IP Based Geo Location

By

Ajmal Hussain **2006-NUST-BIT-02**

Farrah Batool **2006-NUST-BIT-02**

A Project report submitted in partial fulfillment
of the requirement for the degree of
Bachelors of Information Technology

NUST School of Electrical Engineering & Computer Science
National University of Sciences & Technology
Islamabad, Pakistan

2010

SLAC
NATIONAL ACCELERATOR LABORATORY

IP Based Geo Location

Final Report

"Geo location is the process of geographically locating any internet connected node across the world. Geo location techniques attempt to find the latitude and longitude of internet connected nodes."

Dr Ali Khayam
Dr Aimal Tariq
Farrah Batool
Ajmal Hussain

6/11/2010

Table of Contents

Introduction to Geo location:

Geo location is the process of geographically locating any internet connected node across the world. Geo location techniques attempt to find the latitude and longitude of internet connected nodes.

With the growing trend of internationalization and with the expansions of markets to global level with the boom in information technology, now the importance of location based products and services has increased manifold. Online businesses are in intense need of customer location awareness to market their products fulfilling the needs of the customers in the best ways. Providing location specific products and services to the customers, driving location based advertisement campaigns and mirror selection in peer to peer systems increase the need of geo location of internet connected systems.

Types of Geo location:

There are two basic ways of geo locating internet connected rows:

i) **Static** Geo location using static information from various databases like GeoIP database.

ii) **Dynamic** geo location using RTT measure to algorithmically find the location of the nodes.

Static Geo location:

There are a large number of database dependent static geo location tools available which provide location information on the basis of IP addresses or domain names. A few example of such static geo location tools are GeoIP Tool (http://www.geoiptool.com), geo tool (http://www.geotool.com), hostIP Info (http://www.hostip.info) and NetGeo . All these tools mainly obtain location information from the user entered data in websites with registration forms.

Geo locating nodes statically is not adequate because of the following reasons:

- There is a rapid growth in number of internet connected nodes and it has become very difficult to maintain such large location information in the databases. Tackling the dirty data is a cumbersome task.
- IP addresses are dynamic in nature and nodes can move without changing IP addresses so it puts a big question mark on accuracy of Database driven geo location techniques.

Dynamic Geo location:

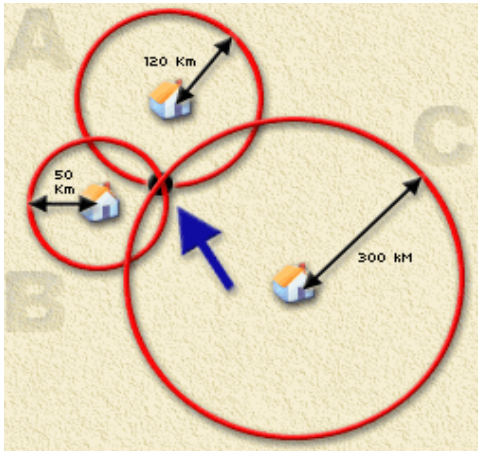
The above stated flaws in static database driven geo location techniques give rise to the need of dynamic geo location process without relying on hard to maintain databases of location information. So we intend to use RTT delay measurements to estimate the geographic locations in a more systematic manner. RTT based geo location requires no huge infrastructure and simply uses Ping RTT values from a large number of known location landmarks to a unknown location target to find its location.

Geo Location Algorithms:

There are various dynamic geo location algorithms which try to estimate the location of internet connected nodes in different ways. Examples of such algorithms are Trilateration, CBG, Apollonius and TBG. All these algorithms take RTT delay as the key ingredient in the estimation process. Two algorithms considered by us in our project are Trilateration and Apollonius. Following is the explanation of both algorithms.

Trilateration:

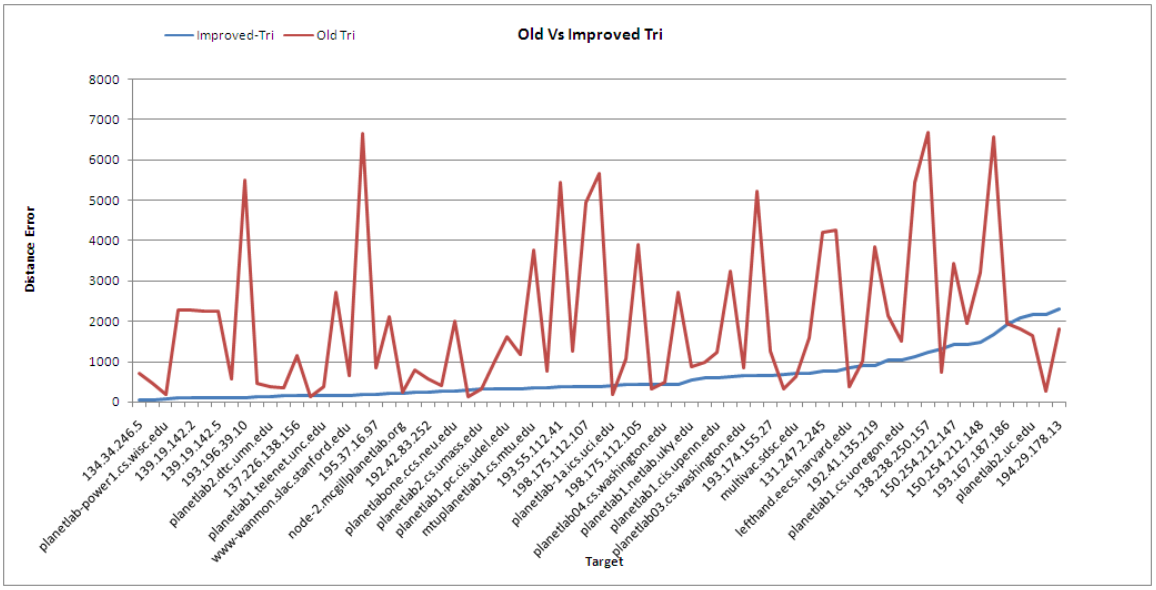
As the name suggests, Trilateration techniques uses three RTT values to estimate the coordinates of target node. In Trilateration technique, when several landmarks ping the specified Target, three landmarks with smallest RTT values (probably nearest located landmarks) are selected and then algorithm tries to estimate the location on the basis of overlapping region of the circles drawn around these selected landmarks. Figure.1 explains this in pictorial format. In Trilateration technique, the center of the overlapping region is considered as the location of the target node.



There are many ways to implement trilateration technique. Linear Least Square Method, Nonlinear Least Square Method , Circles intersection with Clustering and trilateration in 2D and 3D.

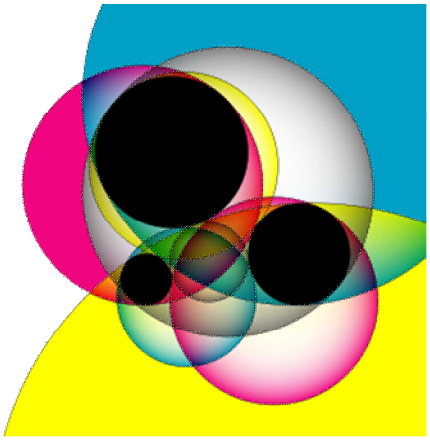
Previously we were implementing Linear Least Square Method to implement trilateration. With this method the 50% of targets showed Distance Error above 1000 km. going through paper on "Performance evaluation of a TOA-based trilateration method to locate terminals in WLAN" which is discussing different Positioning Algorithms. In this paper it is mention that *Linear Least Square Method* is not very accurate and it just provide initial position which can be used in other positioning algorithms (i.e. *Nonlinear Least Squares* and *Independent Time GPS Least Squares*) as the initialization value for their iterations.
Going through wiki for Trilateration implemention we found that method to be easy and straight forward.(http://en.wikipedia.org/wiki/Trilateration). But the algorithm mentioned on wiki is solving trilateration in 3D plane. We have implemented it in 2D plane.

This performance of trilateration was quite improved as shown

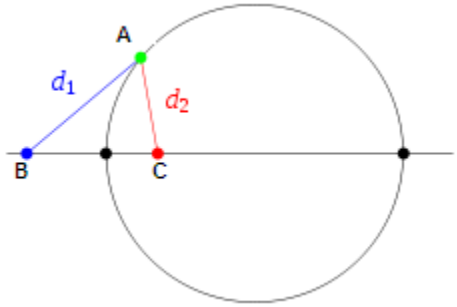


Apollonius:

Another quite efficient geo location technique is Apollonius which also relies on RTT delay data from the known location landmarks to the unknown location target to estimate the target location. The main difference between Trilateration and Apollonius is the way in which the two techniques estimate the location. Contrary to Trilateration, Apollonius technique doesn't take the overlapping region of the three circles into consideration; instead it draws the tangent circles touching all three landmark circles. In this way, Apollonius can result into formation of multiple (or sometimes even no) solution circles and a key part is to select one candidate circle out of these multiple circles. In this technique, the center of finally selected Apollonius circle is considered as the target location. The circle selection in Apollonius is a critical part and is further explained later.

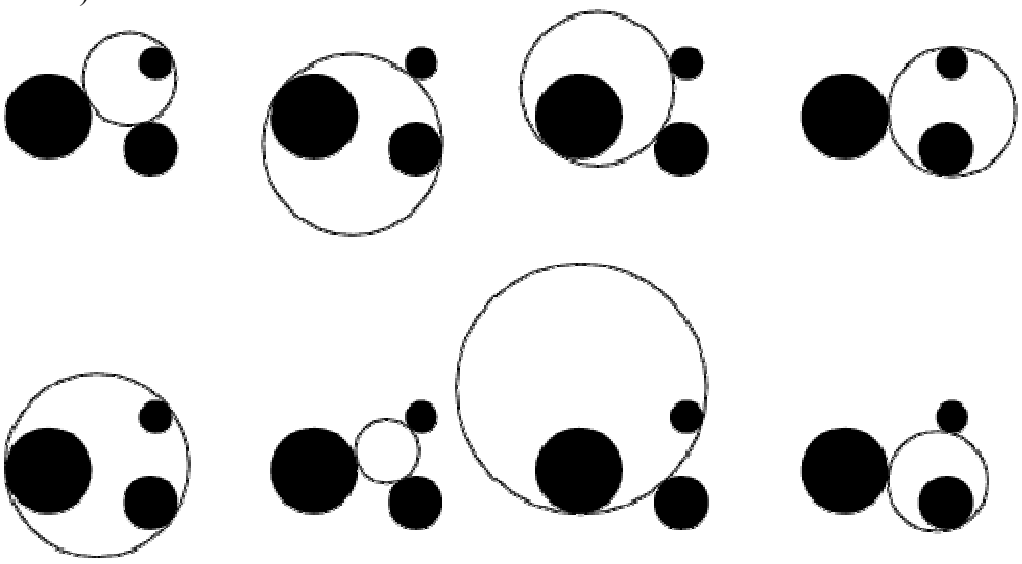


Apollonius circles are three special circles say c1,c2,c3. The set of points with a constant ration of distances d_1/d_2 to two fixed points know as foci is one of the circle.
c1 is unique circle passing to vertex A of triangle ABC (figure) that maintain constant ratio of distances to other two vertices B and C. similarly c2 is passing to vertex B of triangle and maintain constant ratio of distances to other two vertices A and C . and so is the case with c3.

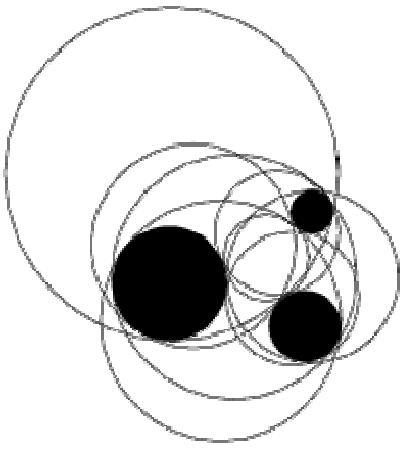


Other definitions:
The circle that touches all three excircles of a triangle and encompasses them (Kimberling 1998, p. 102).
One of the eight circles that is simultaneously tangent to three given circles (i.e., a circle solvingApollonius' problem for three circles).

Apollonius circles are up to eight in number that are tangent to three circles (defined above).



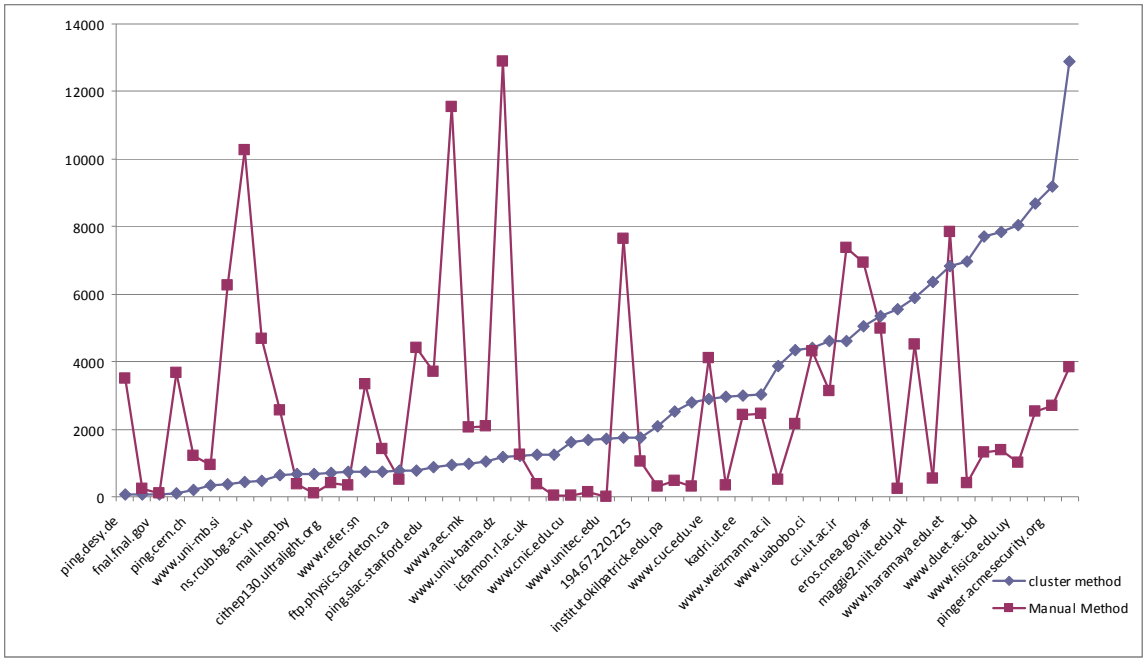
Or



In our algorithm to solve IP based Geolocation we have tried to embed this technique to find our required target. To get three circles of Apollonius we used our 3 landmarks giving minRTT after pinging target node. This minRTT is then converted to 1 way delay which is then used my distance=RTT*alpha to get radius of the circle. In this way we get three circles of Apollonius which are used by our algorithm to find circles that are tangent to these three circles. The resultant circles are up to eight in number or less.
After getting eight(or minimum) circles we need to find one circle which is near our target to find. We have used cluster approach to find this one circle and then compared this technique with manual method ( ) to view the performance of our cluster approach.

In manual method we find the centre point of each circle (out of eight or minimum circles) and convert this x,y point to latitude longitude and compare its Distance Error with the actual location (latitude longitude) of target (given by geoIPtool.com) and circle giving minimum distance error is considered to be the required circle.
This approach cant be used as it is using static database (geoIPtool) which is not reliable in few cases i.e may conatain dirty data. And this approach was just to check performace of our algorithm by getting our required circle each time as no circle other then chosen by geoIPTool could give minimum distance error (if data is not dirty in geoIPtool).

In cluster approach we have tried to find the cluster of circles of Apollonius. After getting eight circles(maximum) of Apollonius we try to locate the cluster of circles. Let's suppose, we obtained n circles using apollonius. Now, we took each of these n circles and for each circle s, we calculated the distance between it's center and the center of all other n-1 circles and summed it. Let we call this sum as Sum(si). Then we took average of this value.
In this way we calculated the average distance value for each of circle and finally one circle with smallest average value was considered as Apollonius circle.

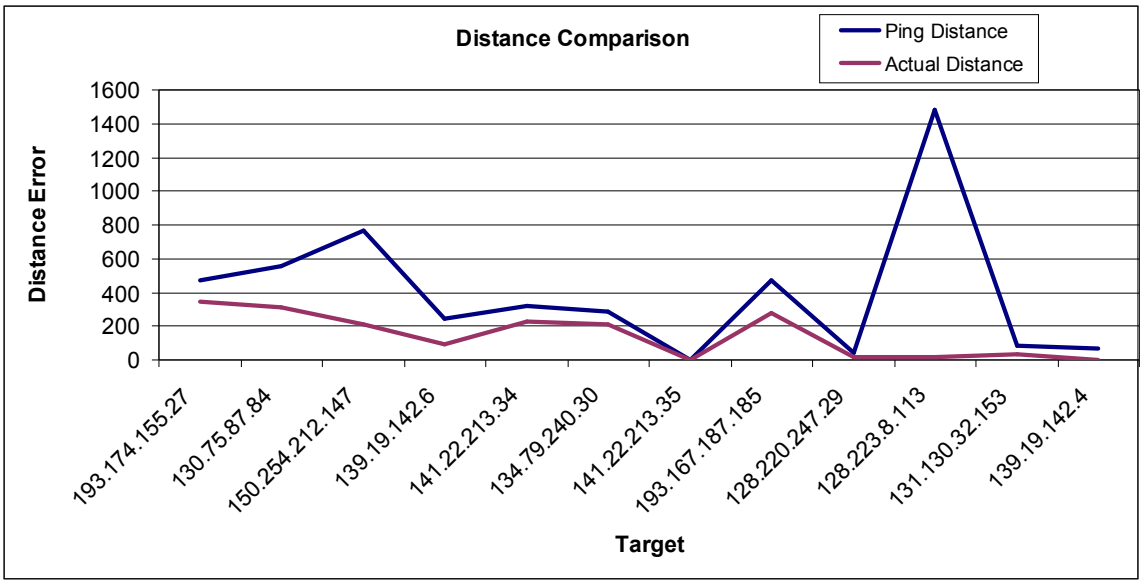


The cluster based approach takes one circle as apollonius circle and the final coordinates are calculated on the basis of this circle while in the other approach of using all the circles for calculating latitude and longitude uses all the circles, so cluster approach can perform equal to the manual approach only in the best case scenario. However, the main advantage of cluster based approach is that it uses a proper mathematical scheme to select the apollonius circle and doesn't solely rely on Geo IP results. On the other hand, the manual approach relies on the Geo IP results and takes Geo IP results as reference.
In some cases cluster approach is performing better then manual approach which is negating the fact that manual approach choses the circle with minimum error distance then how this could be minimum then 'minimum error distance circle' its because landmarks keep on changing due to which clusters keep on changing so minimum error distance circles keep on changing.

Performance of Apollonius

Like in trilateration, in Apollonius we are using distance=RTT*Alpha to get the radius of circles of Apollonius. The way couldn't check performance of Apollonius technique. So in order to see the actual performace of Apollonius technique we performed a test.
We selected known landmarks and a target (which was also a landmark .. but acting as target for other landmarks) then calculated distance (http://www.movable-type.co.uk/scripts/latlong.html) between each landmark to that target and hardcoded this distance to code i.e replaced
e.g distance= 786.00.

By doing so we no need RTT's. Purpose of doing this task was to test Apollonius efficiency in case it do not had to suffer the pinging error.



Doing analysis on results showed that 60-70% of Distance Error is because of pinging error. Hence Apollonius can perform far better if there is no pinging error.

Delay to Distance Mapping:

A critical and basic part of a dynamic geo location technique is the process of mapping RTT delay value into a proper distance value. The correct estimation of the location of any node depends on the correct mapping of delay into distance.

The distance value that is obtained after converting RTT into distance is used in the core of the geo location algorithm. By constructing circles of radius equal to distance value obtained after mapping. In geo location techniques, such circles are drawn around the finally selected landmarks (landmarks with smallest RTT values) and the overlapping region of all the circles is used to find the location of the target node. Now, if during delay to distance mapping, we under estimate the distance then there is a possibility of finding no overlapping region resulting in failure in finding the location of the target. This is depicted in the following figure:

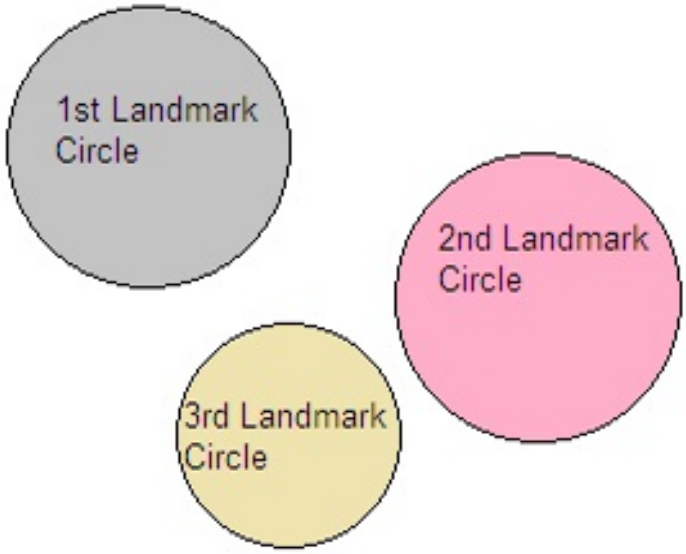


Figure 1: No overlapping regions of circles to estimate location

And the following figure depicts a situation in which there is an overlapping region of the three circles and hence we can estimate the latitude and longitude of the target node.

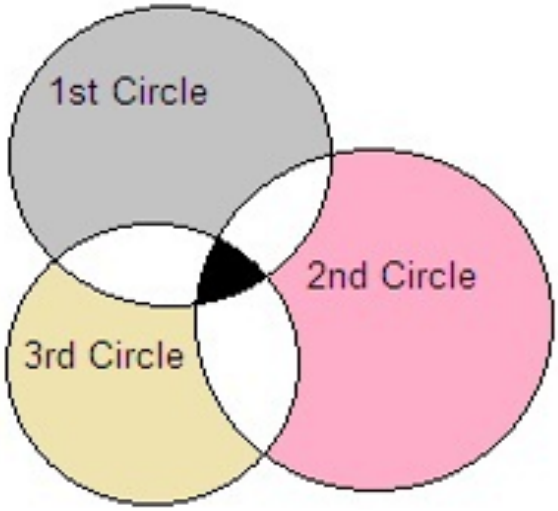


Figure 2: An overlap region available for location estimation

From the above figures, the significance of a correct mapping of RTT into Distance becomes clear. Although a larger value of distance (distance over estimation) will make sure the availability of an overlapping region but using un-necessarily large distance value will result in incorrect overlapping region resulting in the wrong estimation of the target location. Hence, for any geo location algorithm, delay to distance mapping is a key part and success of such techniques greatly relies on this important factor.

In order to convert RTT into Distance (which is then used as the radius of the circles), a conversion factor is used known as alpha. So using this conversion factor Alpha, we get the distance using following formula:

Distance=RTT*Alpha---------- (1)

However it may be noted that in this formula we take one way delay to calculate distance.

Selection of Alpha Value:

The RTT value obtained from various landmarks to the target node depends on various factors like queuing delay, availability of direct path, connectivity type (wired or wireless) etc. The connectivity infra structure in different parts of the world is not uniform and hence it becomes very difficult to reach to a global alpha value to be used for all the landmarks of the world.

In order to find out some reasonable alpha value for different region of the world, we performed a detailed alpha analysis on the basis of following data:

- We had Ping RTT data from around 170 landmarks scattered across the world to the targets (targets were also from the same landmarks—one at a time). This data was collected from the tests of several days and in order to nullify the effect of any incorrect measurements, data of various tests was averaged out.
- We categorized data in various parts on the basis of various geo graphical regions namely North America, Latin America, South Asia, Europe, Middle East, East Asia, Russia and Australia.
- Then we performed Alpha analysis on each region in order to reach to some Acceptable alpha value for various regions.

In the following part results of Alpha analysis are depicted:

Alpha value analysis for landmarks of different regions:

In this part of analysis, we analyzed the Alpha values for all the targets in all regions from a particular landmark (taking one landmark at a time and doing alpha analysis for it against all the targets in the world).

The graph below is for three **African** landmarks. From the graph we can see that for majority of the targets, Alpha value lies below 100. In fact, for any of the three targets, around 98% of the Alpha values were below 100. It's also clear from the graph that for the targets at a smaller distance (possibly in the same region as of Landmark) alpha value is quite smaller and then it increases slightly with the increase in distance. As we know that in the geo location techniques, finally selected landmarks are those which are nearest to the target (with smallest RTT values) so the targets within a small range of distance are of more interest for our analysis.

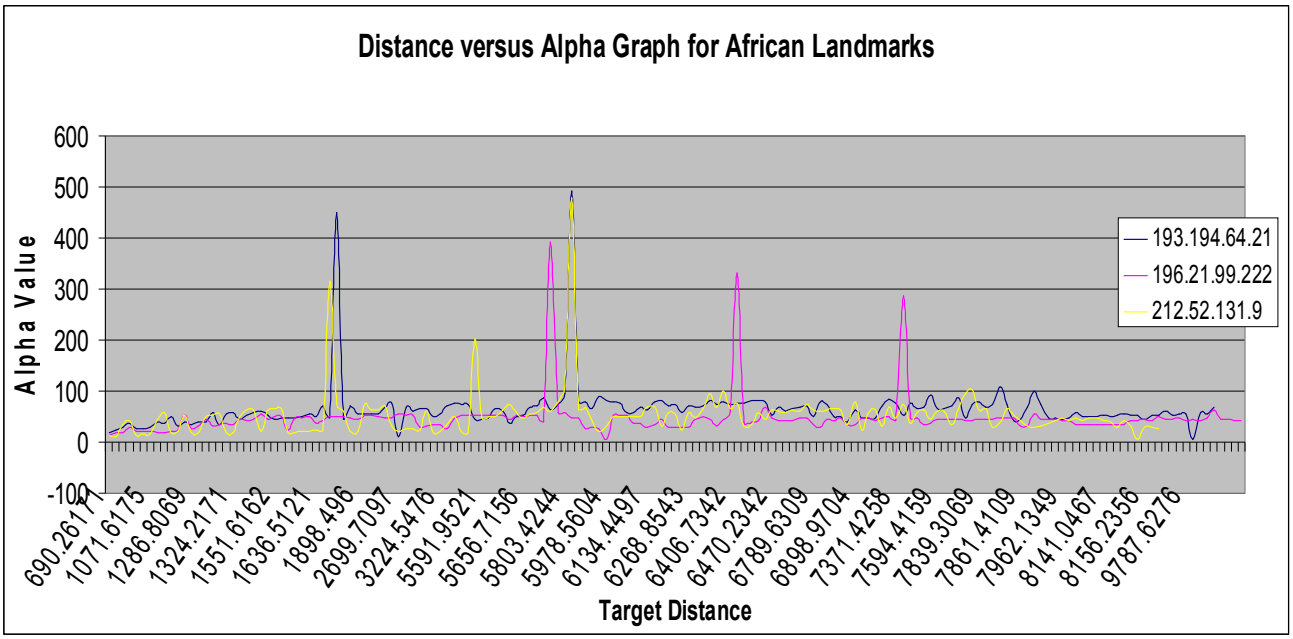


Figure 3: Distance versus Alpha Graph for African Landmarks

Now after getting a clear idea that distance has no significant impact on Alpha value (at least for these African landmarks), in the following graph, average alpha values and Median Alpha values are plotted. The purpose of taking Median into consideration is to see an alpha value not disturbed by the outliers.

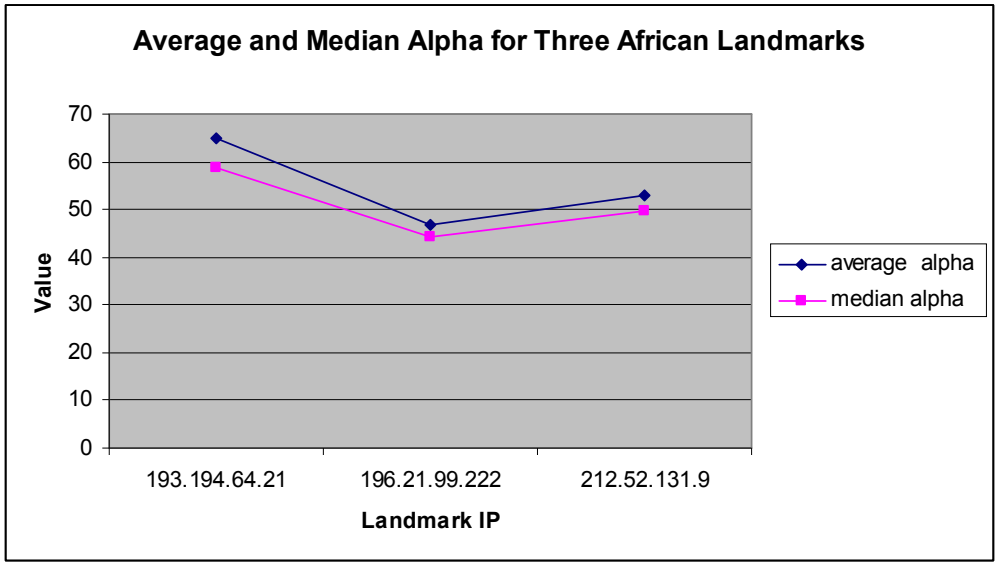


Figure 4: Average and Median alpha for different African landmarks

From the above graph we can make a rough judgment that a reasonable alpha value to be used is between 40 and 60 (while in our TULIP implementation, we were using an Alpha value of 100).

However, to reach to any conclusion, we need to take into consideration the results of other region's landmarks as well.

Next, I did same analysis for North American landmarks. Below is the graph for three North American landmarks which also shows that there is just a slight increase in alpha with increasing distance. Again there is a concentration of alpha values between 50 and 100 and there are just a few values above hundred.
(Note: To keep the graph observable, I have plotted three landmarks but rough plot of graph showed that the behavior of other landmarks is also same).

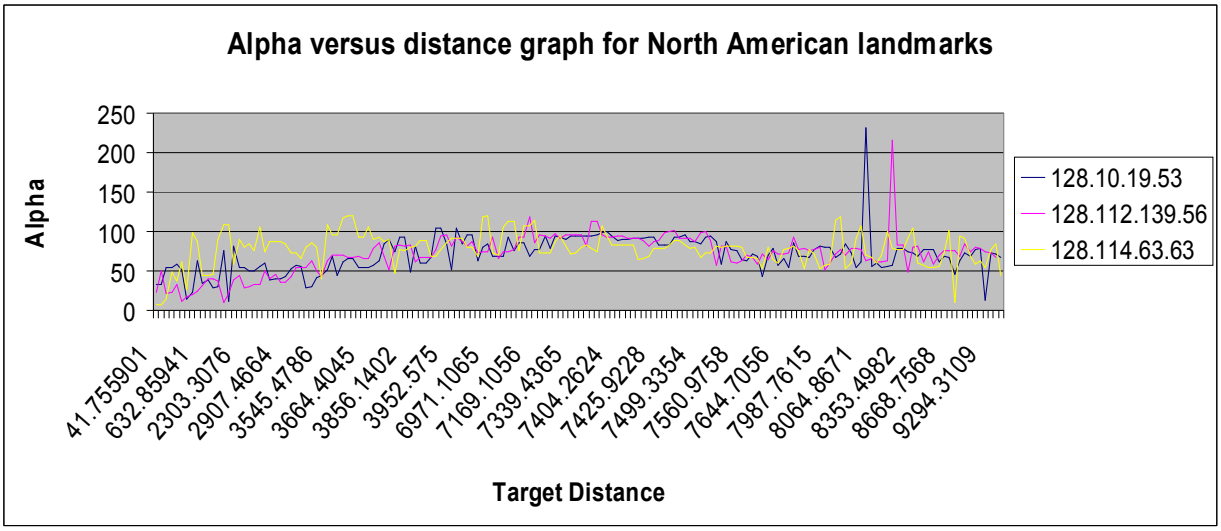


Figure 5: Distance versus Alpha graph for North American landmarks

As it's obvious from the above graph that alpha value is lying between 50 and 100 for majority of targets, so to reach to a better approximation of alpha value, I calculated median and average alpha values for each of the landmark (for all targets around the world). The resultant values are shown in the graph below. Again the graph reinforces our belief of having alpha value in between 60 and 80. In this graph, closeness of average and median values is due to absence of outliers.

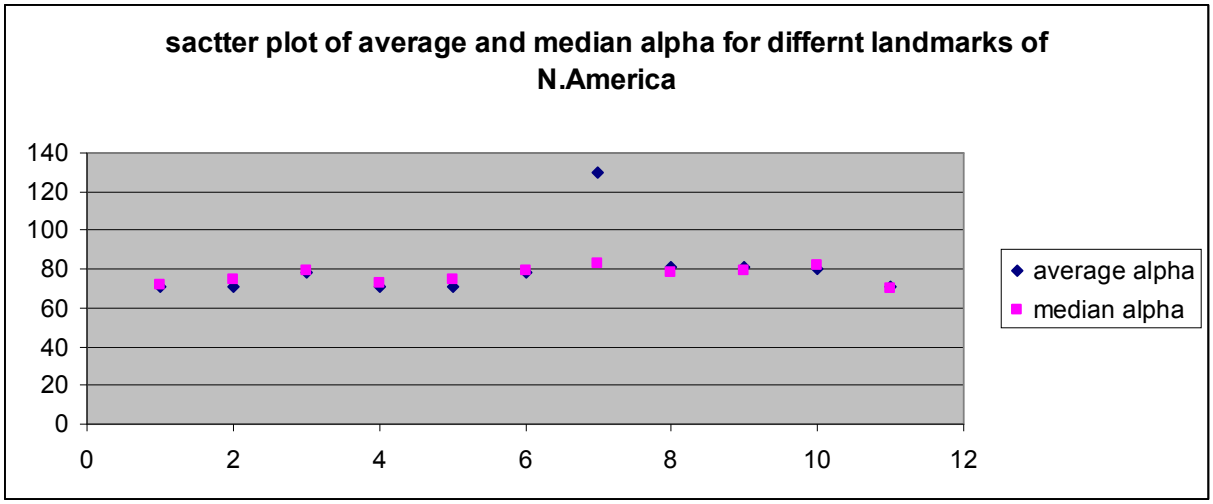


Figure 6: Scatter plot of average and median alpha for North American landmarks

So far from two regions, I found no big impact of distance on alpha values and to make sure that this is true for landmarks of any region, I have analyzed the landmarks of other regions as well similarly.

Following graph shows relationship between distance and alpha value for **European landmarks** (for targets in different regions) and again we can see that growing distance doesn't change alpha value and alpha values oscillate below 100. So again, the graph suggests use of same alpha value in all regions irrespective of distance.

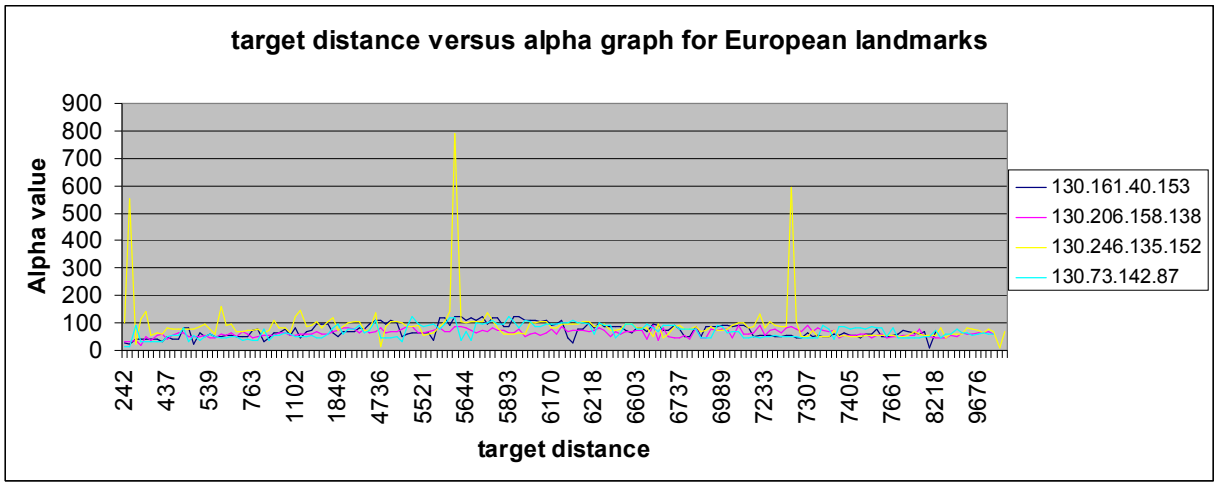


Figure 7: Distance versus alpha graph for European landmarks

Same graph for **Latin American** Region's landmark is shown below which again shows that majority of Alpha values irrespective of distance, oscillates below 100 (but we can observe a slight increase in alpha value with increase in distance especially in the targets closer to the landmark).

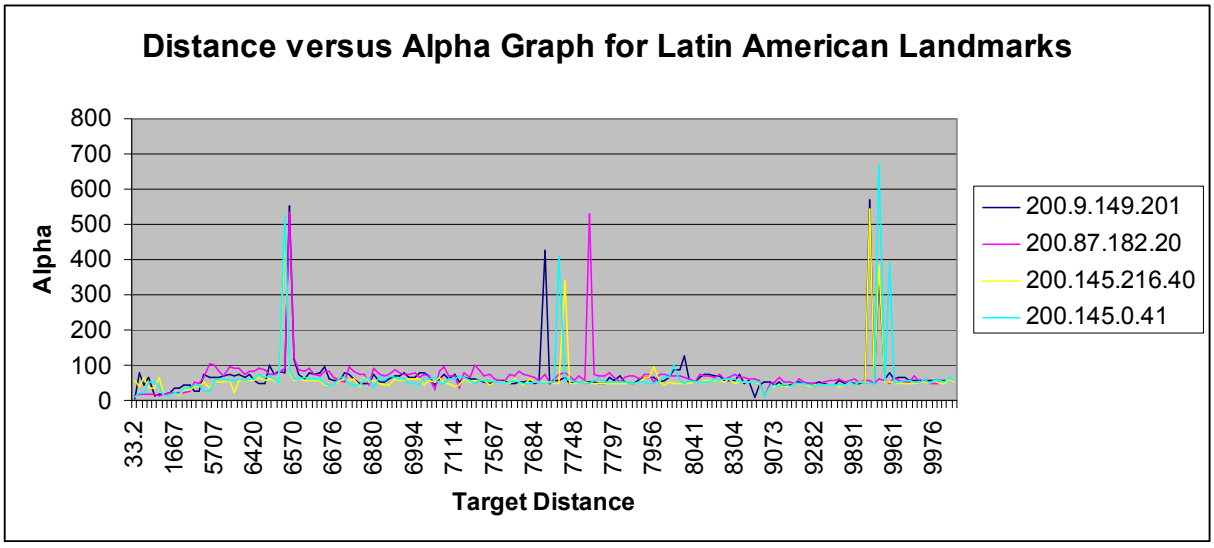


Figure 8: Distance versus alpha graph for Latin American landmarks

In order to make it sure that similar alpha patterns are obtained for all region's landmarks, I have plotted same distance versus alpha value graphs for landmarks of Russia and South Asian landmarks as well which are shown below:

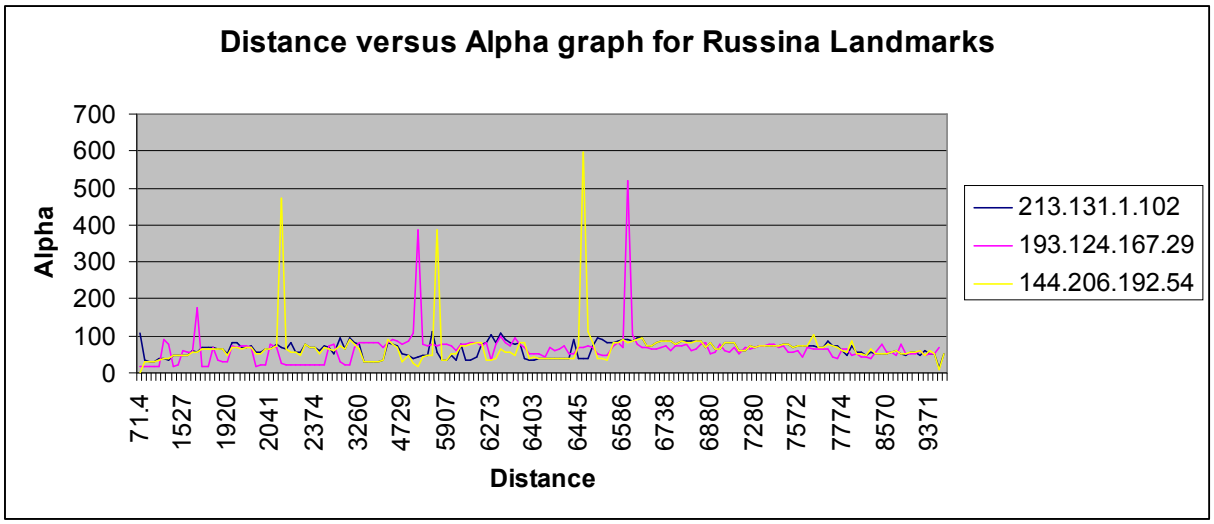


Figure 9: Distance versus alpha graph for Russian landmarks

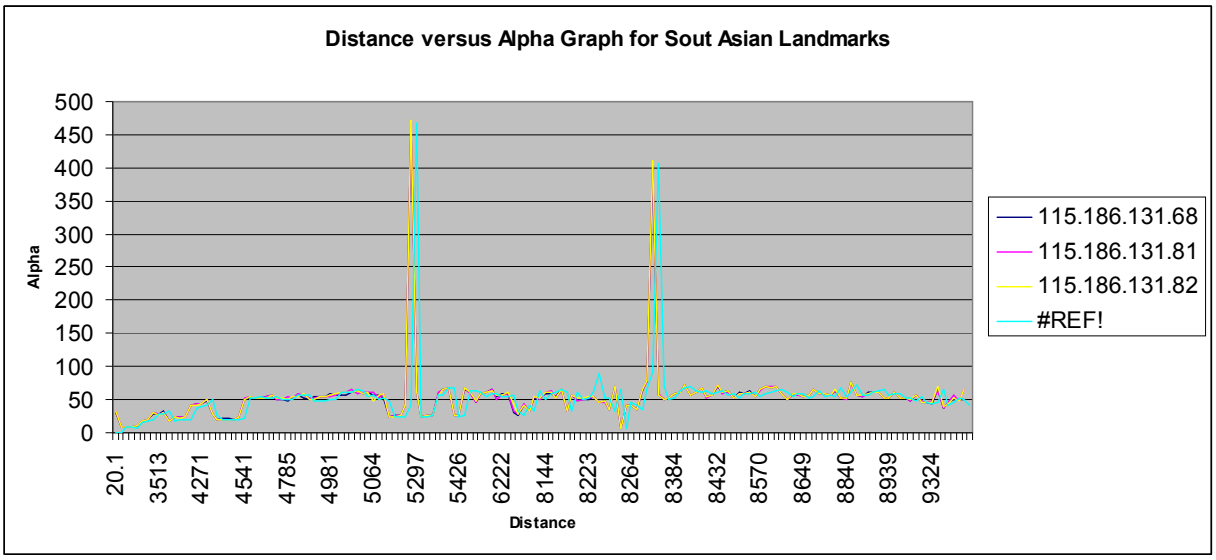


Figure 10: Distance versus alpha graph for south Asian landmarks

Analyzing the Standard Deviation, Median and Average of Alpha for different landmark regions:

In order to reach to a reasonable alpha value to map delays into distances, in this part of analysis, I have analyzed the alpha values of different targets from a particular landmark on the basis of Average and Median Alpha value.

Average Alpha value within 5000 Km distances:

Keeping in mind the fact that in our geo location techniques, finally selected landmarks are those which are most nearest to the target, I analyzed landmark to targets alpha values (average and median) within a distance of 5000Km. The reason of taking a relatively high distance to analyze is to ensure that I get enough sample space to analyze to avoid any distortion in results due to a minority values.

Below is the graph showing average and median alpha values for different landmarks of Africa region (for targets within 5000Km distance). Considering median to be more robust and less affected by outliers, we can say that for African landmarks, an alpha value between 30 and 50 can correctly map delay to distance.

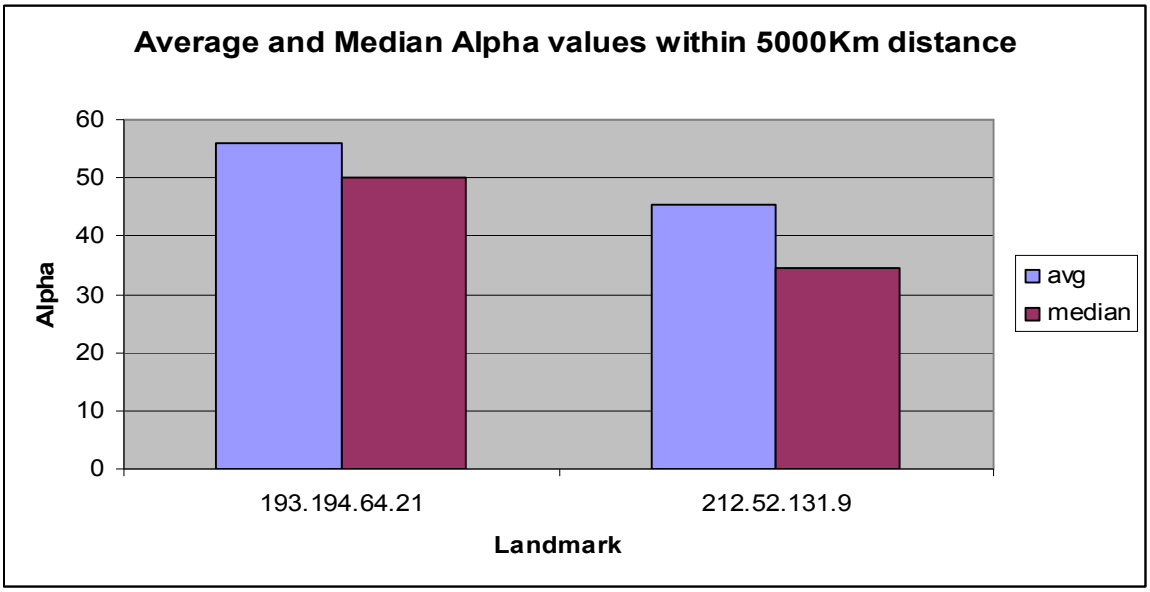


Figure 11:average and median alpha for African Landmarks within 5000Km distance

I did the same analysis for European landmarks and following graph shows that for European landmarks, median alpha values are in 40-80 range. Note that in the previous graph of African landmarks, there were only two landmarks to analyze and hence, for such a small sample space we can't reach to any conclusive point but in this graph there are results for 10 European landmarks and can provide better results. From the graph we can say that in any case, alpha value lies in the range of 45-80.

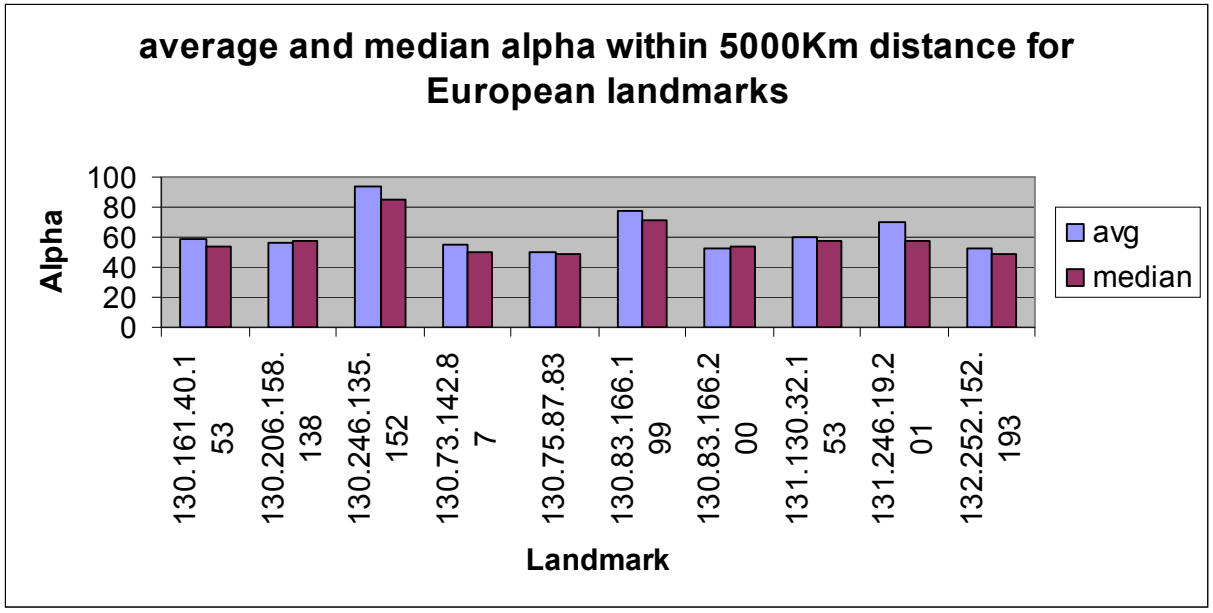


Figure 12:Avg and Median Alpha for European landmarks within 5000Km distance

From the above two analyses, again the idea of having different alpha values for different targets looks more logical. To further confirm this hypothesis, I have plotted same median and average alpha values graph for North American landmarks (shown below) and from the graph we can see that a reasonable alpha value for North American region lies in the range of 45-85.

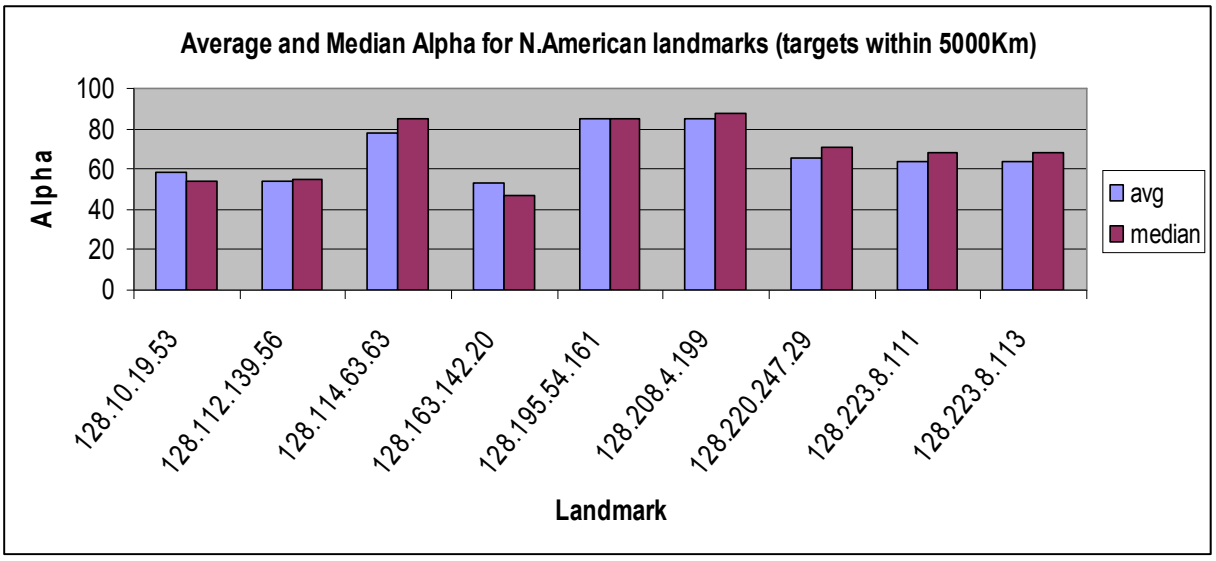


Figure 13:Avg and Median alpha for N.American landmarks within 5000Km distance

When I did the same average and median Alpha analysis within a distance of 5000Km for Latin American landmarks, it was observed that for the Latin American landmarks, we get lower alpha values than other regions because of the fact that even for less distant targets; from Latin American landmarks we are getting higher RTT values. The graph suggests us to use an alpha value in the range of 20-40 for this region's landmarks.

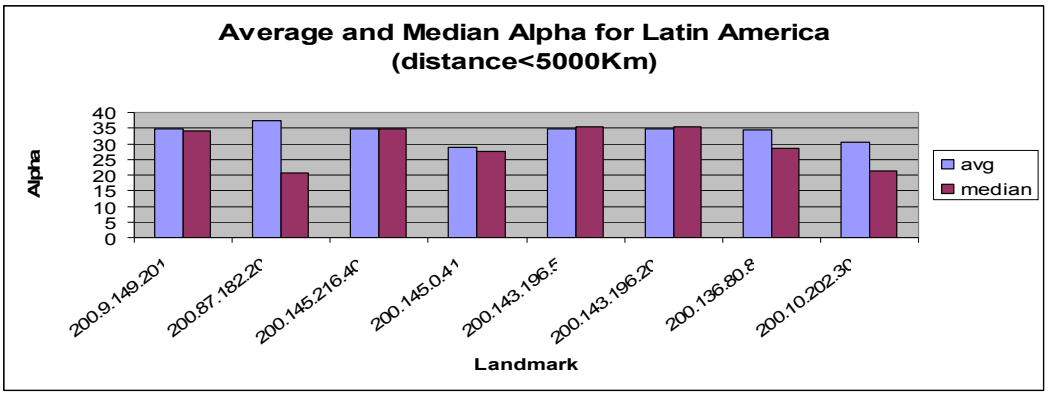


Figure 12: Avg and median alpha for Latin American landmarks

If we do the similar analysis for South Asian landmarks, we get the following graph. Again just like Latin American landmarks, we are getting low alpha values because of higher RTT values even at lower distance. For this region alpha values lie between 20 and 40 so for this region a reasonable alpha value to use might be some value in this range (like 30).

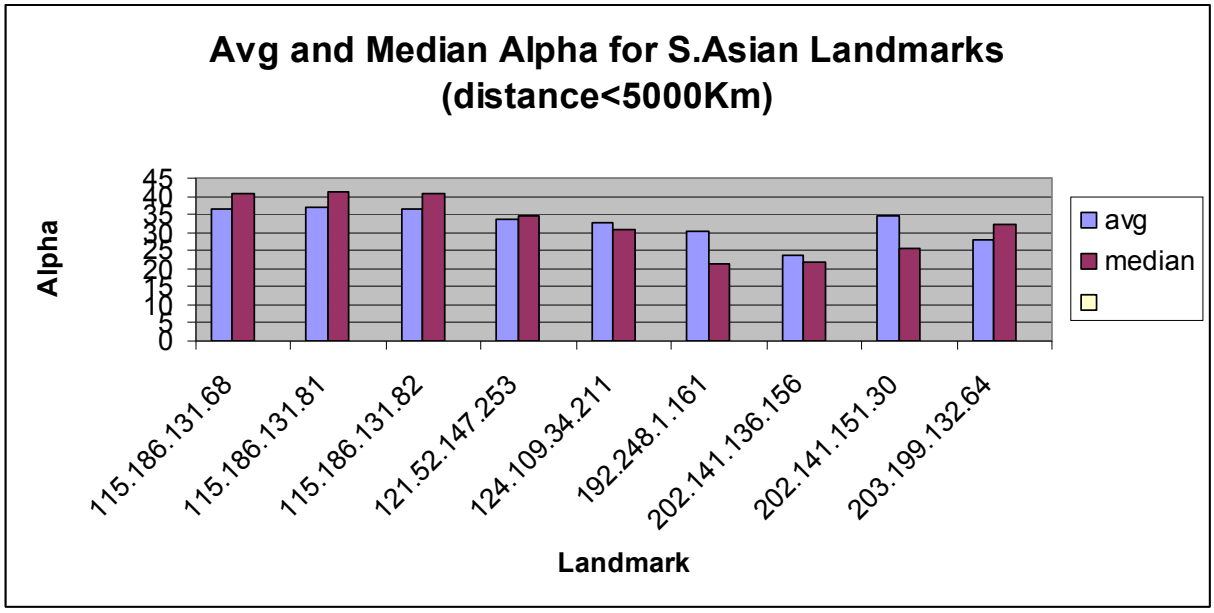


Figure 14: Avg and median alpha values for S.Asian landmarks

After conducting a thorough analysis on the data for Alpha selection of various regions of the world, It was observed that Average alpha value was the best candidate for final

selection as average alpha was producing the best results in terms of Distance error. This is clear from the figure 12 where we can see that for majority of the targets, average alpha is giving the best results (for 18 targets in this case).

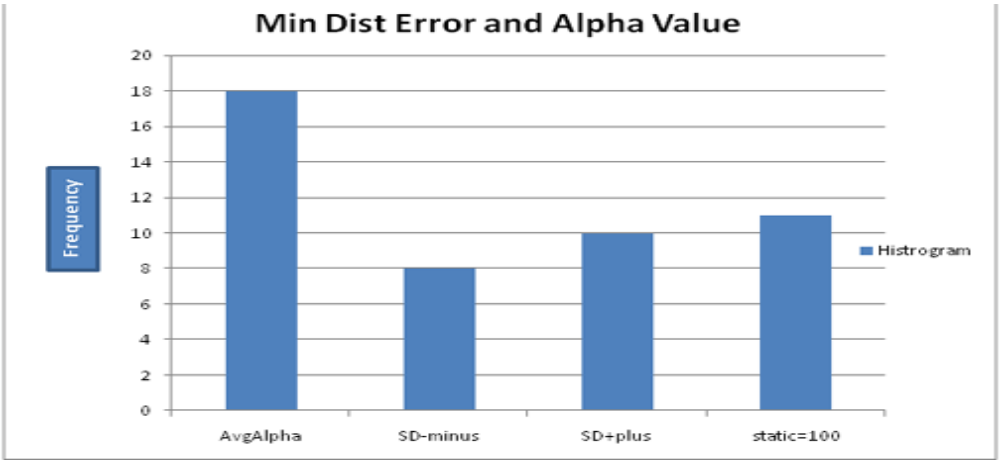
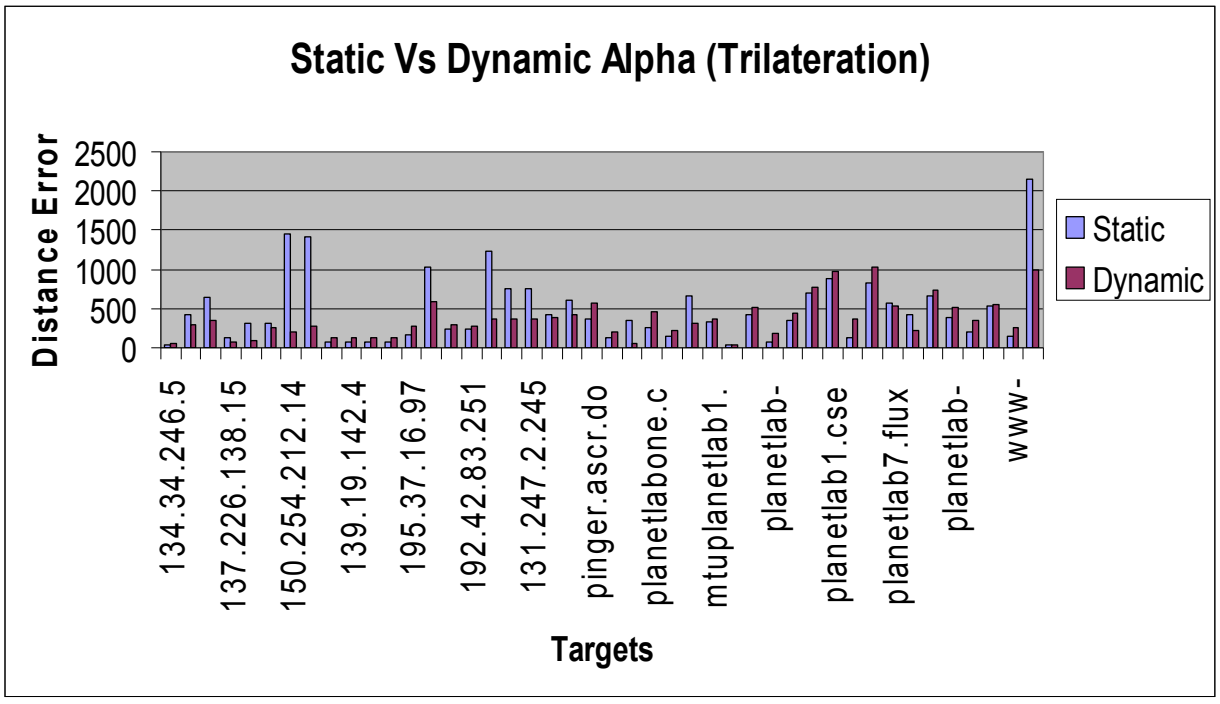


Figure 14: Behavior of various alpha choices for targets (frequency of better results)

Geo location Results with Static and Dynamic Alpha:

After selection of dynamic regional alpha for each region, the next logical step was to analyze the impact of regional alpha on the results of Trilateration and Apollonius. It was observed that the results were drastically improved after applying regional dynamic alpha for delay to distance mapping. Figure 15 shows the comparison of Trilateration results with static fixed and dynamic regional alpha selection.



The improvement in the results with dynamic alpha reveals the fact that region with varying internet connectivity states and varying routing schemes can't have same conversion factor for delay to distance mapping. Each region requires a different alpha value based on the connectivity in that region.

Impact of Landmark Density on Geo location Results:

After optimizing the Alpha value for different regions, when tests were run for targets of various regions, it was observed that results for only North America and Europe were improved and for other regions, Error distance still mounted to thousand of Kilometers. One observation was that this is mainly due to instability of internet connectivity in these regions and better results in Europe and North America are due to better and stable connectivity in those areas. However, another thing observed during the analysis was that these Europe and N. America each also possess a large number of landmarks compared to other regions. So the impact of landmark's density can't be ignored. It was observed that most of the times poor results appear mainly because of the fact that no landmark is found in close vicinity of the target and hence eventhe three final selected landmarks with minimum RTT may have RTT values above 300ms or so.

So to analyze this geographical distance impact on Error distance, I have conducted a region by region analysis in which I have done following main tasks:

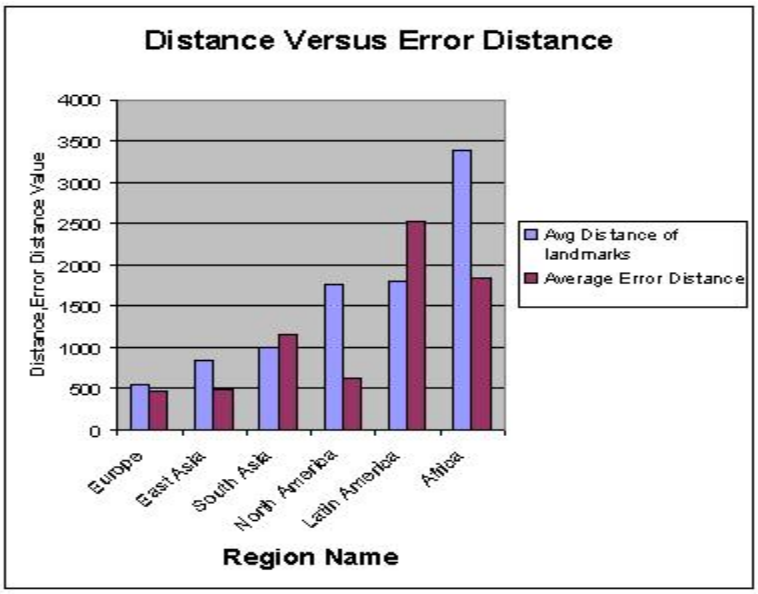
- First, I found distance between different landmarks of a region (their inter-distance) and then calculated the average distance for each region. E.g. if there are 10 landmarks in a certain region, I calculated distance between a landmark and all other landmarks in that region and in this way inter-distance of all landmarks was calculated and then averaging out these results, I got one average distance value for each region.
- Then I took targets of each region, and run tests using the Apollonius geolocation technique for them (using average alpha value of that particular region). In this way I got error distance for each target and then we calculated average Error distance for each region.
- Now the analysis began. I wanted to investigate that what behavior will be there in error distance when the average geographical distance is higher and what it will be when average geographical distance is lower. Of course, an ideal situation was that when there is high geographical distance between landmarks, i.e. landmarks are far apart from each other then the error distance should be higher compared to the case when we have high number of landmarks in a particular region (i.e. we have low average geographical distance).
- Following table shows the average geographical distance and average error distance for each region.

Region	"Avg Distance of Landmarks"	"Avg Error Distance"
Europe	555.6404153	440.0434656
East Asia	841.0220395	2966.42115

South Asia	1006.405099	6379.043878
North America	1756.877965	804.486103
Latin America	1790.503606	5431.530291
Africa	3387.983944	4743.33271

Based on the above data, a graph was plotted to see the relationship between distance and Distance Error in a graphical manner. The graph is shown below:

Same result shown in the form of bar graph is:



From the graph, we see that in Europe where the number of landmarks is high and there is less average distance between landmarks, then we get good results in terms of error distance. The trend looks fine when we move to next regions where average geographical distance is comparatively higher (error distance has also increased for these regions) but we see a different trend for North America where average geographical distance between landmarks is higher than other regions like East Asia and South Asia but yet, the Error distance is quite lower. In fact this result depicts that even though landmarks are geographically dispersed and not closely located in this region, the connectivity in this region is so better that we get quite reasonable results even then.

For all other regions, we see some impact of geographical distance and from this we can say that hopefully results for those regions can also improve if we could have more landmarks in those regions.

It's clear from the result that for all those targets for which Error Distance is less than 1000Km, the MinRTT of the first selected landmark is less than even 60ms. This means

that if for any target we are able to find a nearest landmark (with MinRTT<60), then we will be getting good results and smaller this minRTT is better are the results. From the attached results, we can also see that for all those targets for which we got error distance above 2000Km, the MinRTT (column D) was above 100ms. From these results, we can roughly conclude that if we mange to have our landmarks distributed in the world in such a way that we get low MinRTT for any target, the error distance can drastically reduce.

References

http://en.wikipedia.org/wiki/Circles_of_Apollonius
http://mathworld.wolfram.com/ApolloniusCircle.html
http://upload.wikimedia.org/wikipedia/commons/2/2a/Apollonius_circle_definition_labels.svg
http://www.cs.rit.edu/~ark/543/module05/trilateration.pdf
http://galileo.cs.telespazio.it/liaison/download/Performance%20evaluation%20of%20a%20TOA-based.pdf
http://en.wikipedia.org/wiki/File:Trilateration.svg