A Sample-Based Approach to Data Quality Assessment in Spatial Databases with Application to Mobile Trajectory Nearest-Neighbor Search

Bagher Saberi
Department of Electrical and Computer Engineering,
Isfahan University of Technology
m.saberishahrbabaki@ec.iut.ac.ir

Nasser Ghadiri
Department of Electrical and Computer Engineering,
Isfahan University of Technology
nghadiri@cc.iut.ac.ir

Abstract: *Spatial data is playing an emerging role in new technologies such as web and mobile mapping and Geographic Information Systems (GIS). Important decisions in political, social and many other aspects of modern human life are being made using location data. Decision makers in many countries are exploiting spatial databases for collecting information, analyzing them and planning for the future. In fact, not every spatial database is suitable for this type of application. Inaccuracy, imprecision and other deficiencies are present in location data just as any other type of data and may have a negative impact on credibility of any action taken based on unrefined information. So we need a method for evaluating the quality of spatial data and separating usable data from misleading data which leads to weak decisions. On the other hand, spatial databases are usually huge in size and therefore working with this type of data has a negative impact on efficiency. To improve the efficiency, we need a method for shrinking the volume of data. Sampling is one of these methods, but its negative effects on the quality of data are inevitable. In this paper we are trying to show and assess this change in quality of spatial data that is a consequence of sampling. We used this approach for evaluating the quality of sampled spatial data related to mobile user trajectories in China which are available in a well-known spatial database. The results show that sample-based control of data quality will increase the query performance significantly, without losing too much accuracy. Based on this results some future improvements are pointed out which will help to process location-based queries faster than before and to make more accurate location-based decisions in limited times.*

Keywords: Algebraic operations, Data integration, Quality metrics, Location data, Trajectory Analysis

1. Introduction

In the last decade, we witnessed the public use of spatial data. Google Earth is a good example which is becoming increasingly popular among the general public as well as being used by specialists to develop new applications. Spatial data and spatial reasoning are playing an increasingly important role in today's advanced computing demands such as context-aware systems, pervasive computing and decision support systems [1]. Geographic information is a prerequisite for many specific fields and collecting geographic data needs a lot of labor, resources and money. Moreover, it is impossible to re-collect a part of the geographic data. Therefore, guaranteeing the quality of spatial data is of great importance in enhancing the scientific aspect of the relevant decision-making [2]. On the other hand, huge increase in the volume of spatial data, the routine but complex nature of spatial operations and provision of non-expert driven computer information systems have resulted in vastly increased potential for error and misleading information in spatial data products [3].

Assessing and reporting quality of spatial data is a very important aspect of producing and using this type of information. Quality assessing method has a great impact on the efficiency and usability of data. Although there are many methods for this job but we tried to find a way that has a better efficiency.

The sample-based method, regarding its static or periodic nature, can lead to a great efficiency improvement in spatial data quality assessment algorithms. In this approach quality of data sources is estimated by sampling the base data sources and using the quality of these samples to assess quality of any information product.

The primary contributions of this paper are: (1) introducing the sample-based spatial data quality assessment method, (2) evaluating the effect of sampling on the quality of spatial data, and (3) improving the performance of trajectory nearest-neighbor search controlled by the proposed data quality assessment method. The paper yields insights as to how quality estimates for the base data sources can be used to provide quality estimates for information products generated from them. Thus, it is neither necessary nor useful to consume a lot of resources to inspect entire databases.

The rest of the paper is organized as follows: Section 2 explains the basic concepts of data quality and spatial data quality. Section 3 develops our proposed method. Section 4 introduces data sources, tools that we used in this study, experiments and results. Section 5 points out some open research areas in this context and section 6 concludes the paper.

2. Basic Concepts

In recent years data quality has gained more and more importance in science and practice. This is mainly due to an extended use of data warehouse systems, cooperative information systems and a higher relevance of customer relationship management. In this section, we give an overview of data quality assessment in general, followed by more specific field of spatial data quality assessment. The concept of trajectory nearest neighbor query is also introduced in last part.

2.1. Data Quality

In order to assess data quality, first of all a clear picture of data quality is needed. Many researchers noticed that the benefit of data depends heavily on completeness, correctness, consistency and timeliness. These properties are known as DQ dimensions. One of the major causes for the failure of information systems to deliver can be attributed to data quality. The major challenge of data quality research is to define data quality from the consumer's point of view in terms of fitness for use and to identify dimensions of data quality according to that definition. More than 100 data quality dimensions were uncovered in early research works but many of these dimensions are not very useful in many areas. Some of most important measures of data quality are relevance, accuracy, timeliness, accessibility, utility, interpretability, completeness, coherence and comparability.

On the other hand, we need a method to assess data quality and then deliver the results to the user. Quality reporting is the preparation and dissemination, on a regular or irregular basis, of reports conveying information about the quality of data. A quality report provides information on the main quality characteristics of an information product so that the user should be able to assess product quality. In the optimal case quality reports are based on quality indicators. Metadata is considered to be the main quality report method but many other methods have been proposed [16,17,21].

The data quality assessment template specifies the information that each reviewer should assemble for the target data system. The template organizes the metadata and other information, by data system and assessment process. It also serves as an outline for the assessment report.

There are many quality assessment methods. Cross-domain analysis can be applied to data integration scenarios with dozens of source systems. It enables the identification of redundant data across tables from different, and in some cases even the same, sources. Data validation algorithms verify if a value or a set of values is found in a reference data set. Domain analysis can be applied to check if a specific data value is within a certain domain of values. Matching algorithms are used to identify duplicates such as two customer records that refer to the same customer and many other methods that are currently in use [17].

2.2. Spatial Data Quality

The context within which geospatial data are used has changed significantly during the past ten years. Users have now easier access to geospatial data but typically have less knowledge of the geographical information domain. So they have limited knowledge of the risk related to the use of low quality geospatial data. Moreover, spatial data and many spatial data products are being used today for making very important decisions that may have very strong impact on people's lives. Using spatial data with low quality may even have legal complexities and therefore liability is another aspect of data quality [15].

Many methods for assessing the quality of spatial data have been proposed over the years. In general, these methods are classified in two major groups. (1) Methods that are based on internal quality. Internal quality corresponds to the level of similarity that exists between the data produced and the "perfect" data that should have been produced. In practice, the evaluation of internal quality does not use the perfect data that has no real physical existence since it is an "ideal" dataset, but uses a dataset of greater accuracy than the data produced, which is called "control data" or "reference data". (2) Methods that use external quality. The concept of external quality corresponds to the level of concordance that exists between a product and user needs, or expectations. External quality implies that quality is not absolute and the same product can be of different quality to different users. While methods exist to evaluate internal data quality, the evaluation of external quality remains a field that has been explored very little.

Many of these approaches use artificial intelligence and expert systems [4,5,13,14,18,19,20,22]. However, efficiency is not the primary concern in many of such sophisticated methods. Meta data is almost the most common method for data quality assessment and dissemination.

Among the quality assessment methods, sample-based approach can have a huge impact on efficiency. This is a new method that almost has been neglected in the past but with a good design can be very useful. The large amount of spatial data leads to complexity of data quality measurement, which in turn will cause performance impairment. To solve this problem, data quality control based on the sampling method can be used. In other words, we can use a carefully selected subset of our big dataset to assess its quality. However, shrinking the volume of spatial data and in the same time preserving the quality can be very important. For example accuracy, completeness and topological consistency can be lost when we apply a sampling method on a spatial dataset. So this method should be used very carefully and samples should be selected precisely.

2.3. Trajectory Nearest Neighbor Search

Finding nearest neighbor is a common operation in spatial data context. Clearly we have a spatial object, like a point or a route, and we want to find one object in a collection that has the minimum distance. Some spatial objects, like routes, can be very large. Meaning that they could consist of so many spatial points that spatial operations, like finding the nearest neighbor and its distance could be very time consuming. So if we can find a way to shrink the size of them and at the same time preserve the quality, we would improve the efficiency.

The essence of this study includes conducting a well-known spatial operation, trajectory nearest-neighbor search, using the complete spatial datasets and objects and then sampling the same data sources, performing the same operation and evaluating its effect on change of the quality level of results. Sampling is carried out in different rates so we can observe the effect of this parameter. Using this method, decision makers can determine whether the quality of spatial data after sampling meets their needs or not.

We applied our method on a dataset of cell phone trajectories as a spatial data source. Then we performed a particular spatial operation, like finding nearest neighbor, on them. In the last level, the same routine was applied using the sampled data. This paper shows the change in quality level of these spatial data after applying the sampling procedure.

3. The Proposed Method: Sample-Based Spatial Data Quality

Our approach is to take samples from original data, determine quality of the data in these samples, and use that information to estimate the quality of any information product that can be derived from original data. Sampling is carried out only once or on some predetermined periodic basis. The major advantage of this approach is that only the base data need to be sampled. The quality of data or number of identified deficiencies can be context dependent. Thus, the quality measure used for a given data will vary according to its use.

Any information product that can be derived from original data is result of some combination of algebraic operations. Therefore quality of any information product can be calculated by applying some algebraic operation on quality of base data. So, the sampling method can also be used for measuring quality of data, even if we have multiple tables of data as sources. Every data that might be requested from a database is result of a set of operations like restriction, Cartesian product, union and projection on basic tables of data. So we can calculate quality of every information product if we have a view of quality of base tables [4]. Furthermore this idea is extendable to the multiple independent databases and data integration [5]. This could be a wide area for future research too.

For example if we should use Restriction to be applied to a table in order to get the desired data units, then we use the sample taken from the source table to assess quality of information product [4]. Or in the case of Union of two tables suppose that no duplicates exit and the first table has N data units and P_1 is the estimate for the acceptable data units and n_1 is the size of sample and M, P_2 and n_2 are corresponding factors for the second table then an estimate of the fraction of acceptable data units in the union of two tables is shown as below [4]:

$$P = (n_1 * P_1 + n_2 * P_2) / (n_1 + n_2) \qquad (2)$$

Determining the quality of an information product when there are duplicate data units is considerably more difficult but possible.

It's possible to say that for all of the other algebraic operations, like Join, Difference and Product, we can obtain an estimate of information product quality by applying some operations on the assessed quality of source data. This is a very useful property of sample based approach in the context of spatial data.

Under this circumstance, a large enough sample needs to be taken so that defective records appear in the sample [4]. A standard rule of thumb is that the sample should be large enough so that the expected value of the number of defective items is at least two. Since sampling with replacement is a binomial process, then n, the size of the sample, must satisfy the inequality [5]:

$$n \geq 2 / (1 - P) \qquad (1)$$

Where p represents the true proportion of acceptable data units. Clearly, there needs to be some estimate for the value of P in order to use this inequality. One way of estimating *P* is by taking a preliminary sample before initiating the first round of sampling.

4. Experimental Results

In this section, we present the results of applying our proposed sample based spatial data quality assessment as a control mechanism for efficient processing of trajectory nearest neighbor search.

4.1. Datasets

For this case study we used 36 mobile phone trajectories from Microsoft GeoLife[1] as our source data. These trajectories are from different users, in different times, and in different sizes so we can neglect the effect of this factors. Each one of these trajectories shows a path that a user has fallowed during an arbitrary time of day and hence includes a number of spatial points and each point is consisting of its spatial attributes like latitude, longitude and altitude. It also contains some collateral information like time and date. Table 4.1 shows one of these trajectories in its complete shape. This trajectory is actually the smallest one and only consists of 25 points. Size of trajectories is variable from very small, containing about 25 points, to very large trajectories that include tens of thousands of points.

The GeoNames[2] database is another geospatial database that presents useful spatial data from many countries in the world including longitude, altitude and latitude, population and administrative division of important points, like cities, rivers, lakes and many geographical entities. For this paper we took a part of source data from this database. Figures 4.1 and 4.2 show samples of these two databases.

[1] http://research.microsoft.com/en-us/projects/geolife/default.aspx
[2] http://www.geonames.net

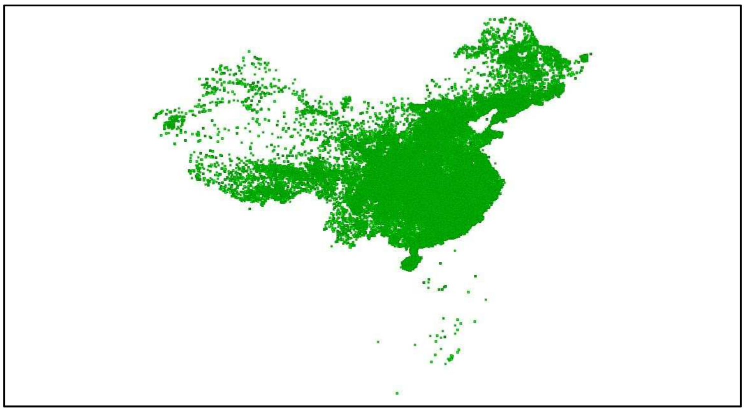

Figure 4.1: China in GeoNames database

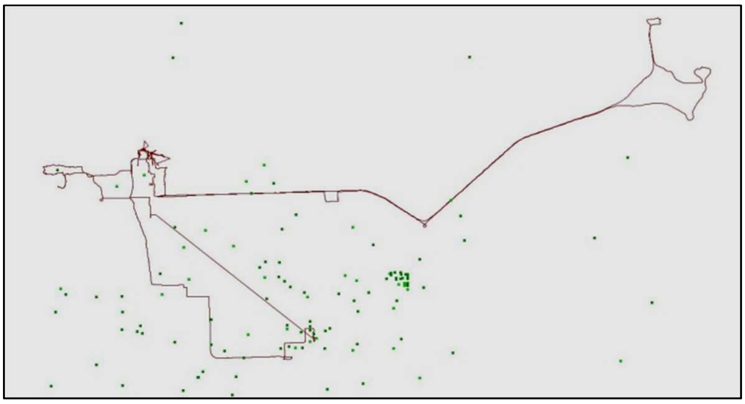

Figure 4.2: A GeoLife trajectory and its neighbors in GeoNames database

4.2. Experiments

Sampling algorithm for spatial data quality is of great importance and serves as a promising research area [2,8,10]. In this paper we used the periodic method for sampling. So in the first step our sampling rate is ½. Table 4.2 shows the results for the smallest trajectory that was mentioned earlier. The sampling procedure was applied for 6 other different sampling rates including 1/3, 1/5, 1/10, 1/15, 1/20 and 1/30, to all 36 trajectories. You can find the results for the smallest trajectory in the tables 4.1 to 4.7. Figure 4.3 depicts the same trajectory in its complete and sampled forms visually.

Sampling rate in this context is an open area for research too. If, for example, the actual error rate is 1 percent and a sample of size 10 is taken, then only occasionally will an error show up in the sample.

We tried to estimate the effect of sampling on the quality of spatial data and operations and demonstrated the results. First we found nearest neighbor to the complete trajectory and measured the spheroid distance between our trajectory and its nearest neighbor. Then we took samples from trajectories in different sampling rates and fallowed the same procedure to observe the effect of different sampling rates on quality of finding nearest neighbor and its distance.

For our study we used some of spatial data quality metrics like positional and attribute accuracy, precision and logical consistency to examine the quality. In sample based approach all of spatial data quality measures are affected. For instance if a spatial entity is up to date, its samples may be not. Or if an entity is topologically consistent, it’s samples, because does not include all of its original data, may be inconsistent. As for the updateness factor, because some of the information is lost in the sampling process, updateness is affected too.

Table 4.1: The original data

DATA_ID	LATITUDE	LONGITUDE	ALTITUDE	DATE_PASED	DATE_CURRENT	TIME_CURRENT
1	40.007409	116.321086	207	39904.84966	04/01/2009	20:23:31
2	40.007474	116.320949	152	39904.84972	04/01/2009	20:23:36
3	40.007477	116.320802	145	39904.84978	04/01/2009	20:23:41
4	40.007493	116.320649	138	39904.84984	04/01/2009	20:23:46
5	40.007507	116.320515	141	39904.8499	04/01/2009	20:23:51
6	40.007528	116.320364	150	39904.84995	04/01/2009	20:23:56
7	40.007535	116.320204	156	39904.85001	04/01/2009	20:24:01
8	40.007452	116.319977	142	39904.85007	04/01/2009	20:24:06
9	40.007458	116.319847	141	39904.85013	04/01/2009	20:24:11
10	40.00757	116.319821	138	39904.85019	04/01/2009	20:24:16
11	40.007656	116.31977	145	39904.85024	04/01/2009	20:24:21
12	40.007684	116.319675	153	39904.8503	04/01/2009	20:24:26
13	40.007683	116.319678	155	39904.85036	04/01/2009	20:24:31
14	40.007678	116.319686	155	39904.85042	04/01/2009	20:24:36
15	40.007678	116.319686	155	39904.85045	04/01/2009	20:24:39
16	40.007681	116.319686	155	39904.85047	04/01/2009	20:24:41
17	40.007721	116.3197	154	39904.85059	04/01/2009	20:24:51
18	40.007708	116.319719	154	39904.85065	04/01/2009	20:24:56
19	40.007675	116.31976	153	39904.85071	04/01/2009	20:25:01
20	40.007627	116.319788	152	39904.85076	04/01/2009	20:25:06
21	40.007553	116.319809	152	39904.85082	04/01/2009	20:25:11
22	40.007585	116.319709	149	39904.85088	04/01/2009	20:25:16
23	40.007625	116.319627	148	39904.85094	04/01/2009	20:25:21
24	40.007654	116.319563	147	39904.851	04/01/2009	20:25:26
25	40.007664	116.31951	146	39904.85105	04/01/2009	20:25:31

Table 4.2: First data sampling

DATA_ID	LATITUDE	LONGITUDE	ALTITUDE	DATE_PASED	DATE_CURRENT	TIME_CURRENT
2	40.00747	116.3209	152	39904.85	04/01/2009	20:23:36
4	40.00749	116.3206	138	39904.85	04/01/2009	20:23:46
6	40.00753	116.3204	150	39904.85	04/01/2009	20:23:56
8	40.00745	116.32	142	39904.85	04/01/2009	20:24:06
10	40.00757	116.3198	138	39904.85	04/01/2009	20:24:16
12	40.00768	116.3197	153	39904.85	04/01/2009	20:24:26
14	40.00768	116.3197	155	39904.85	04/01/2009	20:24:36
16	40.00768	116.3197	155	39904.85	04/01/2009	20:24:41
18	40.00771	116.3197	154	39904.85	04/01/2009	20:24:56
20	40.00763	116.3198	152	39904.85	04/01/2009	20:25:06
22	40.00759	116.3197	149	39904.85	04/01/2009	20:25:16
24	40.00765	116.3196	147	39904.85	04/01/2009	20:25:26

Table 4.3: Second sampling

DATA_ID	LATITUDE	LONGITUDE	ALTITUDE	DATE_PASED	DATE_CURRENT	TIME_CURRENT
3	40.00748	116.3208	145	39904.85	04/01/2009	20:23:41
6	40.00753	116.3204	150	39904.85	04/01/2009	20:23:56
9	40.00746	116.3198	141	39904.85	04/01/2009	20:24:11
12	40.00768	116.3197	153	39904.85	04/01/2009	20:24:26
15	40.00768	116.3197	155	39904.85	04/01/2009	20:24:39
18	40.00771	116.3197	154	39904.85	04/01/2009	20:24:56
21	40.00755	116.3198	152	39904.85	04/01/2009	20:25:11
24	40.00765	116.3196	147	39904.85	04/01/2009	20:25:26

Table 4.4: Third sampling

DATA_ID	LATITUDE	LONGITUDE	ALTITUDE	DATE_PASED	DATE_CURRENT	TIME_CURRENT
5	40.00751	116.3205	141	39904.85	04/01/2009	20:23:51
10	40.00757	116.3198	138	39904.85	04/01/2009	20:24:16
15	40.00768	116.3197	155	39904.85	04/01/2009	20:24:39
20	40.00763	116.3198	152	39904.85	04/01/2009	20:25:06
25	40.00766	116.3195	146	39904.85	04/01/2009	20:25:31

Table 4.5: Forth sampling

DATA_ID	LATITUDE	LONGITUDE	ALTITUDE	DATE_PASED	DATE_CURRENT	TIME_CURRENT
10	40.00757	116.3198	138	39904.85	04/01/2009	20:24:16
20	40.00763	116.3198	152	39904.85	04/01/2009	20:25:06

Table 4.6: Fifth sampling

DATA_ID	LATITUDE	LONGITUDE	ALTITUDE	DATE_PASED	DATE_CURRENT	TIME_CURRENT
15	40.00768	116.3197	155	39904.85	04/01/2009	20:24:39

Table 4.7: Sixth sample

DATA_ID	LATITUDE	LONGITUDE	ALTITUDE	DATE_PASED	DATE_CURRENT	TIME_CURRENT
20	40.00763	116.3198	152	39904.85	04/01/2009	20:25:06

Table 4.8: Seventh sample

DATA_ID	LATITUDE	LONGITUDE	ALTITUDE	DATE_PASED	DATE_CURRENT	TIME_CURRENT
20	40.00763	116.3198	152	39904.85	04/01/2009	20:25:06

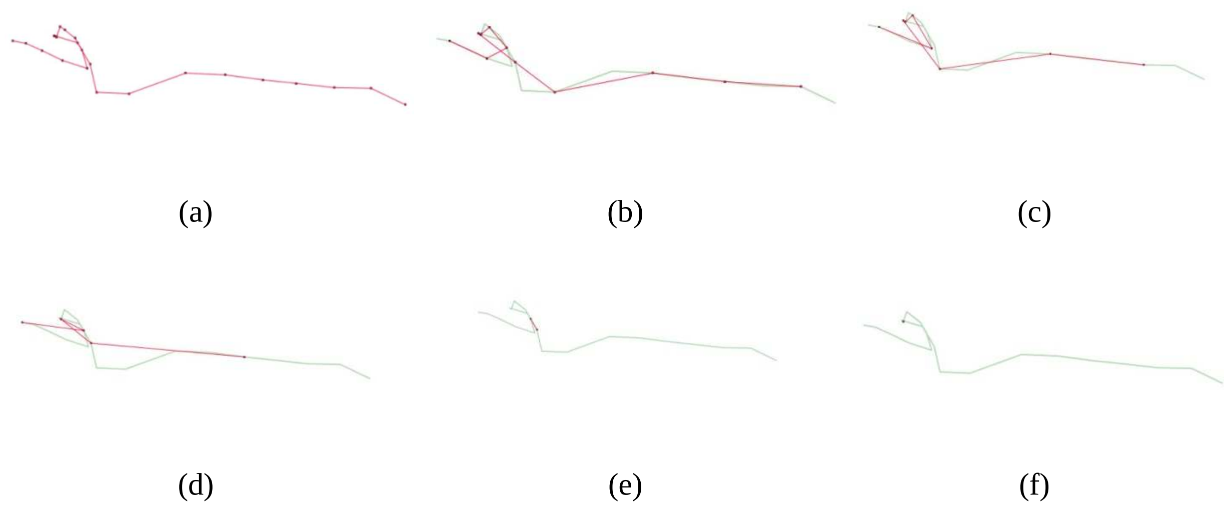

Figure 4.3: (a) Complete trajectory. (b) Sampled route with rate ½. (c) Sampled route with rate 1/3. (d) Sampled route with rate1/4. (e) Sampled route with rate1/10. (f) Sampled route with rates1/15,1/20,1/30.

At first we applied the sample based approach to datasets to find out whether they are of good quality and suitable for use or not. This is essentially the meaning of fitness for use quality measure. According to this measure a dataset that is of good quality for a particular user may be not so good for another one. Actually this measure is so important that over the years many methods have been proposed to somehow quantize it [12]. Therefore we decided to check whether this database is good for this purpose or not.

In order to specify whether data available in GeoNames is of good quality, we made a dataset including about 120 records and then a 20 record uniformly distributed random sample was taken. We chose this size so that at least two errors appear in the sample. Then we measured positional accuracy according to internal quality approach, or by comparing longitude and latitude in our samples to a reference dataset captured from many other credible sources including Wikipedia and Google Earth. For the cities if the fault was under 0.01 we considered the record as acceptable. For the other points the error must be under 0.001 in order to be viable.

The threshold for accepting or rejecting a sample is very important and could be subject of further research. In this work we considered 90% and 10% to accept or reject a sample. Figure 4.4 demonstrates the results for the first sample.

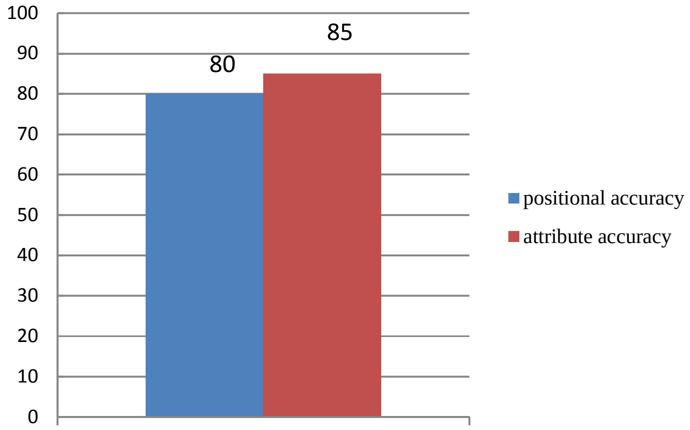

Figure 4.4 : Results of the first round of quality assessment

Because of our standard we needed a second sample. In the second round of our experiment, we took a 30 record random sample and measured two aforementioned quality factors. The second sample is also a uniformly distributed random sample.

The second sample results are depicted in Figure 4.5 that shows the results of second round of quality assessment. The method of measuring accuracy of positional data is the same for first and second round and is based on comparison with the reference dataset acquired from credible databases.

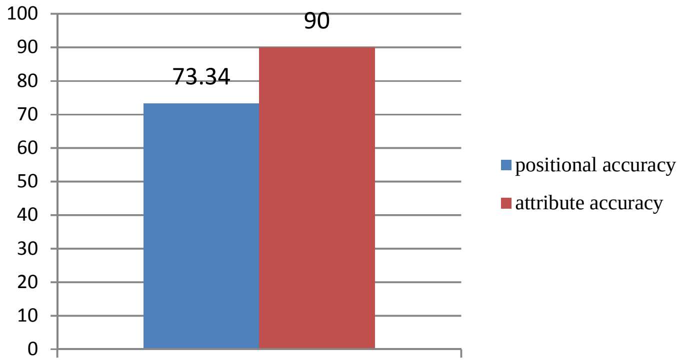

Figure 4.5: Results of the second round of quality assessment

One of the important quality metrics of spatial data is updateness. In GeoNames database date of last data update is available. Users know how old data are and they can choose to use them or not. It's predictable that there is not any pattern in updateness of data. This is understandable because GeoNames provides possibility of entering or editing data for the users.
Suppose that a user wants longitude and latitude of cities and hotels in both Iran and France. In GeoNames this data is stored in two separate tables. This information product can be derived from applying Union operation on two tables. Hence using sample based approach we can present an estimate of data quality in information product in an efficient way.
From previous section we measured positional accuracy of data of Iran to be about 73.34%. In a similar way positional accuracy of data of France is 86.6. So according to eq. (2) positional accuracy of information products can be calculated:

$$(30 * 73.34 + 30 * 86.6) / (30 + 30) = 79.9\%$$

So, a user can decide if quality of data of every needed information product is enough or not.
After assessing the quality of our data sources now we can examine quality of operations on sampled data and compare it with operation on complete data. In the other word we want to know that if we operate on sampled data instead of complete data, like a sampled route instead of complete route, our spatial results how much error will have. For instance if we measure the distance between a particular point and a complete route, and then measure the same quantity but with the sampled route in different sampling rates, how different the results will be. Spatial operation in this study is finding distance. First we calculated the distance between a complete trajectory and its nearest neighbor. The database environment for this job was PostGIS 1.5 and our spatial data visualization is done by OpenJUMP software. You can find the necessary SQL code for constructing the database and necessary conversions in the appendix A. Then the same procedure was applied to the trajectories and we loaded 36 trajectories to a database. The code for this job is available in Appendix A. In the next step because each trajectory was a sequence of spatial points, we had to convert them to a spatial route. So, at this point our data is ready for performing our experiments. The process of finding the trajectories' nearest neighbor and the distance between this two, forms the main part of this study. This task was done using PostGIS spatial operations that you can find this program in appendix A.
The final level of our study includes sampling trajectories with different sampling rates, also converting these samples to routes and ultimately finding the distance between sampled routes

and their nearest neighbor. All of the programs that we used in this study is available in appendix A. It's obvious that by sampling the trajectories the quality or in particular accuracy of calculated distance between these trajectories and their nearest neighbor is affected or to be precise reduced. Figure 4.6 shows this important issue by showing a complete trajectory and its nearest neighbor and also the same trajectory in sampled form and the same nearest neighbor. Process of finding nearest neighbor was performed using PostGIS spatial function ST_Distance(), which returns spheroid minimum distance between two geographic points in meters. For maintaining integrity, all of the lengths in this study were measured using spheroid system so that our results were more realistic. Also all of our measurements are in the metric and we converted all of these results to kilometers so we can compare them. Though, all of the numbers were in a very good precision and we measured the lengths with 12 decimal points.

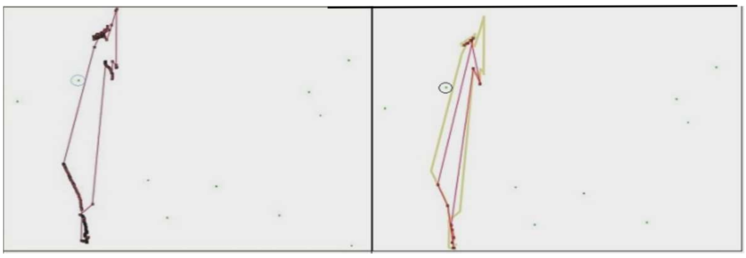

Figure 4.6 : A complete trajectory and its nearest neighbor, (a) original trajectory, (b) sampled trajectory

By conducting the previously explained steps, we observed the fallowing results. First it's obvious that with increase in the sampling rate, error in the calculated distance increases too. Figure 4.7 that shows mean error confirms this.

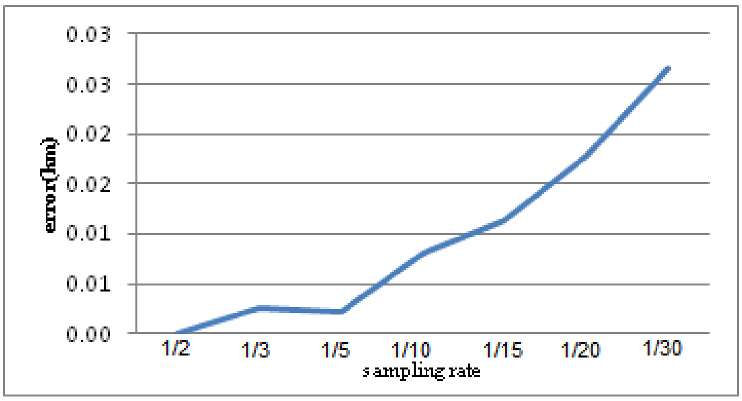

Figure 4.7: Mean error for all trajectories

Figure 4.8 shows the mean error for all of the sampling rates. You can see that error average is changing in a narrower domain for larger trajectories. This means that you may apply sampling on larger trajectories and get more accurate results. Because if the trajectories are too short with a little change in the length of route, you may get a huge change in the error size.

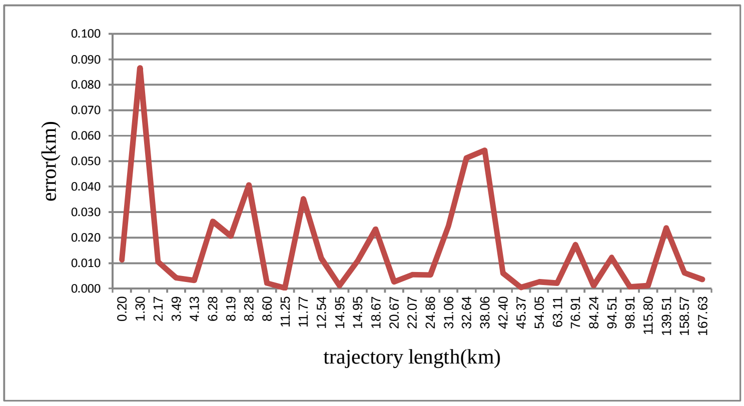

Figure 4.8: Mean error

Also from figure 4.9, you can see that for longer trajectories the ratio of error size to the trajectory length is lower, so for very long trajectories you can neglect the sampling error.

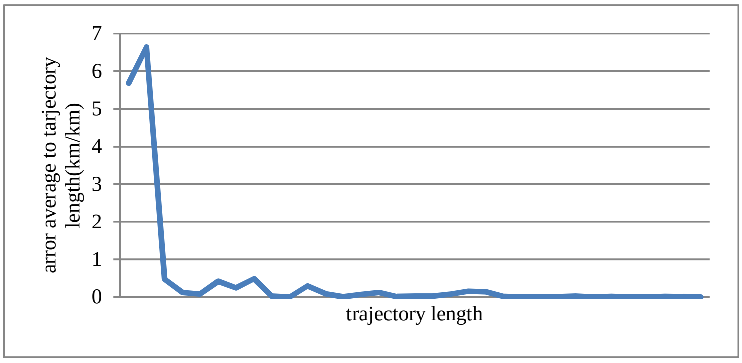

Figure 4.9: Mean error to trajectory length

About the size of the sampling error it's more complicated. it can be seen from figure 4.10 that for smaller trajectories the absolute size of error increases very faster and therefore for the small trajectories with low point number it's not very reasonable to apply the sample based approach. But for longer ones, based on our results this method can be used.

For the other quality factors it is obvious that completeness is very damaged by sampling. Assuming that we have a complete dataset as our source, by sampling we actually are deleting information. So our result database may not be so complete.

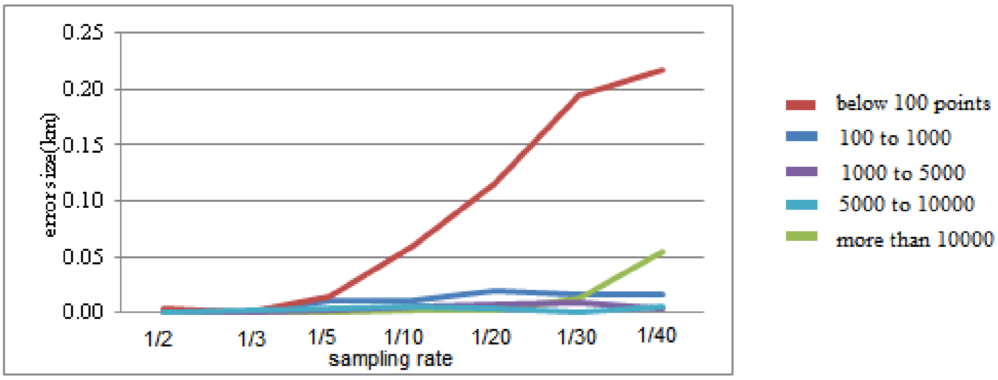

Figure 4.10: Error size for trajectories

In the time scope the results are interesting. Figure 4.11 shows the execution time of finding nearest neighbor operation for larger trajectories and figure 4.12 for smaller ones. In both figures each line indicates the execution time for a trajectory and the legend shows the number of spatial points in it. It seems that for all trajectories, large or small, execution time of finding nearest neighbor is proportional to size of the sample. If we consider the tradeoff between time gain and quality loss it seems that with large and medium size the sampling method could yield better results. Because when we decrease the size of the sample for larger trajectories regarding the huge gain in execution time, the quality loss is reasonable.

Also in figures 4.13 and 4.14 you can see the effect of sample size on execution time for different trajectory sizes. These demonstrations confirm the mentioned results too.

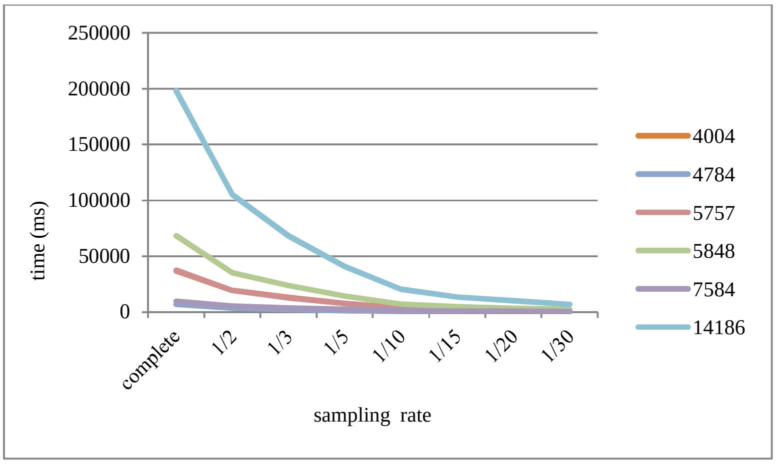

Figure 4.11: Execution time of finding nearest neighbor for large trajectories with more than 3000 points

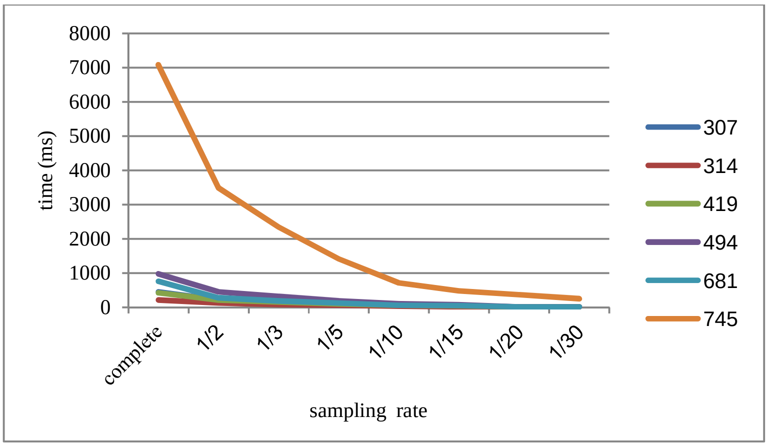

Figure 4.12: Execution time of finding nearest neighbor for small trajectories with less than 1000 points

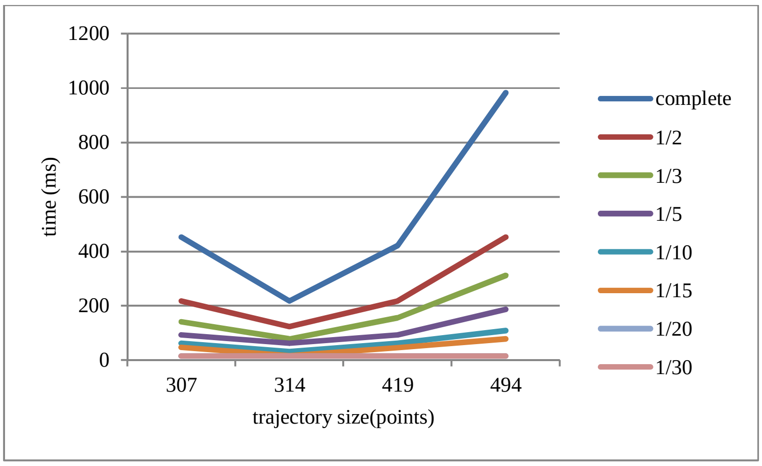

Figure 4.13: Execution time vs. trajectory size

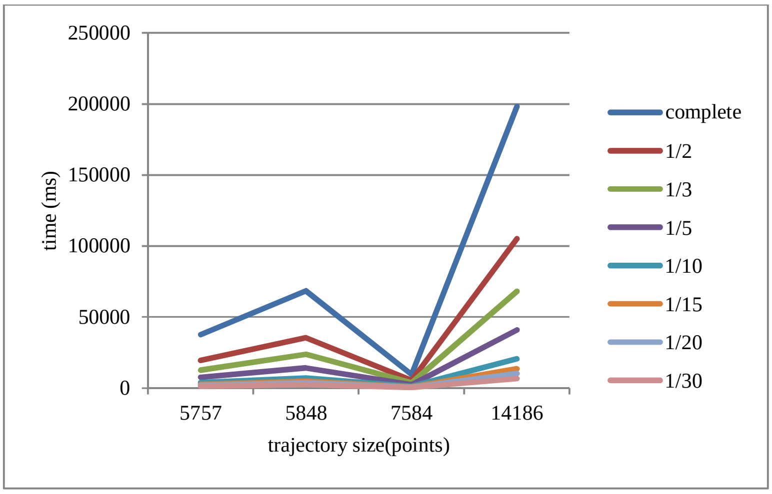

Figure 4.14: Execution time vs. trajectory size

5. Open Issues

Though there is a tremendous amount of work on spatial data quality [3] [7] [8], but still many unsolved problems and research possibilities can be found in the context of practical methods for modeling spatial data quality. Also implementing other data quality assessment methods on spatial databases can be an interesting subject for research.

Applying sample based method to the spatial data, especially in Iran can be subject of much more research work as mentioned in earlier sections.

Formalizing an efficient sampling method, applying sample based data quality assessment to the spatial data integration of heterogeneous data sources and customizing these methods for spatial data from Iran with multiple aspects and application areas such as natural resources, agriculture and urban design are some of the interesting research areas.

6. Conclusion

The importance of spatial data quality is under-estimated especially in Iran and considering the volume of decisions that are made using this type of information, the research work in this area is essential for credibility of such decisions. Hence we decided to apply one of the most efficient methods for assessing data quality to available spatial data from china and tried to quantify the quality and evaluating the accuracy of such data.

Sample based data quality assessment is an applicable and trustworthy method for estimating quality of spatial data and can be customized and used for deciding whether data available in global or local spatial databases are suitable for using under various circumstances or not.

Also because of huge volume of spatial data, many spatial operations on them would be very slow and time consuming. In this study we showed that instead of using all of data, we can use samples such that quality of our operation in many cases is acceptable.

In this approach samples were collected in a static or periodic manner and because of this unique property, if you are working in a distributed or network environment, the efficiency privilege is considerable and therefore this method is applicable in data quality assessment as well.

References

[1] N. Ghadiri, "Spatial Databases: Overview of Technologies,Techniques and Trends", October 2006.

[2] R. Devillers and R. Jeansoulin, Fundamentals of Spatial Data Quality, London: ISTE&John Wiley, 2006.

[3] H. Xie, X. H. Tong and Z. Q. Jiang, "The Quality Assessment And Sampling Model For The Geological Spatial Data In China", *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2008.

[4] R. Devillers, Y. Bedard, R. Jeansoulin and B. Moulin, "Towards Spatial Data Quality Information Analysis Tools For Experts Assessing The Fitness For Use Of Spatial Data", *International Journal of Geographical Information Science*, vol. 21, no. 3, pp. 261-282, 2007.

[5] D. P. Ballou, I. N. Chengalur-Smith and R. Y. Wang, "Sample-Based Quality Estimation of Query Results in Relational Database Environments", *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 5, pp. 639-650, 2006.

[6] S. Dustdar, P. Pichler, V. Savenkov and H. Truonh, "Quality-aware Service-Oriented Data Integration:Requirements, State of the Art and Open Challenges", *SIGMOD*, vol. 41, 2012.

[7] B. K. Boateng and I. Yakubu, "Assessing the Quality of Spatial Data", *European Journal of Scientific Research*, vol. 43, no. 4, pp. 507-515, 2010.

[8] R. Devillers, A. Stein, N. Chrisman, P. Fisher and W. Shi, "Thirty Years of Research on

Spatial Data Quality: Achievements, Failures,and Apportunities", *Transactions in GIS,* vol. 14, no. 4, pp. 387-400, 2010.

[9] R. Devillers, A. Stein, Y. Bédard, N. Chrisman, P. Fisher, "Spatial Data Quality: From Metadata To Quality Indicators And Contextual End-User Manual", *OEEEPE/ISPRS Workshop on spatial Data Quality Management*, 2002.

[10] G.J. Hunter, S. H., Z. Sadiq, A. Boin, "Next-Generation Research Issues In Spatial DataQuality", *Proceedings of SSC 2005 Spatial Intelligence*, 2005.

[11] H. Tastan, M. Altan, "Spatial Data Quality", *Proceedings of the Third Turkish-German Joint Geodetic Days, 1999*.

[12] G. Hunter, A. Agumya, "A Risk-Based Approach to Assessing the 'Fitness for Use' of Spatial Data", *URISA*, 1999.

[13] M. Duckham, J. Drummond, D. Forrest, "Spatial Data Quality Capture Through Inductive Learning", *Spatial Cognition and Computation,* Volume 2, Issue 4, pp 261-282, 2000.

[14] M. Mostafavi, G. Edwards, R. Jeansoulin, "An ontology-based method for quality assessment of spatial databases", *Third International Symposium on Spatial Data Quality*, Vol 28, pp. 49-66, 2004.

[15] H. J. Onsrud, "Liability for Spatial Data Quality", 2009. 18

[16] M. Goodchild, " Beyond Metadata: Towards User-Centric Description Of Data Quality", *ISSDQ,* 2007,

[17] A. Borek, P. Woodall, M. Oberhofer, "A Classification Of Data Quality Assessment Methods", Proceedings Of 16th International conference on information quality, 2012.

[18] A. Zargar, R. Devillers, " An Operation-based Communication of Spatial Data Quality", Advanced Geographic Information Systems & Web Services. GEOWS '09. International Conference on , 2009.

[19] J. Vizhi, T. Bhuvaneswari, " Data Quality Measurement On Categorical Data Using Genetic Algorithm", International Journal of Data Mining & Knowledge Management Process, vol. 2, p.p 33-42, 2012.

[20] R.M. Henriksson, T. Kauppinen, " An Ontology-Driven Approach For Spatial Data Quality Evaluation", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 34, 2007.

[21] A. Boin, G. Hunter, " What Communicates Quality To The Spatial Data Consumer?", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 34, 2007.

[22] G. Hunter, " New Tools For Handling Spatial Data Quality: Moving from Academic Concepts to Practical Reality", *URISA*, vol. 11,p.p 25-34, 2001.

Appendix A.

a- SQL Code for loading Geonames Database:

```
create schema ch01;
create table ch01.tbl_data_china(
        geo_name_id varchar(10) primary key,
        geo_name varchar(200),
        ascii_name varchar(200),
        alter_name varchar(5000),
        latitude double precision,
        longitude double precision,
        feature_class char(1),
        feature_code varchar(10),
        countery_code char(2),
        cc2 varchar(60),
        admin1_code varchar(20),
        admin2_code varchar(80),
        admin3_code varchar(20),
        admin4_code varchar(20),
        population bigint,
        elevation varchar(10),
        dem integer,
        time_zone varchar(40),
        mod_date date);
copy ch01.tbl_data_china from 'C:\Program Files\PostgreSQL\9.1\data\CN.txt' delimiter '';
alter table ch01.tbl_data_china drop column geo_name_id;
alter table ch01.tbl_data_china add column data_id SERIAL primary key;
alter table ch01.tbl_data_china add column geo geography(point,4326);
update ch01.tbl_data_china set geo = ST_GeogFromText('POINT('||longitude||'
'||latitude||')');
```

```
create index china_data_idx on ch01.tbl_data_china using gist(geo);
```

b- SQL Code for converting trajectories to routes

```
create table trajectory.tbl_routes(
        ID varchar(14),
        route geography
);
insert into trajectory.tbl_routes select 20081103232153,st_makeline(array(select
geo::geometry from  trajectory.tbl_20081103232153))::geography;
```

c- SQL Code for sampling , finding the nearest neighbor and the distance

```
create table samples_20.tbl_20090403011657(
        latitude double precision,
        longitude double precision,
        altitude integer,
        date_pased varchar(30),
        date_current varchar(10),
        time_current varchar(8)
        );
alter table samples_20.tbl_20090403011657 add column data_id SERIAL primary key;
alter table samples_20.tbl_20090403011657 add column geo geography(point,4326);

insert into samples_20.tbl_20090403011657
 SELECT * FROM trajectory.tbl_data_20090403011657 WHERE
trajectory.tbl_data_20090403011657.data_id % 3=0;;

insert into samples_20.tbl_routes select 20090403011657,st_makeline(array(select
geo::geometry from samples_20.tbl_20090403011657))::geography;
```

```sql
SELECT DISTINCT ON (ST_Distance(ch01.tbl_data_china.geo,samples_20.tbl_routes.route))
        ch01.tbl_data_china.data_id,
        ch01.tbl_data_china.geo_name,
        ST_Distance(ch01.tbl_data_china.geo,samples_20.tbl_routes.route)/1000 As dist_km
FROM
        ch01.tbl_data_china
        INNER JOIN
        samples_20.tbl_routes
        ON ST_DWithin(ch01.tbl_data_china.geo, samples_20.tbl_routes.route,5000)
WHERE
        samples_20.tbl_routes.id='20090403011657'
ORDER BY
        ST_Distance(ch01.tbl_data_china.geo,samples_20.tbl_routes.route) asc,
        ch01.tbl_data_china.geo_name asc

limit 10;
```