommunications and Signal Processing (TSP), Vysocany Prague, Czech Republic, 9–11 July 2015; pp. 642–646.
30. Chun, B.; Culler, D.; Roscoe, T.; Bavier, A.; Peterson, L.; Wawrzoniak, M.; Bowman, M. PlanetLab: an overlay testbed for broad-coverage services. *ACM SIGCOMM Comput. Commun. Rev.* **2003**, *33*, 3–12.
31. Komosny, D.; Mrdovic, S.; Ilko, P.; Grejtak, M.; Pospichal, O. Testing Internet applications and services using PlanetLab. *Comput. Stand. Interfaces* **2017**, *53*, 33–38.
32. Voznak, M.; Rezac, F. The implementation of SPAM over Internet telephony and a defence against this attack. In Proceedings of the 32nd International Conference on Telecommunications and Signal Processing, Gyor-Moson-Sopron, Hungary, 26–27 August 2009; pp. 26–27.