Bachelor thesis
Computing Science

Radboud University

Validating the accuracy of the MaxMind GeoLite2 City database

Author:
Mike Schopman
s1007619

First supervisor:
Dr. Ir. H.P.E. (Harald) Vranken
harald.vranken@ou.nl

Second assessor:
Dr. K. (Katharina) Kohls
kkohls@cs.ru.nl

June 3, 2021

Abstract

Mapping a given Internet Protocol address to a geographical location is called IP Geolocation. There are several companies that provide IP Geolocation databases. These IP Geolocation databases contain IP address-location pairs. The GeoLite2 City database from the company MaxMind is one of those databases. This research focuses on validating the accuracy of the MaxMind GeoLite2 City database.

In order to validate the accuracy of the GeoLite2 City database, we assemble a set of IP addresses through the collection of probes from the research institute RIPE Atlas. These IP addresses are cross referenced with the GeoLite2 City database and we note the location differences. Based on these differences we calculate how many IP addresses are correctly located within a given radius. We do these calculations for three different radii, namely a 10 kilometer radius, a 50 kilometer radius and a 250 kilometer radius. The number of correctly located IP addresses for each radius represents the accuracy of that radius.

In the process of comparing the accuracies that we obtained and the accuracies reported by MaxMind, we found that most of the accuracies reported by MaxMind fall outside the 99% confidence intervals that we calculated for our accuracies. If we combine the results for all three radii, then in 72% of the cases, the accuracy reported by MaxMind falls outside the 99% confidence intervals. Furthermore, for all three radii, the average accuracies that we obtained were below the average accuracies of MaxMind. The accuracy difference for the 10 kilometer radius is 4.2% and the differences for the 50 kilometer and the 250 kilometer radii are 11.7% and 11.1% respectively.

These numbers indicate that, according to our research, the accuracy of the GeoLite2 City database as reported by MaxMind does not match the accuracy that we obtained. This means that the correctness of the accuracy reported by MaxMind is questionable.

Contents

Chapter 1

Introduction

The discipline of mapping a given Internet Protocol (IP) address to a geographical location is called IP geolocation. IP geolocation is an area which has been researched for quite some years already. In 2001, Padmanabhan et al. [A1] were one of the first to design three distinct techniques to try to address this problem. Current day, IP geolocation is already very sophisticated and it is being used in a wide range of services. A few of these services are: targeted advertising, displaying regional weather, geoblocking and online fraud detection.

In order for these services to work properly, the used IP geolocation method must be reliable. In other words, the geographical location of the IP address must be correct. One of the methods to obtain IP geolocation data is to make use of a database. These databases are hosted by certain companies, one of which is called MaxMind. MaxMind is a provider of IP intelligence and online fraud detection tools. [C17] MaxMind provides free and paid databases which contain location information about IP addresses. The difference between the paid and the free database is that the paid database is a more accurate version of the free database. The paid database contains more IP addresses and is updated more frequently. One of the free databases is called GeoLite2 City and it contains rough coordinates of a given IP address, which translates to a city.

In order for the database to be reliable, the coordinates attached to the IP address must be from the same city where the IP address is actually located. Unfortunately, IP addresses can change location, so it is difficult to maintain a correct database. This database must be checked and updated on a regular basis to remain reliable. One of the key points of a reliable database is a high accuracy. This means that a high percentage of the IP addresses is mapped to the correct location. The higher the accuracy, the greater the chance that the location of an IP address is correct.

MaxMind is open about their accuracy. They have a large table in which they list the accuracy of both their paid and their free databases. [C18] The accuracy is listed per country and it is possible to check this accuracy for different radii. This raises the question:

- Is the accuracy of the MaxMind GeoLite2 City database correct?

This is a relevant question, because the data within the database is always changing. The world of IP addresses itself is ever changing and therefore the IP geolocation databases need to be held up to date. This means that time plays a crucial role and therefore results from previous research may be out of date.

In order to answer the research question, we first have to find out what information is stored in the GeoLite2 City database. After we know what the structure of the database is, we have to collect IP address-location pairs to build a reliable ground truth. The ground truth will contain both IPv4 and IPv6 addresses and both the country and city in which the IP address is located. When the ground truth is assembled and processed such that we are only left with accurate data, we can start looking up the IP addresses from the ground truth in MaxMind's GeoLite2 City database. If an IP address match is found, we can compute the difference in location between the coordinates of the two IP addresses. Based on the accuracy radius given by MaxMind, we can check whether or not the location difference is indeed within the bounded radius. These calculations will be repeated for both the IPv4 and the IPv6 address set. The calculated results will be analyzed and based on the number of correct locations per country given a certain accuracy radius, we can estimate an accuracy of the GeoLite2 City database. This accuracy will only be an approximation, because we use a rather small dataset compared to all allocated IP addresses.

Ideally the accuracy that we calculate is somewhat the same as the provided accuracy by MaxMind. If our accuracy is higher, then this is good for MaxMind, but if our accuracy is much lower, then this may need to be researched even deeper.

Chapter 2 contains some background information in order to understand this research. In chapter 3 the collection and processing of the data is explained. Chapter 4 deepens on the approach of calculating and analyzing the location differences. In chapter 5 the results of the experiments are given. Chapter 6 reviews and compares some related work. Chapter 7 discusses the results and some assumptions that were made. Finally, chapter 8 concludes the research.

Chapter 2

Preliminaries

2.1 Internet Protocol

The Internet Protocol (IP) is designed for use in a network of systems in order for them to communicate. [B13] The IP provides for transmitting datagrams from source to destination, which are identified by fixed length addresses, also known as IP addresses. An IP address comes in two versions as of 2004, namely the versions IPv4 and IPv6. All IPv4 addresses ran out in February 2011 [C19], therefore IPv6 was introduced. IPv6 is an upgraded version of IPv4, because it uses 128 bits instead of the 32 bits of IPv4. This creates a much larger address space. [B14] IPv6 was deployed to tackle the problem of IPv4 exhaustion. Although all IPv4 addresses have long run out, it is still the most used version today. This meant that in order to continue using IPv4, its lifetime was extended by using tricks like dynamic IP addresses and NAT. [B15]

2.1.1 IP Address Distribution

The Internet Assigned Numbers Authority (IANA) is responsible for global coordination of the Internet Protocol addressing systems, as well as the Autonomous System Numbers used for routing Internet traffic. [C20] IANA is the head of the hierarchical system. IANA allocates pools of unallocated IP addresses to Regional Internet Registries (RIRs). There are five RIRs, namely AFRINIC (Africa region), APNIC (Asia/Pacific region), ARIN (Canada, USA and some Caribbean Islands), LACNIC (Latin America and some Caribbean Islands) and RIPE NCC (Europe, the Middle East and Central Asia). Each of these RIRs is responsible for further allocation to National Internet Registries (NIRs), Local Internet Registries (LIRs) and to the Internet Service Providers (ISPs). Finally, ISPs assign IP addresses to their customers, such that every system connected to the Internet has an IP address.

2.1.2 Static vs Dynamic IP Address

ISPs can make use of two types of IP addresses, static and dynamic IP addresses. The difference between a static and a dynamic IP address is that a static IP address can not change whereas a dynamic IP address can. A static IP address means that whichever IP address was assigned to your system, it will always remain the same. This is particularly useful for large businesses and for dedicated services such as FTP and web servers. A dynamic IP address is most of the time assigned to households. ISPs use this dynamic allocation in order for them to manage the IP addresses in a useful way and to reuse them when possible.

Speaking in terms of IP geolocation, the location of a static IP address will be more stable than a dynamic IP address. This is because a static IP address will stay at the same location, but a dynamic IP address can be transferred from one system to another, possibly situated at another location. The old IP geolocation is now invalid and thus becomes less reliable.

2.1.3 Network Address Translation

Another feature to extend the IPv4 lifetime is the usage of Network Address Translation (NAT). [B15] By using NAT, you can reuse IP addresses within a personal environment. On the most basic level, every router contains a NAT table. The router has a unique IP address that can be seen from the outside, but every system on the inside that is connected to the router has a local IP address. This means that not every local system connected to the router needs to have a unique IP address, which saves a lot of addresses.

This also means that systems from the local network can not be separated from each other, since they all share the same unique IP address seen from the outside. For IP geolocation this means that only the router can be identified and not the individual systems behind the router.

The same approach is also used on a larger scale. This is called Carrier-Grade NAT (CGN). [B16] CGN is used between the ISP and the public internet. In this case, the entire ISP network is local and every resident on the inside can be given an arbitrary IPv4 address. CGN couples this to the public internet by a unique IP address. Implementation of CGN is optional for ISPs, but when they choose to do so, it is called NAT444. This is because it passes through three different IPv4 address domains, namely the resident's private network, the carrier's private network and the public internet.

When CGN is implemented by an ISP it will cause problems for IP geolocation. It means that one IP address can be located at different residents, basically resulting in the fact that one IP address has multiple locations. This means that the chance of displaying the correct location lowers, as there are more options to choose from.

2.2 IP Geolocation

As stated in the introduction, IP geolocation is the area in which a geographic location is deduced from an IP address. By itself, an IP address does not tell anything about a location, it is just a 'number' that references to a specific system. There are several methods in obtaining the geolocation, each with their own advantages. In order to estimate the location of an IP address, multiple methods have been proposed over the years.

2.2.1 Landmarks

A landmark is a system with a known IP address and a known location. Landmarks are used for delay measurements. It is possible to probe a target system from a known landmark to get these delay measurements. This gives the advantage that it is now known how far both systems are situated from one another in units of delay.

Landmark Problem

According to Hillmann et al. the landmark problem is the dilemma of using as much landmarks as necessary but as few as possible. [A2] Most of the IP geolocation methods rely on active landmarks, the more landmarks available, the better the accuracy. Whenever there are few landmarks around, or whenever the landmarks are not evenly distributed, the inferred location will be inaccurate. This means that in order to achieve a high accuracy, there must be a sufficient number of landmarks around.

2.2.2 Methods of IP Geolocation

CBG

Constraint-Based Geolocation (CBG) was proposed in 2006 by Gueye et al. [A3] This method determines the location of an IP address by making use of multilateration. Multilateration is the process of inferring a location based on a sufficient number of distances to some fixed points. These distances are calculated by measuring the delay between the target IP address and a landmark. This delay can be converted to a rough distance, drawn as a circle around the landmark. The intersecting area of all circles must contain the location of the target.

TBG

Topology-Based Geolocation (TBG) was also proposed in 2006 by Katz-Bassett et al. [A4] TBG makes use of the traceroute tool to obtain a network topology. Traceroute can be used on the landmarks to probe each other and to probe the target IP address. This will provide round-trip measurements

to the target and it will identify the intermediate network interfaces. The key of this method is that it uses end-to-end delays, inferred per-hop latencies and the network topology to locate the target.

Octant

Octant was proposed by Wong et al. one year later, in 2007. [A5] Octant improved the median accuracy, because it was two times more accurate than previous methods. Octant became the new best framework. It combines adaptions of both CBG and TBG. The advantage of Octant is that next to using positive constraints, is also uses negative constraints. A positive constraint is an area in which the target must be located and a negative constraint is an area in which the target is definitely not located. Combining these two constraints can result in a smaller area than only using positive constraints. On top of that, Octant uses Bézier Curves to represent large and complex areas in a precise way.

In the following years, more and more methods extending on these concepts were introduced that reduced the accuracy error, examples are Posit [A6], Street-Level [A7] and Dragoon. [A2]

2.2.3 Databases

All previously mentioned methods are so called active geolocation methods, because they use algorithms to infer a location. [A8] A completely different method is the so called passive method, which is the database driven geolocation method. This method makes use of a database in which IP addresses are collected. The database consists of blocks of addresses, also known as prefixes, that have a location attached to them. These databases are partially filled by hand and they need to be updated on a regular basis. There are many providers of databases, one of which is MaxMind. [C17] An advantage of a database is that they contain many IP addresses and that the location of a given IP address can be called very fast. On the other hand, the disadvantage is that they need to be updated very often, because of the dynamic IP address problem.

2.3 Ground Truth

The ground truth is a set of IP address-location pairs from which we know that the location of the IP address is correct. The ground truth is important to be correct, since it will be used to compare other locations against. Given the fact that the ground is correct, any deviations from the ground truth are labeled as incorrect. This means that by using the ground truth in an experiment, it is possible to define an approximate correctness of another dataset.

2.4 Usage

As shortly stated in the introduction, IP geolocation is used for many online services. Here are a few of these services explained in order to show the importance of correct IP geolocation data.

- *Targeted advertising*
  Targeted advertising uses IP geolocation data to check in which country or city the user currently is. [C21] This provides information about what language to use, which location based recommendations to give and to promote local companies.

- *Geoblocking*
  Geoblocking is used as a measure to block users from certain continents, countries or regions. [C22] This is for example applied to websites that may only be seen by specific users, blocking them to be seen from the outside. In this case it is about region specific content. It is also used by streaming services. Some series and movies may only be seen in certain countries. In this case we are talking about hiding content for legal purposes.

- *Online fraud detection*
  Another important usage of IP geolocation is with online fraud detection. [C23] If an online company sells goods over the internet, then they also need to handle online payments. If for example the billing address differs a lot from the location of the IP address, then this could be seen as a fraudulent attempt.

Chapter 3

Data Handling

3.1 GeoLite2 City

We start by looking into the GeoLite2 City database. Although MaxMind also offers databases like a country database and an autonomous system database, the city database is what we are most interested in. They offer two types of this database. The first type is a binary MMDB database, which stands for *MaxMind Database*. This database can be queried locally when downloaded from their website. We downloaded the MMDB database on January 25, 2021 and it will be used to perform the comparisons in chapter 4. The second type is a CSV database, which stands for *Comma Separated Value*. This database can be opened by a supported program in which it is possible to see the actual values.

We use the CSV database to check which information is stored in the database. GeoLite2 City has a lot of information stored per IP address. In the database, both IPv4 and IPv6 addresses are stored. Table 3.1 shows the total number of IP addresses that MaxMind has stored in this database.

Table 3.1: MaxMind IP address coverage

	Blocks MaxMind	Addresses MaxMind	Allocated Addresses	Coverage
IPv4	9,000,024	3,690,208,761	3,707,764,736	99.5%
IPv6	2,172,755	$2.5 \cdot 10^{34}$	$8.3 \cdot 10^{34}$	30.4%

The mentioned number of blocks and addresses of MaxMind are based on private communication. The number of allocated addresses is based on [C24], [C25] and [C26].

MaxMind has a high coverage of the IPv4 address range. The reason that not all 256 /8 blocks of IPv4 are allocated is because IANA has 35 of these blocks still stated as *Reserved*. These are the the private 000/8, 010/8

and 127/8 blocks and the multicast blocks from 224/8 up to 255/8. The coverage of the IPv4 address space is highly plausible. This is because the IPv4 address space has long run out, so every IP address in the allocated space is used. However this is not the case for the IPv6 address space. MaxMind has a somewhat low coverage of the IPv6 address range. The number of IPv6 addresses is based on the number of blocks that have been allocated to different RIRs, but this time it is not possible to know how many of the allocated addresses are actually in use. This means that even though MaxMind only covers 30% of the allocated IPv6 addresses, it may as well be the case that currently less than 100% of the IPv6 addresses are in use, meaning that the actual coverage may be higher.

For every block of addresses, MaxMind also stores a reference to a city, a latitude, a longitude and an accuracy radius in kilometers. Next to these values there is some more information stored in the GeoLite2 City database, but this information is not important for this research. The stored city reference points to another table that contains all sorts of information about the city, for example the continent and the country in which it is situated. The term 'City' could be misleading because there is more information stored in the database than only the actual city. The list of cities contains roughly 123,000 city entries from all over the world. MaxMind does not want to report on how they acquire their data. They state that this is their trademark, so it has to remain confidential.

3.2 RIPE Atlas

As mentioned in Chapter 2, the ground truth is a set of IP address-location pairs from which we know that the location of the IP address is correct. It is favorable that these IP address-location pairs are collected from reliable research systems, such that we know that these pairs are accurate. For this research, we use RIPE Atlas. [C27] *'RIPE Atlas employs a global network of probes that measure Internet connectivity and reachability, providing an unprecedented understanding of the state of the Internet in real time'* is its description. This says that they perform internet measurements to map network structures and data flows in order to better understand the internet. RIPE Atlas uses so called 'probes' to perform these measurements. Probes are small, USB-powered hardware devices that hosts connect to an Ethernet port on their router. This can be done anywhere, for example in a private residence or within a large company. A single probe is only a small piece of the entire RIPE Atlas network, but it extends the amount of data that can be gathered. Hosting a probe is an advantage for RIPE Atlas as it extends their measurement capabilities, but it also gives permission to the host to perform their own measurements. When a probe is installed by a host, the host must report the exact location of its probe in order for RIPE Atlas to

know where this probe is located. Due to privacy reasons, the exact location is only known to RIPE Atlas. The location that is given to us is modified by a few hundred meters in order to preserve the privacy of the host. These probes will be used as a ground truth because both the IP address and the location is reported to RIPE.

3.3 Data Collection

Although RIPE Atlas is a publicly available platform, gathering the necessary data is not an easy task. The information about the probes that they present on their web page is not enough as it lacks the specific coordinates. Without the specific coordinates we can not use the probes as the ground truth, because then the location is inaccurate. Luckily RIPE Atlas also provides a web API from which we can gather the data more efficiently. Within this API all probes are listed that have ever been sent to a host by RIPE. Every probe in the list has a bunch of information attached, but the most important information to us is the IPv4 and IPv6 addresses, the country, the coordinates and the connection status. The API consists of a few hundred pages of probes. Each page can be downloaded in four different formats, namely three *JSON* formats and a text format. We chose to download every page individually in the *GeoJSON* format on February 15, 2021. Using this format, we can easily extract the information that we need by using the python programming language.

In order to efficiently store the extracted information, we set up a local MySQL database using the phpMyAdmin tool. [C28] After creating a table in the database that will hold the probes, we were able to write a python script (see Appendix D.1) that imported the necessary data from all downloaded pages into a single database. This is very useful, because it makes processing the data very easy. Now we have the information from all probes of RIPE Atlas in one table. This gives the advantage that we can easily filter unwanted entries and that we have a good overview of all available probes.

3.4 Data Processing

At this point, we have to filter the valid probes for the ground truth. The raw ground truth database consists of 33,547 entries. This means that when the data was downloaded from RIPE Atlas, they have had 33,547 probes that have been sent to hosts in the past. Unfortunately not all probes that have been sent by RIPE have been connected by the host. This means that in this case no information is stored about those particular probes. Next to that, some probes are private. This means that the host chose to keep information like the IP addresses private. It is not possible to check an IP

address that is not known, so all entries in the ground truth without an IPv4 address have to be deleted. This can be done without the loss of IPv6 addresses, because every probe without an IPv4 address does not have an IPv6 address either. Unfortunately, there are 12,193 of these cases. After removing the probes without an IPv4 address, the ground truth consists of 21,354 entries. This is not the end of the story, next to an IP address, the location of the probe is also important. Without a location, the IP address can not be checked for correctness. Therefore all probes without a location also need to be removed from the ground truth. This is the case for 160 probes. After removing these cases we are left with 21,194 entries.

The last aspect that we need to account for is the status of the probe. There are four types of status, namely *Connected*, *Disconnected*, *Abandoned* and *Never Connected*. The *Connected* status means that the probe is currently still reachable, which is favorable and therefore it is reliable for the ground truth. *Disconnected* means that the probe is currently unreachable, but it has been reachable within the last three months. We assume that this time frame is still acceptable and therefore this category of probes is also considered to be valid for the ground truth. *Never Connected* means that a probe has never been connected by the host and therefore no information about this probe is known by RIPE Atlas. These probes are not in the ground truth because they have already been omitted when removing the probes that did not have an IPv4 address. The last status is *Abandoned* and this means that a probe is unreachable for more than three months. This time frame is very broad and therefore the probes with this status are considered to not be reliable for the ground truth. Probes that have this status may have been disconnected by the host, which releases the dynamic IP address. The original location may not be the same as the current location as it may have been changed by this event. We do not want this to happen, because it would affect the reliability of the accuracy.

This means that we also need to remove the probes from the ground truth that have the *Abandoned* status. Unfortunately, from the 21,194 entries, 10,341 have the *Abandoned* status. After removing these probes, we end up with a ground truth that consists of 10,853 probes. This means that we have 10,853 probes of which we know the IPv4 address and from these probes there are also 5,077 probes that have an IPv6 address next to their IPv4 address. This will be the final dataset of IP address-location pairs with which we will start to perform the comparisons with the GeoLite2 City database.

3.5 Probe Distribution

In order to provide some more information about the distribution of the probes, we can take a look at the global network coverage map of RIPE Atlas [C29] and at the distribution of the probes in our ground truth. In figure 3.1, the distribution of both the *Connected* and *Disconnected* probes are shown according to the map of RIPE Atlas on March 3, 2021. From this figure we can safely state that most of the probes are situated in Europe, where RIPE NCC is active. In second place we can see that the United States also has a lot of active probes.

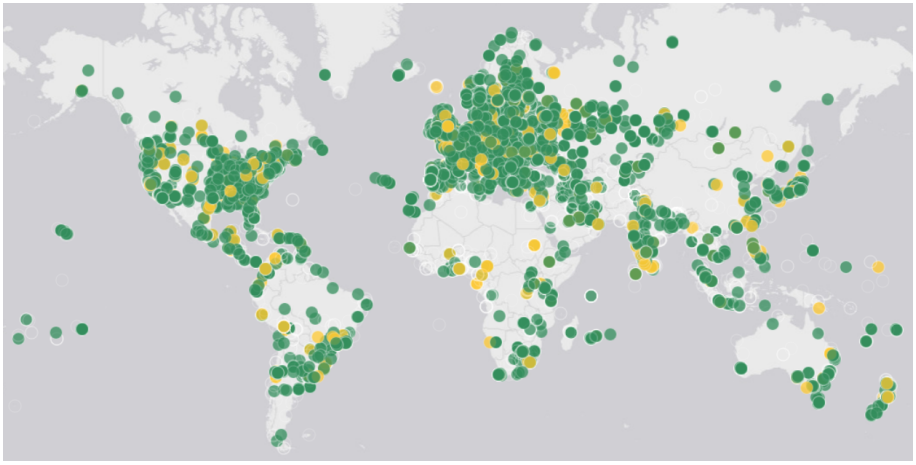

Figure 3.1: Probe distribution. Green is connected, yellow is disconnected.

Now we can take a closer look at the distribution of the probes in our ground truth. In total, our ground truth covers 179 countries from all over the world. However as seen in figure 1, there will most likely be a clustering of probes in Europe and in the United States. The top 3 countries with the most probes in our ground truth are Germany, the United States and France. With 1430, 1383 and 807 probes respectively. Together, these 3 countries make up 33% of all probes in our ground truth.

For this research, we have chosen to only use the top 25 countries that host the most probes in order to assure that no accuracy conclusion is drawn from too little IP addresses for a certain country. We chose to use the top 25 countries because this marks a 100 probe limit. Together, the top 25 countries host 8959 probes from our ground truth. This is 83% of all our probes. Appendix A contains the full list of the top 25 countries, the number of probes they host and the number of IP addresses they cover.

Chapter 4

Research

4.1 Distance Calculations

Now that we are left with a proper ground truth, we have all the information that we need in order to start comparing the ground truth against the GeoLite2 City database. In order to efficiently analyze the results at a later stage, we adapted the MySQL database in such a way that we can store some additional information for both the IPv4 and the IPv6 addresses. This additional information contains the location difference in kilometers, the given accuracy radius by MaxMind and whether or not the location difference falls within the accuracy radius.

We start by comparing the IPv4 address space. We wrote a second script (see Appendix D.2) that takes an IPv4 address from the ground truth and looks it up in the GeoLite2 City database. If no match is found, we skip it and mark it as unresolved. However, if there is a match, we compare the two coordinates of the IPv4 address and calculate the difference in distance between them using the python function *geodesic* from the *geopy* library. We store this difference in the MySQL database alongside the corresponding IPv4 address. Then we look up the accuracy radius for that specific IPv4 address in the GeoLite2 City database and we store this value in our MySQL database. Lastly, we check whether or not the calculated difference in distance is less than or equal to the accuracy radius. If this is the case, then we mark the location as correct, but if it falls outside the accuracy radius, we mark it as incorrect. So after the completion of the IPv4 address space, we have marked every IPv4 address in our ground truth either as unresolved or we have stored the difference in location, the accuracy radius and the correctness of the IPv4 address.

The process for the IPv6 address space went rather similar. The only difference is that not every probe in the ground truth has an IPv6 address next to the IPv4 address. This means that we had to skip the probes without an IPv6 address this time. In the end, every IPv6 address is marked either

as unresolved, or we stored the values for the same three attributes that were also used for the IPv4 address space.

4.2 Data Analysis

In order to answer the research question *Is the accuracy of the GeoLite2 City database correct?*, we need to analyse the results that we obtained in the previous section. We first have to know in which way we need to analyse the results, so we have to take a look at the data that is published by MaxMind. For this, we use the GeoIP2 City Accuracy table, which also contains the accuracy for the GeoLite2 City database. [C18] This table contains the accuracy of their services, listed per country. It is possible to choose from several IP options. Each set of options will return a table that contains the corresponding accuracies. Their original table contains the accuracies of the IPv4 address space and IPv6 address space combined, but they have an option to exclude IPv4 addresses. Another option that can be chosen from is the radius, also called the resolution. The radius can be changed to display the accuracy percentages that correspond with the chosen radius. There are seven possible resolution radii to choose from, namely a resolution of *10 km, 25 km, 50 km, 100 km, 250 km, exact postal* and *exact city*. Lastly, it is also possible to choose whether cellular IPs, broadband IPs or both IPs need to be considered in the accuracy table. Cellular IPs are mobile IP addresses and broadband IPs are the IP addresses assigned by an ISP.

For this research, we only use broadband IPs, so we go with the broadband IPs only option. Since the probes from the ground truth are connected to routers, the cellular IP option is not applicable here. We validate the broadband IP tables for three resolutions and we do not exclude the IPv4 addresses. We use the resolution of 10 km to represent the correctness of a city. MaxMind does not report the radius of the *Exact City* option, so therefore we chose to use the smallest resolution available to represent a city. The second resolution we use is the 50 km radius. Lastly we use the 250 km resolution to represent a country, because this is the largest radius available. We consulted these accuracy tables on February 26, 2021. The accuracies of these tables may change in the future.

4.2.1 Point Accuracy

We start by calculating the point accuracies for the 10, 50 and 250 kilometer radii. As described in section 3.5, we only used the top 25 countries that host the most probes in order to not use too little IP addresses. This means that we only calculate the accuracies for these 25 countries. For every country c, we follow the steps listed below to calculate the percentage of correct IP addresses given a radius ρ, where ρ can be replaced by one of the three radii mentioned above.

1. Calculate the total number of IPv4 addresses of c (ν_c).
2. Calculate the total number of IPv6 addresses of c (μ_c).
3. Take the sum of step 1 and step 2 to calculate the total number of IP addresses of c.
4. Calculate the number of IPv4 addresses of c that have a distance difference of at max ρ kilometer (δ_c).
5. Calculate the number of IPv6 addresses of c that have a distance difference of at max ρ kilometer and the difference is not 0, because then the IPv6 address was marked as unresolved (η_c).
6. Take the sum of step 4 and step 5 to calculate the total number of IP addresses of c within a radius of ρ kilometer.
7. The fraction of the answers to step 6 and step 3 yield the percentage of correct IP addresses of c within a radius of ρ kilometer.

More specifically, we use formula 4.1 for every country c to calculate the accuracy of that country given a radius ρ.

$$acc_c(\rho) = \frac{\delta_c + \eta_c}{\nu_c + \mu_c} \tag{4.1}$$

The steps listed above represent the general approach on calculating the accuracies. For our research, we have to query the ground truth to calculate the accuracies. The queries that we used in these steps are listed in Appendix C. For every country, we have now calculated the percentages of correct IP addresses for all three radii. These percentages represent the accuracy of broadband IPs within the corresponding radius. The fact that step 5 needs to account for not using the unresolved IPv6 addresses is to make sure that the set of IP addresses that we use is a subset of the database of MaxMind.

4.2.2 Confidence Interval Accuracy

Now that we calculated the percentages of correct IP addresses for the three different radii, we need to check the reliability of our own results. This means that we have to perform some statistics to calculate the confidence intervals of the results that we obtained. A confidence interval is an interval that marks the boundaries of the result that will be estimated by repeating the experiment, with a certain level of confidence.

The result of the experiment is based on the fact that an IP address can either be within the given radius or it is outside the given radius. This means that the IP address is either marked as a success or as a failure. The experiment can only have these two outcomes, so it represents a Bernoulli

trial. A Bernoulli trial is an experiment with the outcome of either success or failure. Now that we know that our experiment represents a Bernoulli trial, we are able to use a binomial proportion confidence interval.

A binomial proportion confidence interval is an interval that can be estimated when only the sample size and the number of successes are known. Two examples of a binomial proportion confidence interval are the *normal approximation interval* and the *Wilson score interval*. Wilson's score interval is an improvement over the normal approximation interval, because it can be applied to small samples and it corrects for a sample with a bias. [A9] This makes Wilson's score interval more suitable over the normal approximation interval for our experiment. It is particularly useful, because even when a country has little IP addresses in our ground truth, the Wilson score interval will still estimate a good approximation of the confidence interval of our accuracies. Formula 4.2 calculates Wilson's score interval.

$$(w^-, w^+) \equiv \left(p + \frac{z^2}{2n} \pm z\sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}} \right) \Big/ \left(1 + \frac{z^n}{n} \right) \tag{4.2}$$

In this formula, p represents the calculated accuracy, n is the sample size and z is a fixed number that corresponds with the probability that an element falls within the confidence interval. We calculate the accuracies for both a 95% and a 99% confidence interval, which has a z-value of 1.96 and 2.576 respectively.

In order to not calculate all intervals by hand, we wrote a third script (see Appendix D.3) that uses the Wilson score interval formula and it calculates both the 95% and 99% confidence intervals for every country. These ranges indicate that when we redo the experiment, the results will fall within this range with 95% or 99% confidence. When we increase the level of confidence, the confidence interval becomes larger. This is because by increasing the level of confidence, we need to make sure that more edge cases are identified as correct. The disadvantage of using a higher confidence interval is that the interval becomes larger, but the advantage is that the level of confidence is higher. This means that when the value is still found outside the interval, then it is likely that this value is incorrect.

4.2.3 Comparing the Accuracy

Now that we have calculated the confidence intervals for every country for all three radii, we can start comparing the intervals with the accuracies reported by MaxMind. We check whether or not the accuracies reported by MaxMind fall within the confidence intervals that we calculated before. If the accuracy of MaxMind indeed falls within the confidence interval, then we mark it as correct. If it does not fall within the confidence interval, then the accuracy of MaxMind is either higher or lower than the confidence interval. Both of these cases are marked as incorrect.

Chapter 5

Results

This chapter shows the results that we obtained during the research. The next three sections cover the results for the three different radii that we used in the experiment. Each section presents the results of the corresponding radius in a figure.

The accuracy of Iran is not displayed in the GeoIP2 City Accuracy table of MaxMind. This means that it is not possible to compare the two accuracies of this country. Therefore, in the following sections, Iran is omitted and the remaining 24 countries are discussed.

5.1 10 Kilometer Radius

Figure 5.1 shows the results of the experiment on a 10 kilometer radius, which we used to represent the radius of a city. Within this figure, the green bars present the number of IP addresses that we used for every country. A clear overview of the number of IP addresses can be found in Appendix A. The blue diamonds present the accuracies reported by MaxMind and the black intervals present the 99% confidence intervals of the results of the experiment, where the black dots mark the point accuracies. Appendix B.1 gives an overview of the exact numbers that we obtained for the 10 kilometer radius.

In figure 5.1, the countries are sorted into two groups. The left group presents the countries in which MaxMind reports a higher accuracy than the point accuracy calculated in the experiment. The right group presents the countries in which MaxMind reports a lower accuracy than the point accuracy calculated in the experiment. Both the left and the right group are ordered by the accuracy of MaxMind in decreasing order. This is done to make the figure more readable.

In figure 5.1, the left group contains 15 countries and the right group contains 9 countries. This means that in 15 cases, the accuracy of MaxMind is reported to be higher than the point accuracy that we obtained. In 9 cases the accuracy of MaxMind is lower than the point accuracy that we obtained. If we take the confidence intervals to be a range of valid accuracies, then MaxMind reports a correct accuracy for 10 of the 24 countries.

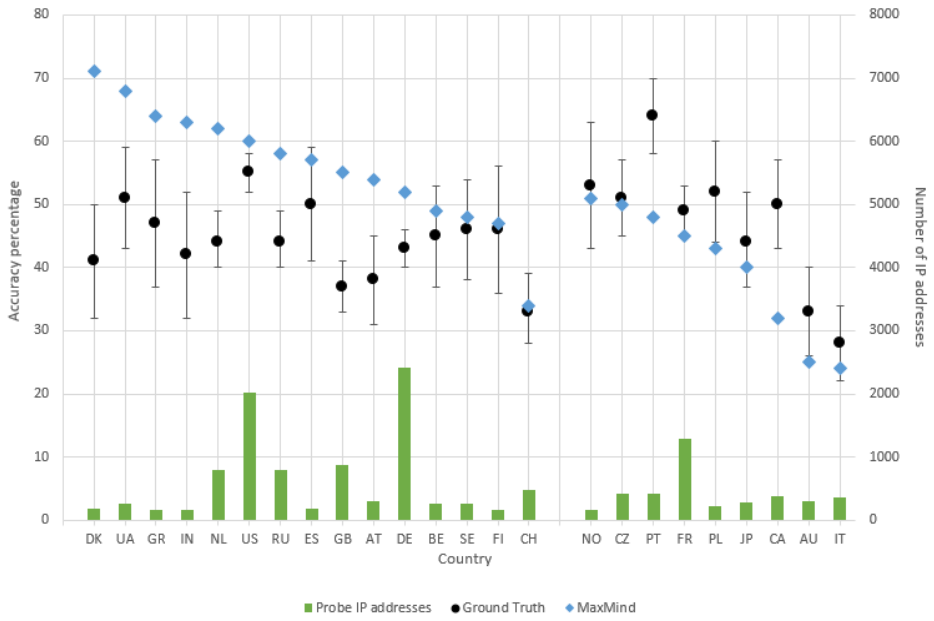

Figure 5.1: 99% CI - 10 KM Accuracy

The largest deviation between the accuracy reported by MaxMind and the point accuracy that we calculated is for the country Denmark, which is the first country in this figure. This accuracy difference is 30%. Although there are some large deviations, there are also some minor deviations. We can calculate the average deviation for a country by using formula 5.1.

$$\sum_{x \in Countries} \Big(\Big| \text{accuracy\_MM}(x) - \text{accuracy\_GT}(x) \Big| \Big) \; / \; \Big| Countries \Big| \quad (5.1)$$

Formula 5.1 basically computes the sum of the deviations from all 24 countries, taking into account to only use the absolute values and then it takes the average to get the average deviation for a country. In this formula, the ground truth accuracy is defined to be the point accuracy that we obtained. Using formula 5.1, the average accuracy deviation for a country is 10.3%. We can also calculate the average deviation between the two entire datasets by using formula 5.2.

$$\left(\sum_{x \in Countries} \text{accuracy\_MM}(x) - \sum_{x \in Countries} \text{accuracy\_GT}(x) \right) \Big/ \left| Countries \right| \quad (5.2)$$

The average accuracy reported by MaxMind for these 24 countries is 50.0% and the average accuracy that we calculated is 45.8%. By using formula 5.2, the average accuracy deviation, for the 10 kilometer radius, between the GeoLite2 City database and the ground truth is 4.2%.

5.2 50 Kilometer Radius

Figure 5.2 shows the results of the experiment on a 50 kilometer radius, which we used because it is the standard radius that MaxMind uses. This figure uses the same representation as described in section 5.1. An overview of the exact numbers that we obtained for the 50 kilometer radius can be found in Appendix B.2.

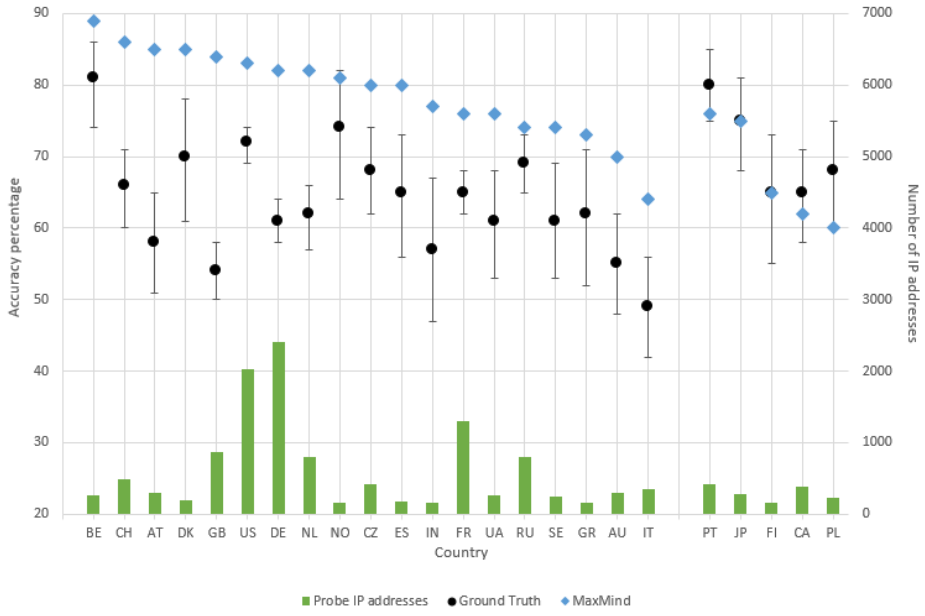

Figure 5.2: 99% CI - 50 KM Accuracy

In figure 5.2, the left group contains more countries than the left group of figure 5.1. This time the left group contains 19 countries and the right group contains 5 countries. This also means that in 19 cases, the accuracy of MaxMind is reported to be higher than the point accuracy that we obtained. In 5 cases the accuracy of MaxMind is lower than the point accuracy that we obtained. If we again take the confidence intervals to be a range of

valid accuracies, then MaxMind reports a correct accuracy for 6 of the 24 countries. Which is 4 countries less than for the 10 kilometer radius.

For this radius, the largest deviation between the accuracy reported by MaxMind and the point accuracy that we calculated is for the country Great Britain. This accuracy difference is again 30%. We can calculate the average deviation for a country by using formula 5.1. This results in an average accuracy deviation of 12.8%. This is a little higher than for the 10 kilometer radius.

The average accuracy reported by MaxMind for these 24 countries is 76.6% and the average accuracy that we calculated is 64.9%. By using formula 5.2, the average accuracy deviation, for the 50 kilometer radius, between the GeoLite2 City database and the ground truth is 11.7%. This deviation is much higher than for the 10 kilometer radius.

5.3 250 Kilometer Radius

Figure 5.3 shows the results of the experiment on a 250 kilometer radius. This is the last radius we used and it was used to represent the radius of a country. This figure also uses the same representation as described in section 5.1. Appendix B.3 gives an overview of the exact numbers that we obtained for the 250 kilometer radius.

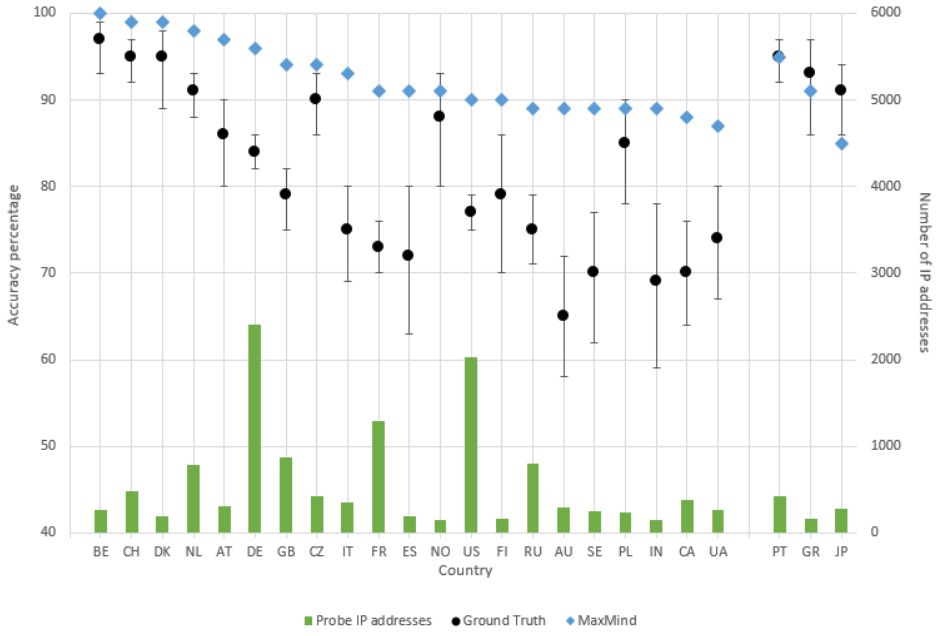

Figure 5.3: 99% CI - 250 KM Accuracy

The left group of figure 5.3 is the largest of the three radii. The left group contains 21 countries and the right group only contains 3 countries. This means that in 21 cases, the accuracy of MaxMind is reported to be higher than the point accuracy that we obtained. In only 3 cases the accuracy of MaxMind is lower than the point accuracy that we obtained. If we take the confidence intervals to be a range of valid accuracies, then MaxMind reports a correct accuracy for only 4 of the 24 countries. This is the lowest number of the three radii.

This time, the largest difference between the accuracy reported by MaxMind and the point accuracy that we calculated is for the country Australia. This accuracy difference is 24% and it is less than the difference for the previous two radii. We can again calculate the average deviation for a country by using formula 5.1. This results in an average accuracy deviation of 10.9%.

The average accuracy reported by MaxMind for these 24 countries is 92.3% and the average accuracy that we calculated is 81.2%. Again, by using formula 5.2, the average accuracy deviation, for the 250 kilometer radius, between the GeoLite2 City database and the ground truth is 11.1%. This deviation is higher than for the 10 kilometer radius, but about the same as for the 50 kilometer radius.

5.4 Result Analysis

When we look at figure 5.1, the values of the ground truth and the values of MaxMind seem to converge to the middle, whereas in figures 5.2 and 5.3, the values of the ground truth mostly stay below the values of MaxMind. This indicates that the larger the radius, the more countries fall in the left group because their values become less than those of MaxMind. Since most of the countries fall in the left group, the average accuracies of all three radii are below the average accuracies of MaxMind. The summary of these averages can be found in table 5.1.

Table 5.1: Summary of the accuracy results

	Correct	Average MaxMind	Average Results
10 km	10 / 24	50.0%	45.8%
50 km	6 / 24	76.6%	64.9%
250 km	4 / 24	92.3%	81.2%

Figures 5.1, 5.2 and 5.3 also contain the number of IP addresses that we used for every country. Although the scale is different in each figure, the number of IP addresses used for each radius is the same. The scale only differs to line up with the other results. We included the number of used

IP addresses to see if there is a correlation between the number of used IP addresses and the obtained accuracy.

In order to check if there is indeed a correlation, we can make use of the Pearson correlation coefficient. This coefficient ($-1 < \rho < 1$) expresses how much two datasets correlate with each other. The closer ρ is to 0, the less the two datasets correlate, the closer ρ is to either 1 or -1, the more the two datasets correlate.

For each of the three radii, we take a look at the correlation between the number of IP addresses that we used for each country (ν) and the accuracy that we obtained for each country (α). We also take a look at the correlation between the number of IP addresses that we used for each country (ν) and the deviation from the accuracy of MaxMind for each country. The deviation for a country is calculated by subtracting the accuracy reported by MaxMind (μ) from the accuracy obtained in this research (α). Define Γ to be the set of all 24 countries. We look at the following two correlations:

(1) $\rho_1 = \text{corr}\Big(\{\nu_c \mid c \in \Gamma\}, \{\alpha_c \mid c \in \Gamma\}\Big)$

(2) $\rho_2 = \text{corr}\Big(\{\nu_c \mid c \in \Gamma\}, \{(\alpha_c - \mu_c) \mid c \in \Gamma\}\Big)$

Table 5.2 presents the calculated correlation coefficients.

Table 5.2: Correlation coefficients

	ρ_1	ρ_2
10 km	0.08	-0.05
50 km	-0.02	-0.24
250 km	-0.08	-0.17

All six correlation coefficients are much closer to 0 than to either 1 or -1. This indicates that there is minor correlation between the number of IP addresses that we used and the accuracy that we obtained. There is also minor correlation between the number of IP addresses that we used and the deviation from the accuracy of MaxMind. This means that the number of IP addresses that we used in the experiment does not influence the results.

Chapter 6

Related Work

Accuracies of IP geolocation databases have been validated in the past. However, there are few papers that go into detail on the accuracy of a single database. Prior research only focused on calculating the average accuracies of multiple databases. Only the related work on this topic after 2014 is considered, because before 2014, the GeoLite2 database did not exist yet.

In 2016, Kester compared the accuracy of IPv4 and IPv6 geolocation databases. [A10] The aim of this research is to map the difference between the accuracy of IPv4 and IPv6 in geolocation databases. However, we can still use the results from this research to get a better understanding of the accuracy of these databases in the past.

In this paper, Kester uses three different databases to calculate the difference between the accuracy of IPv4 and IPv6 addresses. The databases are: DB-IP, IP2Location DB5-Lite and MaxMind GeoLite2. In order to perform the calculations, Kester also built a ground truth. This ground truth is assembled by collecting IP address-location pairs from the research institutes CAIDA and RIPE Atlas. Since Kester wants to compare the statistics of the IPv4 and IPv6 address sets, he only uses nodes in the ground truth that both have an IPv4 and an IPv6 address. There are 3206 nodes in his ground truth that comply with the restriction. Kester then uses the ground truth to calculate the average deviations between the locations in the ground truth and the locations provided by the databases. Kester also presents the accuracies of the databases given a certain radius. Both calculations are done for the IPv4 and IPv6 address-sets separately.

Note that the ground truth from Kester slightly differs from our ground truth. Kester used two research institutes and he is only able to use nodes with both an IPv4 and IPv6 address. Kester has less usable IP addresses in the ground truth and the IP addresses can differ from the ones in our ground truth.

Kester calculated that for the GeoLite2 database, the average distance difference between the IPv4 address set and the ground truth is 264.4 kilometer. The average distance difference between the IPv6 address set and the ground truth is 1098.7 kilometer.

In our research, the distance difference between the IPv4 address set and the ground truth is 176.7 kilometer and the average distance difference between the IPv6 address set and the ground truth is 637.8 kilometer. This means that the average results that we obtained are better than the average results that Kester obtained in his research. Since the average distance difference is lower for both the IPv4 and IPv6 address sets, the overall accuracy of the GeoLite2 database may have increased over time.

Kester also calculated the GeoLite2 database accuracies for eight different radii. These accuracies are calculated for the entire database. Kester has calculated the accuracies for both the IPv4 and IPv6 address sets separately in order to compare the two sets. This means that in order to compare our results to those from Kester, we also need to split the accuracies that we obtained into an IPv4 and an IPv6 accuracy. Table 6.1 contains the comparison of the accuracies of the IPv4 address set and table 6.2 contains the comparison of the accuracies of the IPv6 address set. In both tables, the 2016 column represents the accuracies found by Kester in 2016 and the 2021 column represents the accuracies that we obtained during this research. These separate accuracies have specifically been calculated in order to compare them to the accuracies that Kester obtained.

Table 6.1: Comparison of the IPv4 accuracy of the GeoLite2 database

	GeoLite2 IPv4 2016	GeoLite2 IPv4 2021
10 km	38%	51%
50 km	55%	73%
250 km	77%	88%

Table 6.2: Comparison of the IPv6 accuracy of the GeoLite2 database

	GeoLite2 IPv6 2016	GeoLite2 IPv6 2021
10 km	15%	31%
50 km	23%	46%
250 km	49%	67%

Both table 6.1 and table 6.2 show that all the accuracies that were calculated by Kester in 2016 are lower than the accuracies that we obtained during this research. Although Kester does not discuss how well the accuracies correspond to those of MaxMind, he does state that the obtained

accuracies are on the low end. This is also the case with the accuracies that we obtained during this research.

Kester states that the DB-IP database performs the worst in locating IPv4 addresses. IP2Location and MaxMind perform almost the same in locating IPv4 addresses. In locating IPv6 addresses, MaxMind outperforms the other two databases. According to Kester, this makes the database of MaxMind the best one out of these three databases.

In 2017, Gharaibeh et al. took a look at the accuracy of router based geolocation in popular databases. [A11] Their research includes the following databases: MaxMind GeoIP2, MaxMind GeoLite2, IP2Location DB11-Lite and Digital Element NetAcuity. This research built a ground truth using two methods, namely to geolocate routers by decoding location hints in their hostnames and by using the Round Trip Time to locate the routers. They then cross reference the ground truth with all four databases. They analyze the router geolocation coverage and accuracy at country- and city-level.

According to their research, their country-level geolocation accuracy results over the ground truth data show less accuracy than all database providers report. They state that all database providers report an accuracy higher than 97%, whilst they calculated an accuracy of 89.4% for NetAcuity, 77.5% for IP2Location and 78.6% for the MaxMind databases. Since they do not mention the radius that they use to represent a country, it is not possible to compare their results to the results that we obtained. Gharaibeh et al. conclude their research by stating that the databases are not accurate enough in geolocating routers at country- and city-level.

Lastly, in 2020, Xu et al. performed an experimental comparison of free IP geolocation databases. [A12] This experiment also includes calculating the accuracy for the MaxMind GeoLite2 database. However, they only use Chinese IP addresses. This means that they only calculate the accuracy of China within the GeoLite2 database. The scope of their research is too narrow to compare it to the results from our research.

Chapter 7

Discussion

In this chapter, we discuss some assumptions that we made during the research. These assumptions led to the results as they are presented in chapter 5. We also look at the impact of these assumptions. Next, we take a look at the reliability and the validity of the research. Then we discuss the results and try to think of ways that might explain the results as they are. Finally, we discuss the accuracy tables of MaxMind.

7.1 Assumptions

In chapter 3.4 we made the assumption that a *Disconnected* probe was good enough to use in the ground truth. Therefore, the ground truth consisted of probes with either a *Connected* status or a *Disconnected* status. We chose to use both status in order to increase the number of IP addresses. The question is, could the *Disconnected* probes negatively impact the results? To answer this question, we run the same steps described in 4.2.1 to calculate the accuracies for the three different radii, this time only using the IP addresses from the *Connected* probes. The new results are almost the same. The new accuracies drop at most 3% in accuracy. This means that the average results also decrease, so they do not get closer to the average of MaxMind. This indicates that including the *Disconnected* probes in the ground truth does not negatively impact the results.

In chapter 4.2 we made the assumption that all IP addresses in our ground truth are broadband IP addresses. Therefore we also used the accuracy tables given by MaxMind that only accounted for the broadband IP addresses. However, if we choose to use the accuracy tables given by MaxMind that account for both broadband- and cellular IP addresses, their accuracies drop. These accuracies are closer to those that we obtained in this research.

In chapter 4.2.2 we used the Wilson score interval to calculate the confidence intervals of our results. One condition of using the Wilson score in-

terval on our results is that the IP addresses are equally distributed within every country. We made the assumption that the IP addresses from the ground truth are indeed equally distributed within every country.

Lastly, as mentioned in chapter 3.1, MaxMind does not report in which way they gather their IP address-location pairs. In the same way they do also not provide how they calculate their own accuracies. This means that the process described in chapter 4 might not be the same process that MaxMind uses to calculate their accuracies.

As stated in chapter 3.1, the GeoLite2 City database contains an accuracy radius alongside the IP address. One possibility is that MaxMind uses the accuracy radius as a margin of error. This is the radius in which, according to MaxMind, the real location of the IP address must reside.

In our research, we did not use the accuracy radius, because we calculated the distances ourselves. If the calculated distance was lower than the chosen radius in the accuracy table given by MaxMind, then we simply marked it to be correct. The question is, does MaxMind use the accuracy radius to calculate their accuracies?

For example, we choose the radius of the accuracy table given by MaxMind to be 50 kilometer. We calculated a distance difference between the location given by MaxMind and the location from the ground truth of 40 kilometer, so we marked the location correct for this radius. Now let the accuracy radius be 100 kilometer. Then we still mark the location as correct, but MaxMind might mark the location as incorrect, because the accuracy radius is larger than the chosen radius in the accuracy table. This problem can also occur the other way around. If we calculated a distance difference of 60 kilometer, then we marked the location incorrect for the 50 kilometer radius. But if the accuracy radius was set to be 10 kilometers, then MaxMind might mark the location as correct because the accuracy radius is lower than the chosen radius in the accuracy table.

In order to answer the question if MaxMind uses the accuracy radius to calculate their accuracies, we calculated the new accuracies by only looking at the accuracy radii given by MaxMind. If this is their approach in calculating their accuracies, then the new results must lie closer to those from MaxMind. Using this method, the accuracies are still lower than the accuracies reported by MaxMind and in most cases the new accuracies are also even lower than the accuracies reported in chapter 5. This means that if we add more restrictions on top, the accuracies will only decrease further. Therefore, it is most likely that MaxMind does not use the accuracy radius to calculate their accuracies.

7.2 Reliability and Validity

First we discuss the reliability of this research. The experiment performed in this research is time sensitive. We have downloaded the GeoLite2 City database and the ground truth locally, which at that point was a fixed environment on which we performed the calculations. The dates on which we have downloaded the GeoLite2 City database and the ground truth are mentioned in chapters 3.1 and 3.3. If the research is repeated using the same environment, then the results should be the same. However, if the research is repeated using the most up to date ground truth and GeoLite2 City database, then the results will most likely be different. In this case, it is hard to say if the difference is subtle or large, but they will not be exactly the same.

The validity of the research is more complex. There are three actions during the research that we have to look into. The first action is calculating the distance differences, the second action is calculating the number of correct IP addresses within a certain radius and the third action is calculating the confidence intervals.

We calculated the distance differences by using an in-built python function. This function is proven to return the correct distance, so this action should work as intended. Calculating the accuracies given a certain radius is done by checking how many IP addresses comply with the given restrictions. This action should also work as intended, but as stated in chapter 7.1, the accuracies can also be calculated differently if another method is used. This will result in different results. Lastly, as stated in chapter 7.1, one condition of calculating the confidence intervals is that the IP addresses are equally distributed within every country. We assumed this to be the case, but on closer inspection, there are minor clusters of IP addresses in larger cities. Although there are plenty of IP addresses in the rest of the countries, it might not be totally equally distributed. This could cause the confidence intervals to be less accurate.

7.3 Result Discussion

In chapter 5, the results of the experiments are shown in three figures. In chapter 5 we discussed the results numerically and we checked if there is a possible correlation between the number of IP addresses that we used and the accuracy that we obtained. Now we try to think of other ways that might explain the results as they are.

In each of the following three sections, we first discuss the accuracies of MaxMind, then we discuss the point accuracies obtained in this research and finally we compare the two.

7.3.1 Area of the Country

First we take a look at the area of the countries. More specifically, does the area of the country affect the accuracy? Could it be the case that larger countries have low accuracies and that smaller countries have high accuracies? In order to answer this question, we order the countries by their area in decreasing order and then we take a look at the accuracies for the three radii. Figure 7.1 contains the ordering of the countries and uses the accuracies of MaxMind. Figure 7.2 contains the ordering of the countries and uses the accuracies obtained in this research.

The linear lines through the points are called trendlines. These trendlines give an indication if there is a gradual increase or decrease along the points. In figure 7.1, we see that for all three radii, there is an increase in accuracy. The trendline for the 10 kilometer radius increases slowly and the trendlines for the 50 and 250 kilometer radii increase more steeply. This indicates that, for the accuracies reported by MaxMind, there is a correlation between the area of the country and the accuracy. Namely the smaller the country, the higher the accuracy.

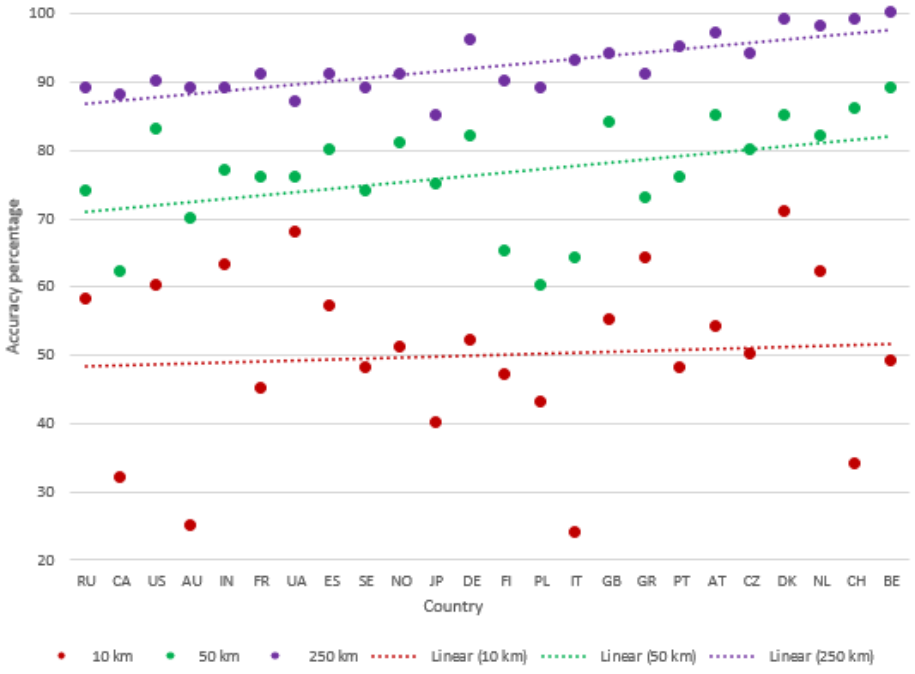

Figure 7.1: Correlation between the area of the country and the accuracy of MaxMind

In figure 7.2 however, we see that for the 10 kilometer radius, there is a very small decrease in accuracy. That would indicate that the smaller the country, the lower the accuracy. The trendline for the 50 kilometer radius has a very small increase in accuracy and the trendline for the 250 kilometer radius increases steeply. This is a good indication that for the 250 kilometer radius, there is indeed a correlation between the area of the country and the accuracy. Namely, the smaller the country, the higher the accuracy.

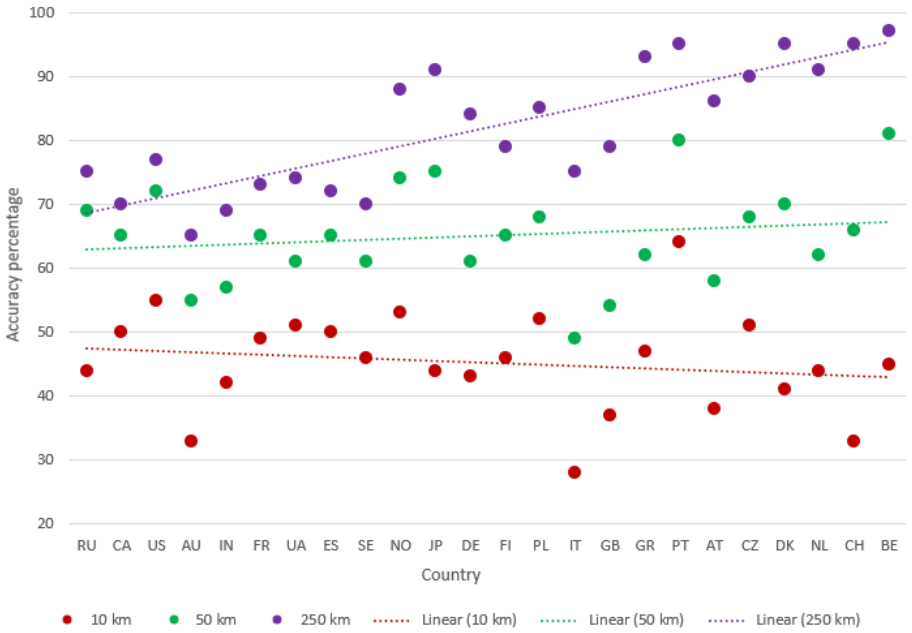

Figure 7.2: Correlation between the area of the country and the accuracy obtained in this research

When we compare the two figures, we see that for most of the radii the trendlines are increasing. This is a good indication that it indeed holds that the smaller the country, the higher the accuracy. In our results we saw that the trendline for the 250 kilometer radius increases steeply. This is probably due to the fact that when the radius is large, then most of the area of the smaller countries fall within the radius. Due to this occurrence, it does not matter where in the country the IP address is located, it will always be marked as correct. The difference between the trendlines in figures 7.1 and 7.2 for the 10 kilometer radius is small, this difference might be explainable by the fact that the confidence intervals are not included in these calculations. These may slightly shift the trendlines. On top of that, in figure 7.1, the dispersion of the 10 kilometer accuracies is large. This means that a small deviation in the accuracies will cause the trendline to shift.

7.3.2 Location of the Country

Next to the area of the country, we can also look if the geographical location of the countries affects the accuracy. More specifically, we can divide the 24 countries in groups, where every group represents a unique RIR. In chapter 3.5 we have seen that most of the IP addresses are situated in Europe, where the RIPE NCC RIR is active. This also means that there are not a lot of countries that can be divided among the other RIRs. We have the United States and Canada that belong to the ARIN RIR. Then we have Australia, Japan and India that belong to the APNIC RIR. Finally, the leftover 19 countries fall under the RIPE NCC RIR. Figure 7.3 and figure 7.4 show the average country accuracies for each RIR for all three radii. Again, figure 7.3 contains the average accuracies reported by MaxMind and figure 7.4 contains the average accuracies obtained in this research. For the next observations we have to keep in mind that there are far more countries in the RIPE NCC group than in the ARIN and APNIC groups.

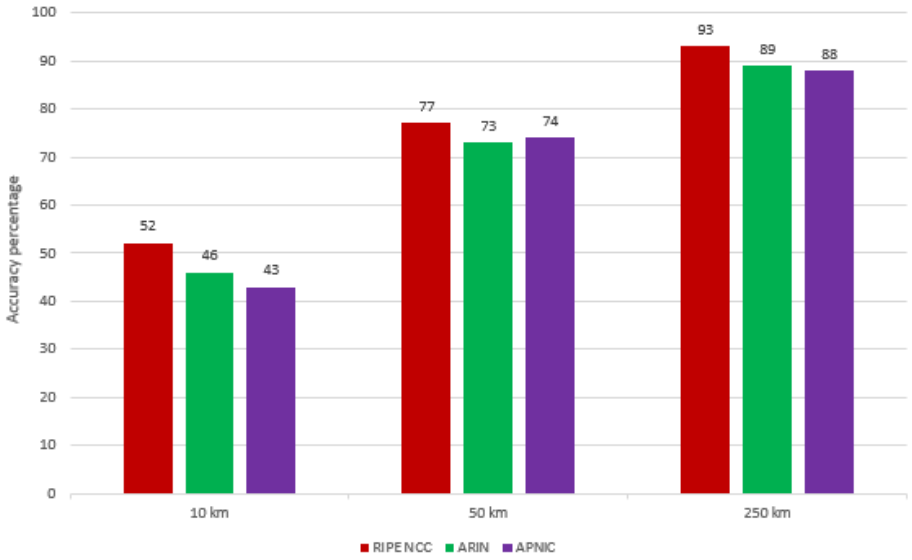

Figure 7.3: Correlation between the location of the country and the accuracy of MaxMind

In figure 7.3 we see that the ARIN and APNIC RIRs have a comparable accuracy for all three radii. The RIPE NCC RIR has the highest accuracy for all three radii. These results are in line with the observations in chapter 7.3.1, figure 7.1. The ARIN RIR consists of the countries in second and third place of countries with the largest area. The APNIC RIR consists of the countries in fourth, fifth and eleventh place of countries with the largest area. In chapter 7.3.1 we stated that larger countries have lower accuracies, therefore these two RIRs have a lower accuracy. In turn, the RIPE NCC RIR

contains a lot of smaller counties. In figure 7.1 all trendlines are increasing, which indicated that smaller countries have higher accuracies, therefore this RIR has a higher accuracy.

In figure 7.4, we do not see major correlations. We do see that the APNIC RIR has the lowest accuracy for the 10 and 50 kilometer radii and that the ARIN RIR has the highest accuracy for these two radii. For the 250 kilometer radius, the RIPE NCC RIR has the highest accuracy. These results are in line with the observations in chapter 7.3.1, figure 7.2. The same principle applies as described for figure 7.3. The only major difference is that in figure 7.2, the 10 kilometer radius trendline is decreasing. This results in a lower 10 kilometer average accuracy for the RIPE NCC RIR in figure 7.4.

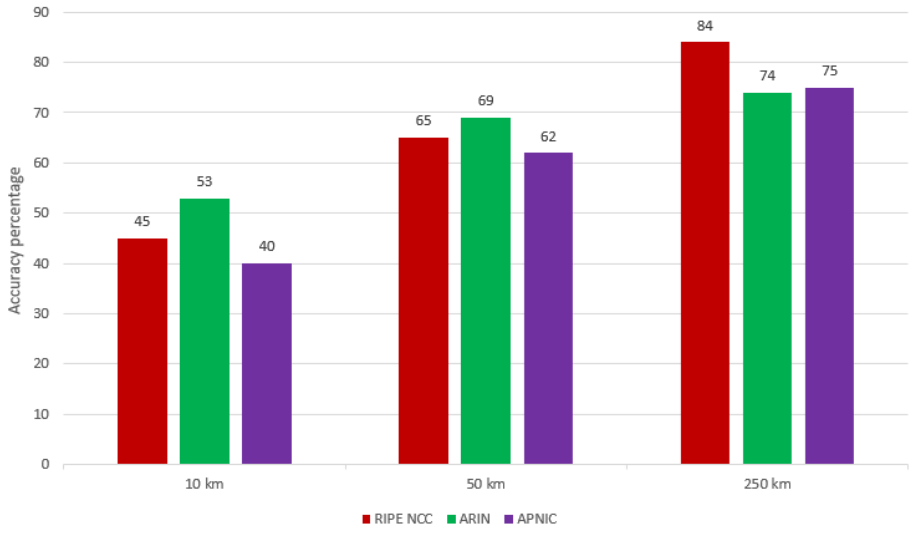

Figure 7.4: Correlation between the location of the country and the accuracy obtained in this research

When we compare the two figures, we see that for the 10 kilometer radius in figure 7.3, the RIPE NCC RIR has the highest average accuracy of the three RIRs, whereas in figure 7.4 the ARIN RIR has the highest average accuracy. In both figures 7.3 and 7.4 we see that for the 250 kilometer radius, the RIPE NCC RIR has the highest accuracy. Again, this means that for large radii, the European countries have a higher accuracy than non European countries due to the fact that most European countries are small.

7.3.3 IP Addresses per Country

Lastly, we look into the number of IPv4 addresses that every country owns. More specifically, does the number of IPv4 addresses that a country owns affect the accuracy? We chose to only use the IPv4 addresses, because our research uses more IPv4 addresses than IPv6 addresses and the IPv4 address allocation is nicely documented. [C30] Based on these numbers we order the countries by the number of IP addresses that they own in decreasing order. Then we take a look at the accuracies for the three radii. Figure 7.5 contains the ordering of the countries and uses the accuracies of MaxMind and figure 7.6 contains the ordering of the countries and uses the accuracies obtained in this research.

In figure 7.5, for all three radii, the trendlines show a small increase along the points. For the 10 kilometer radius, the trendline increases more steeply. This may be an indication that the accuracy increases when a country owns less IP addresses.

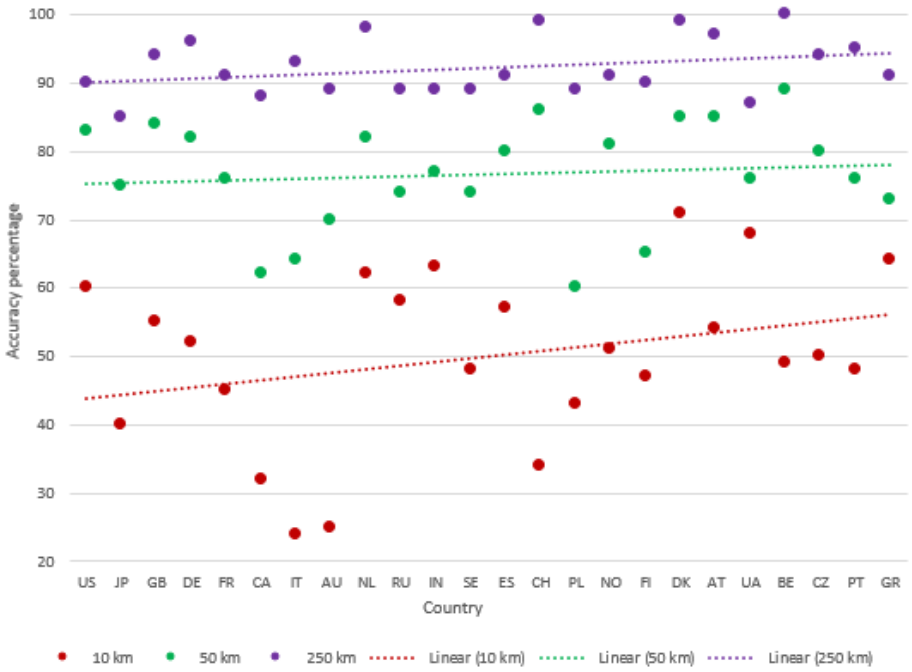

Figure 7.5: Correlation between the number of IP addresses per country and the accuracy of MaxMind

In figure 7.6, we see the same occurrence as in figure 7.5. For all three radii, the trendlines show a small increase along the points. This time the trendline increases more steeply for the 250 kilometer radius.

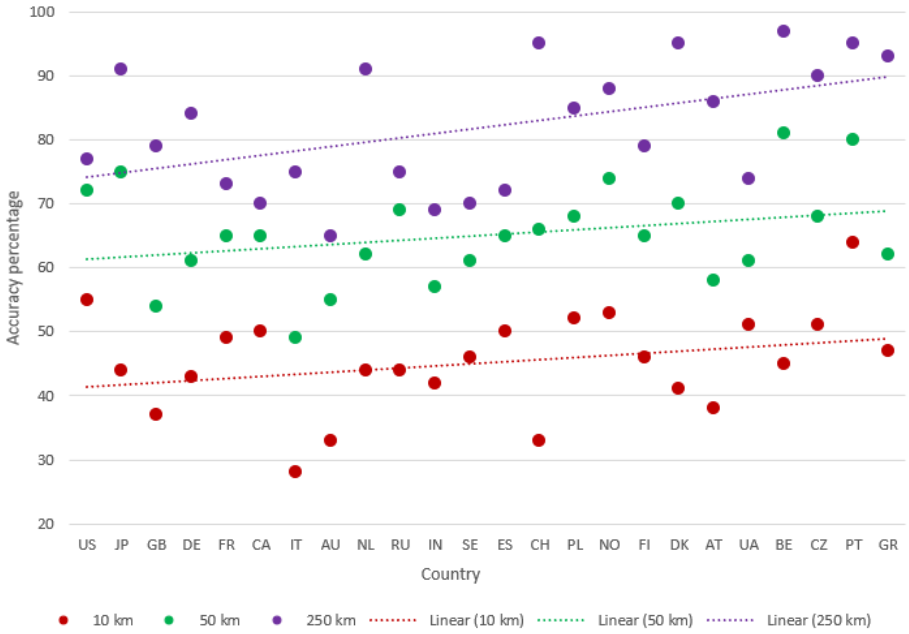

Figure 7.6: Correlation between the number of IP addresses per country and the accuracy obtained in this research

When we compare the two figures, we see that in both figures, for all three radii, the trendlines are increasing. This is a good indication that there is indeed a correlation between the number of IP addresses that a country owns and the accuracy. Namely, the less IP addresses a country owns, the higher the accuracy. It is difficult to find an explanation why this is the case. A possible explanation could be that there are indeed IP address clusters as described in chapter 7.2. Countries with little IP addresses may have clusters in large cities only. Countries with a lot of IP addresses may have IP addresses everywhere in the country on top of the clusters in large cities. It could be the case that the IP addresses in the clusters are all marked correct if they all fall under the same IP address block. This causes a high accuracy for countries with only IP addresses in clusters.

7.4 Accuracy Table

Lastly, we discuss an interesting phenomena in the accuracy tables given by MaxMind. As stated in chapter 4.2, MaxMind offers an option to exclude the IPv4 addresses, resulting in the accuracy tables for the IPv6 addresses. When we do so, most of the accuracies increase. This means that in most of the cases, the IPv4 addresses lower the overall accuracy. Concluding that in most cases the overall accuracy of the IPv4 addresses is lower than the overall accuracy of the IPv6 addresses. However, if we split the accuracies

that we obtained in this research in an IPv4 and an IPv6 component, then in most cases the IPv4 accuracy is higher than the IPv6 accuracy. This is in contradiction with what we see in the accuracy tables given by MaxMind.

A possible explanation to this problem could be the number of IPv4 and IPv6 addresses. As mentioned in chapter 3.1, the number of IPv6 addresses in the GeoLite2 City database is far more than the number of IPv4 addresses. This could mean that the IPv6 accuracy outweighs the IPv4 accuracy. In this research, we have used more IPv4 addresses than IPv6 addresses, which may have shifted the accuracy towards the IPv4 addresses.

Chapter 8

Conclusions

In this research, we have validated the accuracy of the MaxMind GeoLite2 City database for 24 countries. We did so by calculating our own accuracies using a self assembled set of IP addresses collected from the research institute RIPE Atlas and comparing them against the accuracies reported by MaxMind. We compared the accuracies for three different radii.

In the process of answering the research question *Is the accuracy of the MaxMind GeoLite2 City database correct?* we found that, for all three radii, the average accuracies that we obtained were below the average accuracies reported by MaxMind. Among the three radii, the average accuracies of the smallest radius came closest to the average accuracies of MaxMind. The accuracy difference for the smallest radius, the 10 kilometer radius, was 4.2%. The accuracy differences for the 50 kilometer and 250 kilometer radii were larger, namely 11.7% and 11.1% respectively.

This indicates that, according to our research, the accuracy of the GeoLite2 City database as reported by MaxMind does not match the accuracy that we obtained. This means that the correctness of the accuracy reported by MaxMind is questionable.

Bibliography

Literature

[A1] V. N. Padmanabhan and L. Subramanian. An investigation of geographic mapping techniques for internet hosts. In *Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, SIGCOMM '01, page 173–185, New York, NY, USA, 2001. Association for Computing Machinery.

[A2] P. Hillmann, L. Stiemert, G. D. Rodosek, and O. Rose. Dragoon: Advanced modelling of ip geolocation by use of latency measurements. In *2015 10th International Conference for Internet Technology and Secured Transactions (ICITST)*, pages 438–445, London, UK, 2015.

[A3] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida. Constraint-based geolocation of internet hosts. *IEEE/ACM Trans. Netw.*, 14(6):1219–1232, December 2006.

[A4] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe. Towards ip geolocation using delay and topology measurements. In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, IMC '06, page 71–84, New York, NY, USA, 2006. Association for Computing Machinery.

[A5] B. Wong and I. Stoyanov. Octant: A comprehensive framework for the geolocalization of internet hosts. In *4th USENIX Symposium on Networked Systems Design & Implementation (NSDI 07)*, Cambridge, MA, April 2007. USENIX Association.

[A6] B. Eriksson, P. Barford, B. Maggs, and R. Nowak. Posit: A lightweight approach for ip geolocation. *SIGMETRICS Perform. Eval. Rev.*, 40(2):2–11, October 2012.

[A7] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang. Towards street-level client-independent ip geolocation. In *Proceedings of*

the 8th USENIX Conference on Networked Systems Design and Implementation, NSDI'11, page 365–379, USA, 2011. USENIX Association.

[A8] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye. Ip geolocation databases: Unreliable? *SIGCOMM Comput. Commun. Rev.*, 41(2):53–56, April 2011.

[A9] S. Wallis. Binomial confidence intervals and contingency tests: Mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics*, 20(3):178–208, 07 2013.

[A10] J. J. Kester. Comparing the accuracy of ipv4 and ipv6 geolocation databases. In *24th Twente Student Conference on IT*, Enschede, The Netherlands, January 2016.

[A11] M. Gharaibeh, A. Shah, B. Huffaker, H. Zhang, R. Ensafi, and C. Papadopoulos. A look at router geolocation in public and commercial databases. In *Proceedings of the 2017 Internet Measurement Conference*, IMC '17, page 463–469, New York, NY, USA, 2017. Association for Computing Machinery.

[A12] W. Xu, Y. Tao, and X. Guan. Experimental comparison of free ip geolocation services. In *Security with Intelligent Computing and Big-data Services*, pages 198–208, Cham, 2020. Springer International Publishing.

Technical literature

[B13] J. Postel. Internet protocol. *RFC 0791*, September 1981.

[B14] S. Deering and R. Hinden. Internet protocol, version 6 (ipv6) specification. *RFC 2460*, December 1998.

[B15] K. Egevang and P. Francis. The ip network address translator (nat). *RFC 1631*, May 1994.

[B16] Y. Shirasaki, S. Miyakawa, A. Nakagawa, J. Yamaguchi, and H. Ashida. Nat444 with isp shared address. *IETF Tools*, October 2008.

Web Literature

[C17] MaxMind. https://www.maxmind.com/en/company, 2021.

[C18] MaxMind city accuracy comparison. https://www.maxmind.com/en/geoip2-city-accuracy-comparison, 2021.

[C19] Number Resource Organization. Free pool of ipv4 address space depleted. https://www.nro.net/ipv4-free-pool-depleted, 2011.

[C20] Internet Assigned Numbers Authority. https://www.iana.org/numbers, 2021.

[C21] IP2Location 'Hexasoft'. The use of ip geolocation to enable geo advertising. https://blog.ip2location.com/knowledge-base/the-use-of-ip-geolocation-to-enable-geo-advertising/, 2018.

[C22] EmpowerIT 'Ally Roos'. Everything you need to know about geo-blocking. https://www.empowerit.com.au/blog/cybersecurity/about-geo-blocking/, 2020.

[C23] IP2Location 'Hexasoft'. Fraud detection with ip2location ip geolocation. https://blog.ip2location.com/knowledge-base/fraud-detection-with-ip2location-ip-geolocation/, 2018.

[C24] Internet Assigned Numbers Authority. Iana ipv4 address space registry. https://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xhtml, 2021.

[C25] Internet Assigned Numbers Authority. Internet protocol version 6 address space. https://www.iana.org/assignments/ipv6-address-space/ipv6-address-space.xhtml, 2019.

[C26] Internet Assigned Numbers Authority. Ipv6 global unicast address assignments. https://www.iana.org/assignments/ipv6-unicast-address-assignments/ipv6-unicast-address-assignments.xhtml, 2019.

[C27] RIPE Atlas. https://atlas.ripe.net/, 2021.

[C28] phpMyAdmin. https://www.phpmyadmin.net/, 2021.

[C29] RIPE Atlas. Global ripe atlas network coverage. https://atlas.ripe.net/results/maps/network-coverage/, 2021.

[C30] Ip address by country. https://worldpopulationreview.com/country-rankings/ip-address-by-country, 2021.

Appendix A

Probe distribution

Chapter 3.5 covers the distribution of the probes of RIPE Atlas and the distribution of the probes of our ground truth. The ground truth contains probes from 179 countries. The 25 countries that host the most probes are listed in Table A.1. Some of these probes also contain an IPv6 address next to their standard IPv4 address. The total number of IP addresses for every country is also shown in Table A.1.

Place	**Code**	**Country**	**Probes**	**IP Addresses**
1.	DE	Germany	1430	2412
2.	US	United States	1383	2022
3.	FR	France	807	1297
4.	RU	Russia	629	795
5.	GB	Great Britain	579	874
6.	NL	Netherlands	510	792
7.	PT	Portugal	339	425
8.	CH	Switzerland	298	486
9.	IT	Italy	280	356
10.	CZ	Czech Republic	268	426
11.	CA	Canada	263	379
12.	UA	Ukraine	219	270
13.	AT	Austria	200	301
14.	AU	Australia	198	298
15.	SE	Sweden	188	254
16.	PL	Poland	181	231
17.	JP	Japan	168	286

18.	BE	Belgium	166	264
19.	ES	Spain	144	186
20.	DK	Denmark	134	187
21.	IN	India	122	155
22.	FI	Finland	121	169
23.	IR	Iran	120	138
24.	GR	Greece	108	166
25.	NO	Norway	104	156

Table A.1: Top 25 countries with the most probes in our ground truth.

Appendix B

Accuracy tables

Chapter 5 covers the results of the experiment. Whereas in chapter 5 only the diagrams are shown, this appendix shows all the data of the experiment in three tables. Table B.1 has the results of the 10 kilometer radius, table B.2 has the results of the 50 kilometer radius and table B.3 has the results of the 250 kilometer radius. The accuracies of MaxMind were noted from the GeoIP2 City accuracy table on February 26, 2021.

B.1 10 Kilometer Radius

Place	Country	Probes	95% CI	99% CI	MaxMind
1.	DE	43%	41% - 45%	40% - 46%	52%
2.	US	55%	53% - 57%	52% - 58%	60%
3.	FR	49%	46% - 52%	45% - 53%	45%
4.	RU	44%	41% - 47%	40% - 49%	58%
5.	GB	37%	34% - 40%	33% - 41%	55%
6.	NL	44%	41% - 47%	40% - 49%	62%
7.	PT	64%	59% - 68%	58% - 70%	48%
8.	CH	33%	29% - 37%	28% - 39%	34%
9.	IT	28%	24% - 33%	22% - 34%	24%
10.	CZ	51%	46% - 56%	45% - 57%	50%
11.	CA	50%	45% - 55%	43% - 57%	32%
12.	UA	51%	45% - 57%	43% - 59%	68%
13.	AT	38%	33% - 44%	31% - 45%	54%
14.	AU	33%	28% - 39%	26% - 40%	25%
15.	SE	46%	40% - 52%	38% - 54%	48%

16.	PL	52%	46% - 58%	44% - 60%	43%
17.	JP	44%	38% - 50%	37% - 52%	40%
18.	BE	45%	39% - 51%	37% - 53%	49%
19.	ES	50%	43% - 57%	41% - 59%	57%
20.	DK	41%	34% - 48%	32% - 50%	71%
21.	IN	42%	35% - 50%	32% - 52%	63%
22.	FI	46%	39% - 54%	36% - 56%	47%
23.	IR	42%	34% - 50%	32% - 53%	?%
24.	GR	47%	40% - 55%	37% - 57%	64%
25.	NO	53%	45% - 61%	43% - 63%	51%

Table B.1: Accuracy of the top 25 countries in a 10 kilometer radius using broadband IPs only.

B.2 50 Kilometer Radius

Place	**Country**	**Probes**	**95% CI**	**99% CI**	**MaxMind**
1.	DE	61%	59% - 63%	58% - 64%	82%
2.	US	72%	70% - 74%	69% - 74%	83%
3.	FR	65%	62% - 68%	62% - 68%	76%
4.	RU	69%	66% - 72%	65% - 73%	74%
5.	GB	54%	51% - 57%	50% - 58%	84%
6.	NL	62%	59% - 65%	57% - 66%	82%
7.	PT	80%	76% - 84%	75% - 85%	76%
8.	CH	66%	62% - 70%	60% - 71%	86%
9.	IT	49%	44% - 54%	42% - 56%	64%
10.	CZ	68%	63% - 72%	62% - 74%	80%
11.	CA	65%	60% - 70%	58% - 71%	62%
12.	UA	61%	55% - 67%	53% - 68%	76%
13.	AT	58%	52% - 63%	51% - 65%	85%
14.	AU	55%	49% - 61%	48% - 62%	70%
15.	SE	61%	55% - 67%	53% - 69%	74%
16.	PL	68%	62% - 74%	60% - 75%	60%
17.	JP	75%	70% - 80%	68% - 81%	75%
18.	BE	81%	76% - 85%	74% - 86%	89%

19.	ES	65%	58% - 71%	56% - 73%	80%
20.	DK	70%	63% - 76%	61% - 78%	85%
21.	IN	57%	49% - 65%	47% - 67%	77%
22.	FI	65%	57% - 72%	55% - 73%	65%
23.	IR	47%	39% - 55%	36% - 58%	?%
24.	GR	62%	54% - 69%	52% - 71%	73%
25.	NO	74%	67% - 80%	64% - 82%	81%

Table B.2: Accuracy of the top 25 countries in a 50 kilometer radius using broadband IPs only.

B.3 250 Kilometer Radius

Place	**Country**	**Probes**	**95% CI**	**99% CI**	**MaxMind**
1.	DE	84%	82% - 85%	82% - 86%	96%
2.	US	77%	75% - 79%	75% - 79%	90%
3.	FR	73%	71% - 75%	70% - 76%	91%
4.	RU	75%	72% - 78%	71% - 79%	89%
5.	GB	79%	76% - 82%	75% - 82%	94%
6.	NL	91%	89% - 93%	88% - 93%	98%
7.	PT	95%	92% - 97%	92% - 97%	95%
8.	CH	95%	93% - 97%	92% - 97%	99%
9.	IT	75%	70% - 79%	69% - 80%	93%
10.	CZ	90%	87% - 93%	86% - 93%	94%
11.	CA	70%	65% - 74%	64% - 76%	88%
12.	UA	74%	68% - 79%	67% - 80%	87%
13.	AT	86%	82% - 89%	80% - 90%	97%
14.	AU	65%	59% - 70%	58% - 72%	89%
15.	SE	70%	64% - 75%	62% - 77%	89%
16.	PL	85%	80% - 89%	78% - 90%	89%
17.	JP	91%	87% - 94%	86% - 94%	85%
18.	BE	97%	94% - 98%	93% - 99%	100%
19.	ES	72%	65% - 78%	63% - 80%	91%
20.	DK	95%	91% - 97%	89% - 98%	99%
21.	IN	69%	61% - 76%	59% - 78%	89%

22.	FI	79%	72% - 84%	70% - 86%	90%
23.	IR	75%	67% - 81%	65% - 83%	?%
24.	GR	93%	88% - 96%	86% - 97%	91%
25.	NO	88%	82% - 92%	80% - 93%	91%

Table B.3: Accuracy of the top 25 countries in a 250 kilometer radius using broadband IPs only.

Appendix C

Queries database

These are the queries that have been used in section 4.2. Each line number represents the query that has been used in the same step number. In query 4 and query 5, adapt the number *difference* to match the radius that needs to be checked.

```
1  SELECT `country`, COUNT(*) FROM `Probes` GROUP BY `country` ORDER BY 2
   ↪  DESC
2  SELECT `country`, COUNT(*) FROM `Probes` WHERE `address_v6` IS NOT NULL
   ↪  GROUP BY `country` ORDER BY 2 DESC
3
4  SELECT `country`, COUNT(*) FROM `Probes` WHERE `difference` <= 10 GROUP BY
   ↪  `country` ORDER BY 2 DESC
5  SELECT `country`, COUNT(*) FROM `Probes` WHERE `address_v6` IS NOT NULL
   ↪  AND `difference6` <= 10 AND `difference6` > 0 GROUP BY `country`
   ↪  ORDER BY 2 DESC
```

Every query groups by country and orders on the COUNT(*) argument. This is particularly useful, because MaxMind also groups by country. The first query just counts how many IPv4 addresses each country has. This is possible, because every probe in the ground truth has an IPv4 address. The second query needs to account for the fact that not every probe has an IPv6 address. The query for step four checks how many IPv4 addresses are within the given radius and because every IPv4 address is found in the GeoLite2 City database, we do not have to check for unresolved addresses. The last query does need to account for that problem. Not every IPv6 address is resolved in the GeoLite2 City database, so we have to check that the IPv6 address is not unresolved.

Appendix D

Code

D.1 Import probes into database

The following code is used to read all important data from the probe files that were downloaded from the RIPE Atlas API. The data is read from every file and it is pushed to the local database. The parameters, the full SQL queries and some repeated steps are omitted to reduce the code length.

```
1  import mysql.connector as mc
2  import json
3  import os
4
5  #Connect to the database and perform the insert query.
6  def insert_json(PARAMS):
7      query = "INSERT STATEMENT"
8      args = (ARGS)
9
10     try:
11         mydb = mc.connect(CONNECT ARGS)
12         cursor = mydb.cursor()
13         cursor.execute(query, args)
14         mydb.commit()
15
16     except mc.Error as error:
17         print(error)
18
19     finally:
20         cursor.close()
21         mydb.close()
22
23 #Loop through all .json type files.
24 directory = r'/DIR'
25 for entry in os.scandir(directory):
26     if (entry.path.endswith(".json")):
27         with open(entry) as f:
28             data = json.load(f)
```

```
29
30          #Save the important data to a temporary variable.
31          for feature in data['features']:
32              id          = feature['properties']['id']
33              address_v4  = feature['properties']['address_v4']
34              ...
35
36              insert_json(PARAMS)
```

D.2 IP Address comparison

The following code is used to calculate the difference in location of two coordinates. The IP address from the ground truth is queried to the GeoLite2 City database and if a match is found, then the information that is needed is pushed to the local database. Again, the parameters and the full SQL queries are omitted to reduce the code length.

```
1   import mysql.connector as mc
2   import json
3   import geoip2.database as gd
4   from geopy.distance import geodesic
5
6   def processdata():
7       query1 = "SELECT * FROM Probes"
8       args1 = ()
9
10      #Try to make a connection to the local database and execute the query.
11      try:
12          mydb = mc.connect(CONNECT ARGS)
13          cursor = mydb.cursor(buffered=True)
14          cursor.execute(query1, args1)
15          result = cursor.fetchall()
16
17          for res in result:
18              reader = gd.Reader('./GeoLite2-City.mmdb')
19
20              #Look up the IP address in the GeoLite2 City database.
21              try:
22                  response = reader.city(res[1])
23
24                  #Check if the accuracy radius is set within the GeoLite2
                    ↪  City database.
25                  if (response.location.accuracy_radius == None):
26                      mm_accuracy = -1
27                  else:
28                      mm_accuracy = response.location.accuracy_radius
29
30
31
```

```
32                  #Calculate the distance difference between the two
                    ↪  coordinates.
33                  coordinate1 = (response.location.latitude,
                    ↪  response.location.longitude)
34                  coordinate2 = (res[4], res[5])
35                  mm_difference = geodesic(coordinate1, coordinate2).km
36
37                  reader.close()
38
39                  #Push the results back to the local database.
40                  query2 = "UPDATE Probes"
41                  args2 = (ARGS)
42                  cursor.execute(query2, args2)
43                  mydb.commit()
44
45              #If the IP address is not found in the GeoLite2 City database,
                ↪  pass it and then the distance difference remains 0.
46              except geoip2.errors.AddressNotFoundError:
47                  pass
48
49      except mc.Error as error:
50          print(error)
51
52      finally:
53          cursor.close()
54          mydb.close()
55
56  processdata()
```

D.3 Calculating the statistics

The following code is used to automatically calculate the 95% and 99% confidence intervals for every country and every radius. These intervals are pushed back to the local database. Again, the parameters and the full SQL queries are omitted to reduce the code length.

```
1   import mysql.connector as mc
2   import math
3
4   #These are the formulas to calculate the Wilson score interval.
5   def wilson(p, n, z):
6       left     = (((p+((z**2)/(2*n))) -
        ↪  (z*(math.sqrt(((p*(1-p))/n)+((z**2)/(4*(n**2)))))))/(1+((z**2)/n)))*100
7       right    = (((p+((z**2)/(2*n))) +
        ↪  (z*(math.sqrt(((p*(1-p))/n)+((z**2)/(4*(n**2)))))))/(1+((z**2)/n)))*100
8       return "{0}, {1}".format(left, right)
9
10
11
```

```
12  def statistic():
13      query1 = "SELECT * FROM Statistics"
14      args1 = ()
15
16      #Try to make a connection to the local database and execute the query.
17      try:
18          mydb = mc.connect(CONNECT ARGS)
19          cursor = mydb.cursor(buffered=True)
20          cursor.execute(query1, args1)
21          result = cursor.fetchall()
22
23          #Extract country, sample size and accuracies.
24          for res in result:
25              country = res[0]
26              n       = res[2]
27              10km    = res[3]
28              50km    = res[4]
29              250km   = res[5]
30
31              #Calculate the 95% confidence intervals.
32              answer10_95  = wilson(10km, n, 1.96)
33              answer50_95  = wilson(50km, n, 1.96)
34              answer250_95 = wilson(250km, n, 1.96)
35
36              #Calculate the 99% confidence intervals.
37              answer10_99  = wilson(10km, n, 2.576)
38              answer50_99  = wilson(50km, n, 2.576)
39              answer250_99 = wilson(250km, n, 2.576)
40
41              #Push the results back to the local database.
42              query2 = "UPDATE Statistics"
43              args2 = (ARGS)
44              cursor.execute(query2, args2)
45              mydb.commit()
46
47      except mc.Error as error:
48          print(error)
49
50      finally:
51          cursor.close()
52          mydb.close()
53
54  statistic()
```