Epidemic Prevention and Control Based On GeoHash

Youwei Huang (Project Manager)
huangyw@iict.ac.cn
Institute of Intelligent Computing Technology, Chinese Academy of Sciences
Suzhou, Jiangsu, China

Feng Lu (Algorithm Engineer)
lufeng20g@ict.ac.cn
Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China

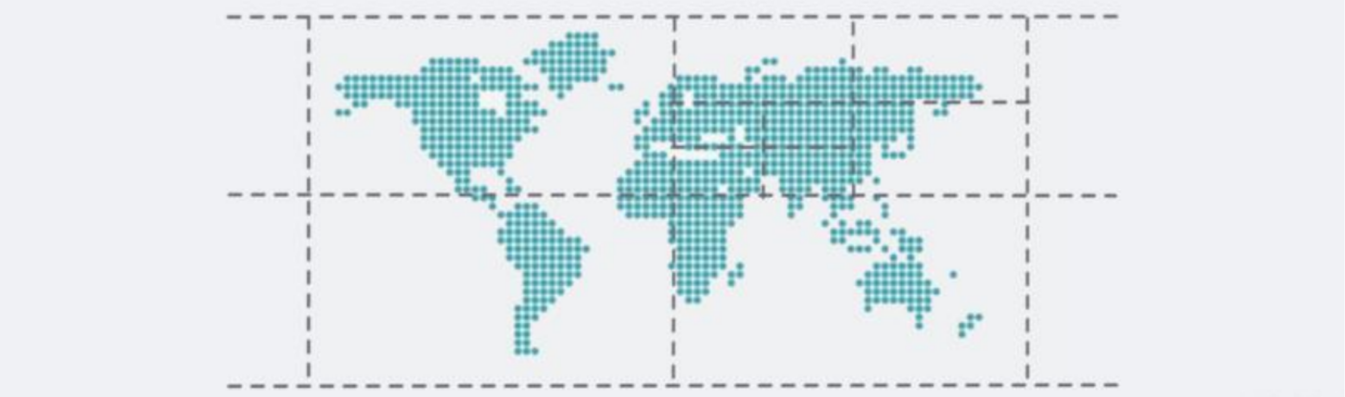


Figure 1. GeoHash for geographic grid division

Abstract

COVID-19 (Coronavirus Disease 2019) which is a contagious disease caused by SARS-CoV-2[**?**] was first detected in Dec. 2019. Until 2021, this virus is still spreading around the world. Before the vaccine is widely vaccinated or the invention of specific medication, many measures have been taken by people to prevent the spread of the epidemic. In a special period, we have to quarantine the high-risk groups and lock down seriously infected regions. Here we propose a kind of dynamic block division technology based on **GeoHash** used to monitor, screen and control the epidemic areas. **GeoHash** is a public domain geocode system invented in 2008 by Gustavo Niemeyer[**?**]. We divide a map into several blocks and use **GeoHash** to encode the information of each block. Through the **GIS**, **GeoHash** can be easily decoded to original visual blocks on a digital map. A map generated by **GIS** is used for epidemic prevention and control, so it can be named **"Epidemic Map"**. Each block on such **Epidemic Map** contains the safety information and other important characteristics which are concerned by medical work. These dynamic blocks on the map can be scaled and represented as various regular geometric shapes. The vital information and the results of quantitative analysis of the data on each block support for decision-making, measures formulation, and effectiveness assessment of COVID-19 prevention and control. Such a kind of geographic information system can be used not only for preventing and controlling COVID-19 pandemic, but also be applicative in instances of other epidemic diseases.

Keywords: GeoHash, COVID-19, GIS, big data, epidemic prevention and control

ACM Reference Format:
Youwei Huang (Project Manager) and Feng Lu (Algorithm Engineer). 2021. Epidemic Prevention and Control Based On GeoHash. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, **??** pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

During the COVID-19 pandemic, many cities over the world were forced to lock down. Wuhan City and the major cities in Hubei, China were put under lockdown on the 23rd and 24th of January, respectively[**?**]. Lockdown meant the whole region was quarantined and cut off physical contact with the outside world. The citizens were forbidden to leave their city or even their home. The national medical team carried out centralized medical observation and treatment in quarantined cities. Research shows, COVID-19 spread became weaker following lockdown[**?**]. However the lockdown of a

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

city can cause huge economic losses. The lockdown of some vital areas can cause irreparable losses, such as financial center, political center, and industrial dependent cities. Another issue that confuses people is how to distinguish whether the area they are in or where they are going is safe. The current regional risk warning or lockdown is based on the administrative divisions as figure 2. Cities, states or provinces all over the world have different sizes and irregular geographic borders. There may be an outbreak in a city, but it does not mean that it spreads to all corners of the city. On the contrary, there may be no epidemic in the center of neighboring cities, but there may already be a huge risk at the border with these surrounding cities. Figure 2 is a map that shows the initial locked down cities in Wuhan province, China, a white block surrounded by red blocks is dangerous, even if it is not locked down.

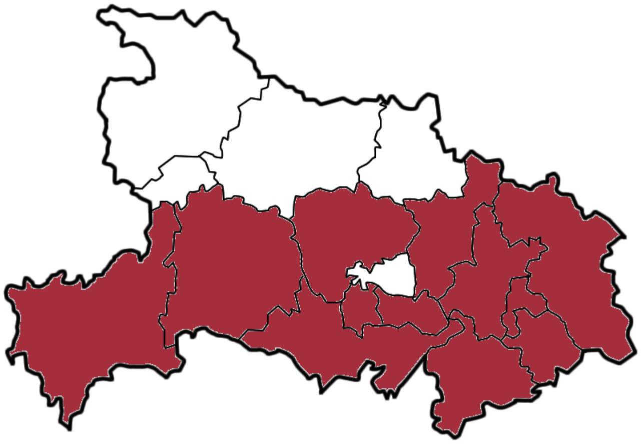


Figure 2. Map of locked down administrative divisions of Hubei. [Public domain], via Wikipedia. (https://en.wikipedia.org/wiki/COVID-19_lockdown_in_Hubei).

We summarize the main weaknesses by using the current dividing measures:

1. The border of a city is irregular and the transmission of virus doesn't follow the administrative division, so the prevention and lockdown can be not accurate.
2. The administrative size of a city is fixed, but the disease is spreadable, so the region of lockdown can not be expanded flexibly.
3. Due to regional differences, the information released in each region is not complete and not uniform.
4. The news released by the local government can be lagging and users cannot get it in real time.

Based on the above statements, it is not the best way to observe and control the epidemic area through the administrative division. For infectious diseases, we have abandoned the common administrative methods. And we have adopted a technology based on GIS (Geographic Information System)[**?**]. **GeoHash** is used to divide the map into several geometric blocks. The 2-dimensional geometric blocks on the map are encoded by GeoHash, they are reduced to 1-dimension and stored as string in any databases. These blocks are presented as regular geometry, but can be scaled according to the needs of different observation scope.

Meanwhile, we do this technology is because it has the following advantages when controlling epidemic situation:

1. The blocks divided by GeoHash are regular geometry, and the shape can be customized by observer.
2. The blocks are generated dynamically and they are scalable according to the scope of infection.
3. By combining with GIS, information about the epidemic situation can be encoded in GeoHash or directly saved.
4. Block data can be easily quantified, as example of generating safety index.
5. When such a GIS is released to the Internet, users get epidemic information in real time.

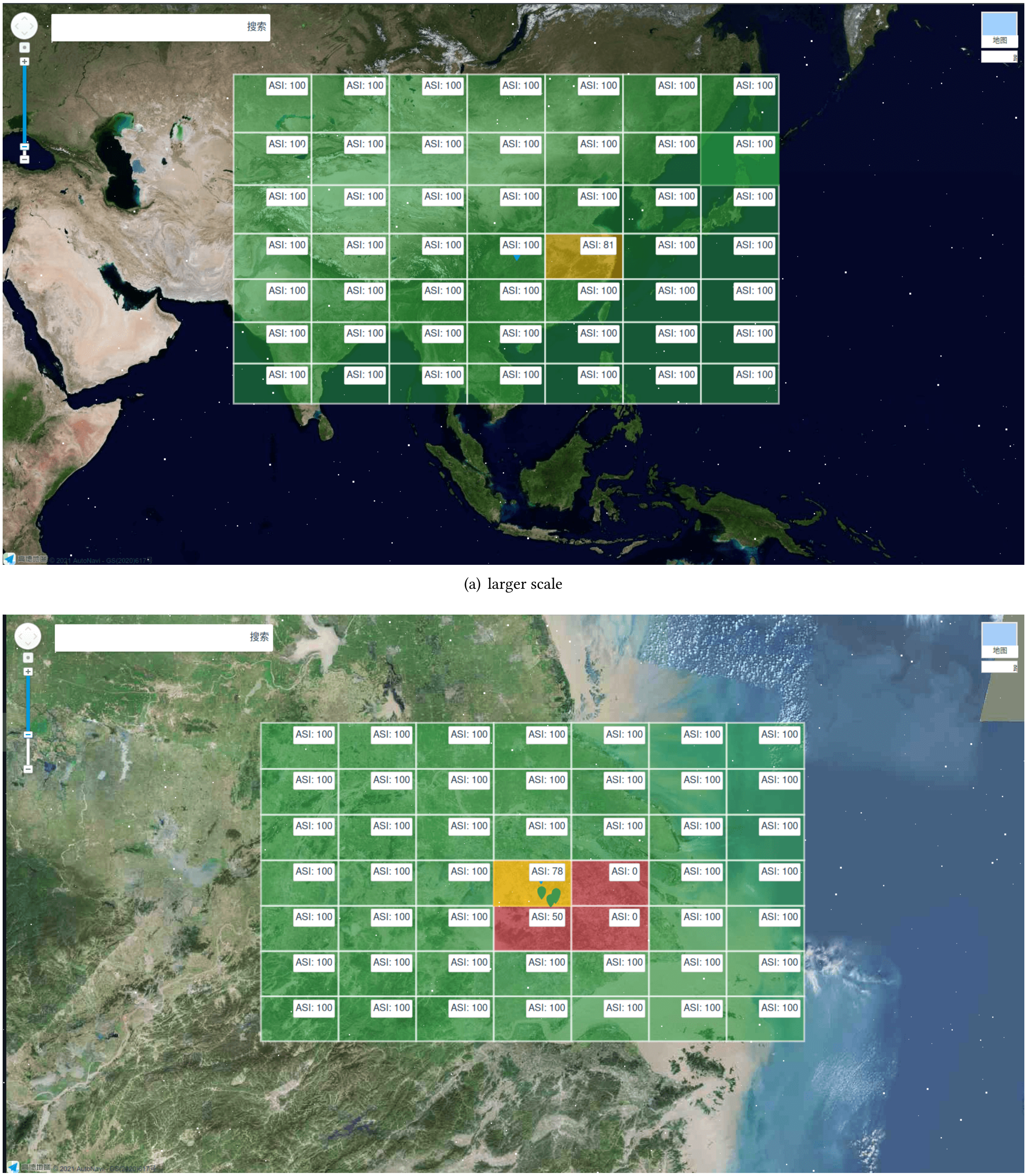
In the practical and experimental scenarios, we use mobile application and web technology to develop such a particular GIS for medical prevention and treatment as shown in Figure 3. It works for medical workers as a visual auxiliary tool and share the results of epidemic data analysis. The **ASI** in Figure 3 is a value of "Area Safety Index". **ASI** represents the risk level of a region. The system contributes to enlighten and support decisions of governments, medical institutions, users, and other researchers who are doing the similar research with us.

The core idea of this technology is dividing the earth to dynamic blocks. A block has the following characteristics:

1. Blocks are regular geometry connected with each other. It can be understood as cellular grids.
2. Blocks can be scaled on the map.
3. Blocks are created only when they are meaningful.
4. Scaling is limited, with the smallest and largest block size.
5. Blocks scaling levels are discrete sizes, not continuous.
6. A block stores structured data, which can be used for computing and analysing various attributes.

The last item in the above list indicates that the purpose of dividing by blocks is to perform quantitative analysis related to the epidemic situation.

We will also introduce other related work in controlling epidemic by using similar information technology and computer visualization technology. Then we will focus on how we use **GeoHash** to divide blocks on the map, and explain the methods of quantifing epidemic data to provide area risk warning. Finally, by using our GIS example, we will give our experimental and test results, and summarize the conclusion.



ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100
ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100
ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100
ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 81	ASI: 100	ASI: 100
ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100
ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100
ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100

(a) larger scale

ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100
ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100
ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100
ASI: 100	ASI: 100	ASI: 100	ASI: 78	ASI: 0	ASI: 100	ASI: 100
ASI: 100	ASI: 100	ASI: 100	ASI: 50	ASI: 0	ASI: 100	ASI: 100
ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100
ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100	ASI: 100

(b) smaller scale

Figure 3. Dynamic blocks with ASI in geographic information system

2 Related Work

There are some mature cases of controlling and treating COVID-19 pandemic by using information technology and Internet data. Many studies on COVID-19 have recently emerged, and various data science applications combating the pandemic have been reported recently[?]. The main functions of these systems or softwares are listed as follows:[?]

- Tracking of people's movements.
- Early warning of high-risk areas.
- Screening of asymptomatic potential infections.
- Drug development.
- Information release and policy support.

2.1 Data Visualization Analysis

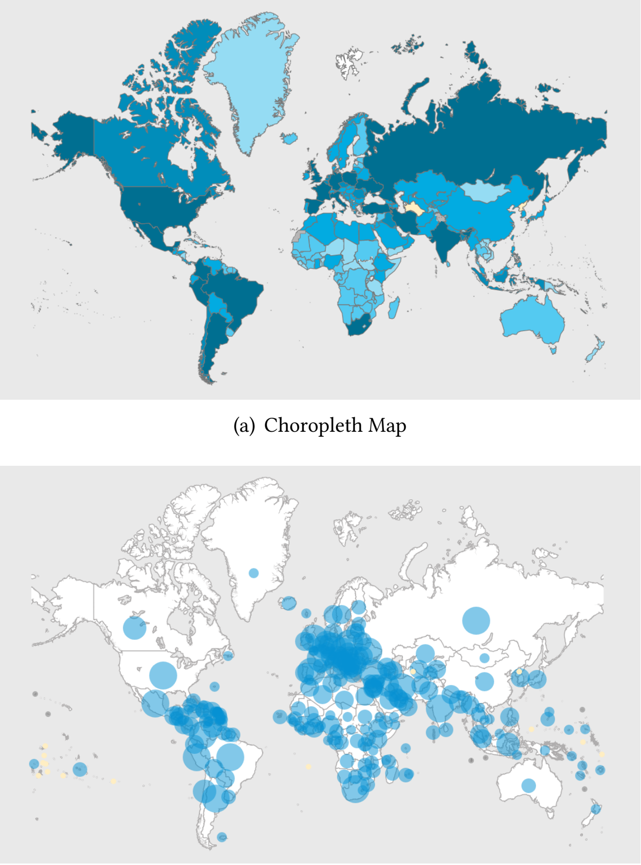
The computer can visualize all kinds of structured data and convey the visual information to users. Visualization technology presents data to users by drawing charts and graphics, in which the data is represented by symbols, such as bar charts, line charts, pie charts, maps and etc[?].

Figure 4 shows the global COVID-19 epidemic situation in the form of map charts. The epidemic maps are updated by WHO (World Health Organization)[1] in real time, to display the number of cases around the world.

Maps in Figure 4 uses two styles of presentation: choropleth and bubble. One uses the depth of color to show the severity of the epidemic situation in each country, and another uses the bubble size to show the number of infections. No matter what kind of map, its role is to help the local people easy to understand the severity of the epidemic, and prompt the local government to take actions to treat and control the epidemic situation. These two figures (Figure 4) are similar to the previous Figure 2, except that Figure 2 only shows the data of one province. Contents in these maps include like: confirmed cases, deaths, historical cases, added cases, and regional lockdown status. Some disadvantages of such kind of maps are discussed in the introduction section (Section 1).

The bar charts and line charts are aslo widely used in epidemic data analysis. These graphs are mostly used to show the trend of epidemic situation and transmission cycle. Figure 5 is an example that uses both bar chart and line chart to perform the trend of daily new cases from Jan. 2020 to Jan. 2021 in the US. The red line in this chart is the 7-day moving average curve. Other trends from different types of data, such as death trends, can be found on the official CDC[2] website.

The line charts and bar charts reflect historical epidemic data from time series, while map charts reflect epidemic data from spatial distribution. In other related work survey, visual data analysis are used to study the relationship between population mobility and the epidemic spreading pattern. During the early outbreak of coronavirus in Wuhan, China, the graphs in a research suggested that the number of confirmed cases in other provinces were directly proportional to the inflow of Wuhan population. The research group also used the pattern derived from the data analysis to predict the number of infections.[?]



(a) Choropleth Map

(b) Bubble Map

Figure 4. WHO coronavirus disease dashboard. [Public domain], via WHO (https://covid19.who.int)

2.2 Geographic Information System

The maps we described in the previous section are charts, the information carried by map charts is limited, unflexible, not automatic analysis and not real-time. Although charts can provide visual perception, users need to analyze the graphs by themselves, but GIS can integrate analysis, prediction and other practical functions. A GIS (geographic information system) is a conceptualized framework that provides the ability to capture and analyze spatial and geographic data[?]. Since the outbreak of the epidemic, a number of geographic information systems have been built or have added real-time epidemic related functions, such as "epidemic map displays", "fever clinic queries" and "passenger information queries".

[1] https://www.who.int

[2] https://covid.cdc.gov/covid-data-tracker

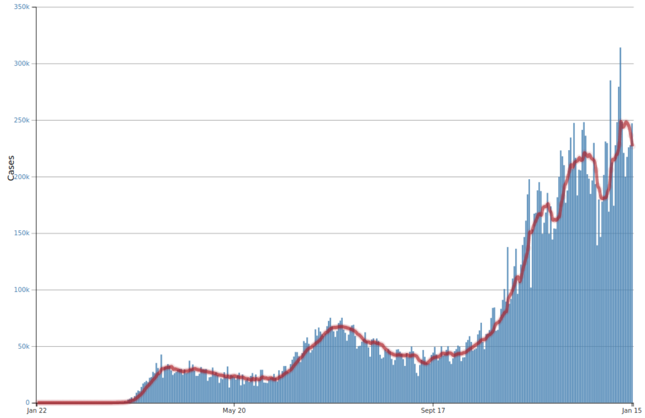
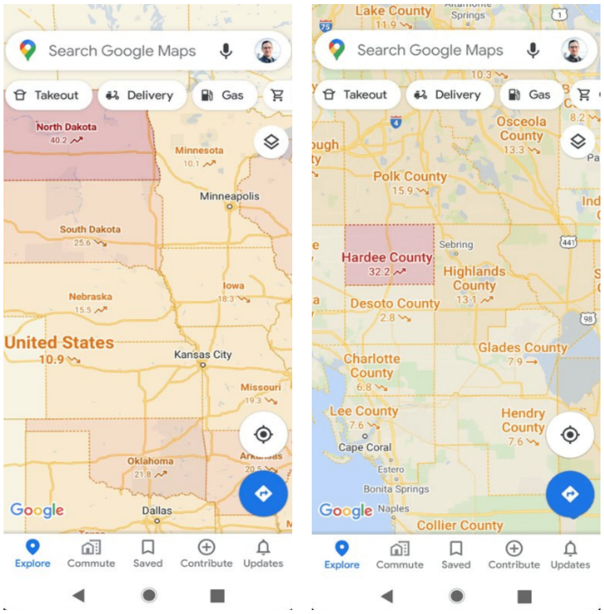


Figure 5. Daily trends in number of COVID-19 new cases in the US reported to CDC. [Public domain], via CDC (https://covid.cdc.gov/covid-data-tracker/#trends_dailytrendscases)

Based on existing commercial GIS softwares, they made important contributions to epidemic prevention and control[**?**]. Some of the systems have the function of dynamic zoom map, which can display the situations of different scaling areas, from state level to community level. For example, Figure 6 shows a map with an epidemic layer on it, it is a GIS tool that shows critical information about COVID-19 cases in an area so the users can make more informed decisions about where to go.



(a) State Level (b) County Level

Figure 6. COVID layer in Google Map of different scales

In Figure 6, **Google Map** adds a COVID-19 layer to the GIS, it also quantities an safety index by using the 7-day average for the number of new cases per 100,000 people. It also indicates whether cases are increasing or decreasing. The layer's colors indicate:[3]

- Grey: Less than 1 case
- Yellow: 1-10 cases
- Orange: 10-20 cases
- Dark orange: 20-30 cases
- Red: 30-40 cases
- Dark red: 40+ cases

Unlike **Google Map**, we perform the quantitative analysis based on dynamic GeoHash blocks instead of administrative regions. The safety quantification algorithm we use not only includes the simple infected cases. In the following chapters, I will focus on our research work about how to use GeoHash to improve the GIS in epidemic prevention and control.

3 Pre-Work

Our challenges are mainly from two aspects: data collection and data quantification. Data collection is integrated in GIS, which requires users' devices to upload the current positions of the users, and the server needs to store the position data submitted by users. Quantitative analysis of information is a more complex process and our work is to quantify the collected messy data in GeoHash blocks.

For data collection, the system can provide a mobile application for users to view the surrounding epidemic situation. At the same time, in order to obtain the surrounding epidemic safety situation, users need to upload their own GPS information. Under the privacy policy, we only collect users' GPS information, but we will not save users' personal informationa. So we don't track or monitor users' positions in the system. This method of collection is carried out anonymously, and each record is stored as a virtual and unknown identity which means only users know their own locations but other users cannot get it.

After the diagnosis, the medical workers mark the confirmed cases through their virtual identities. Other users don't know these users' real identity, but they can browse the surrounding infection. The medical workers know their patients' real identities, but they cannot track their location information without user authorization. To be simple, our solution is to use the virtual identity or encode user information to separate location information and real user information. This is an approach to keep a balance between keeping user privacy and collecting data for epidemic control.

For example, in China, a kind of QR code carried by users called **Health Code** shows one's health status but doesn't tell one's information. Each QR code corresponds to a unique user in the real world. In the real world, people use the health code to know the health status of themselves or others.

For data quantification, we need to divide the map to blocks

[3] https://support.google.com/maps/answer/9795160

first by using GeoHash. At the step of GPS data collection, we get the longitude and latitude values from the user. GeoHash will calculate which block the user belongs to. Suppose a GeoHash block as a set B_1 that includes many users, we quantify the safety index of B_1 based on confirmed cases and the total number of the user. The historical users, "visitors" in a block also need to be considered. These visitors who are diagnosed in the short term will have negative effects on a block. As discussed above (in Section 2.1), a research has proposed that the number of people infected in a region is positively correlated with the inflow of the population from high-risk areas. That means, when collecting position data, we noy only need to save user current position, but also need to save user historical location data. The general storage and calculation process is shown in Figure 7, and the detailed methods and algorithms are described in Section 4.

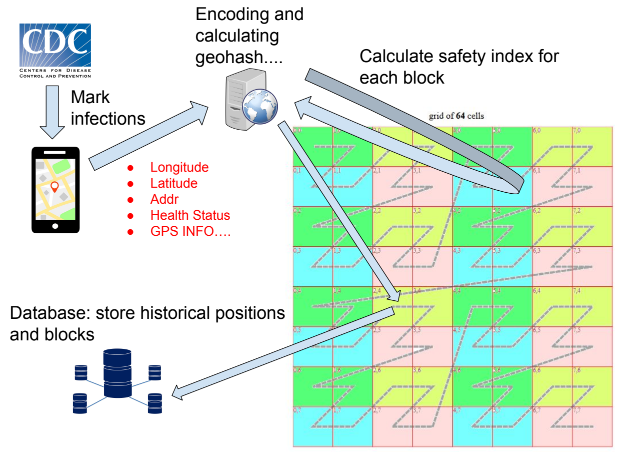

CDC
Centers for Disease Control and Prevention
Encoding and calculating geohash....
Calculate safety index for each block
Mark infections
grid of 64 cells
Longitude
Latitude
Addr
Health Status
GPS INFO....
Database: store historical positions and blocks

Figure 7. The process of data collection and quantification

4 Implementation Methods

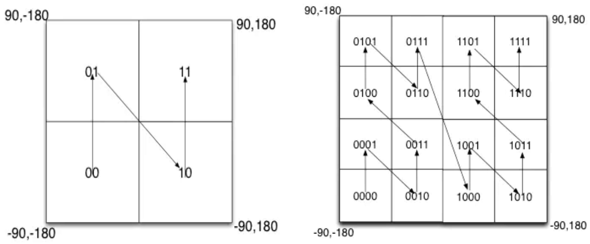
In Pre-Work section, user position information is obtained through the mobile terminal. The information contains some necessary data: userid, longitude and latitude, and **USI**[4]. Based on the above data, we can divide users into different blocks.

4.1 Grid Division

Our first step is to divide the earth into blocks or grids. GeoHash technology can help us achieve this step. GeoHash can perform the following operations:

- GeoHash divides a two-dimensional map into buckets of grid according to the range of longitude and latitude.
- GeoHash encodes a geographic location (longitude and latitude) into a string of binary codes.
- Through the incremental operation of encoded binary code, these grids are chained to a Z-order curve as Figure 8.

How does Geohash achieve the above operations?
We know that the geographical range of longitude is from -180° to 180° and latitude is from -90° to 90° [**?**]. We can divide the longitude into two intervals: [-180°, 0°], [0°, 180°]. We denote the two intervals by binary number "0" and "1" respectively. In an equivalent way, we divide the latitude into two intervals: [-90°, 0°], [0°, 90°] and denote them by "0" and "1".
Then we can use "00", "01", "10", "11" to represent the four grids. Figure 8 gives examples of "divide the earth into 4 grids" and "divide the earth into 16 grids". But in this way, the area of each grid is very large, and the grid precision is too low. The same way is used to further subdivide into 16, 64 grids, etc. We get higher precision by adding the bits of GeoHash binary which means the earth is divided into more and smaller grids.



(a) 4 Grids (b) 16 Grids

Figure 8. Z-order curve to show GeoHash grids division

4.2 Gird Size

The following Table 1 contrasts the geographical length and GeoHash bits length at around 30° latitude. Since the geometric shape of the earth is a sphere, even with the same GeoHash bit, the grids of high latitude contains less geographical length than those of low latitude. Attention: the length data in Table 1 is generated from the grids around 30° latitude, it doesn't represent all the grids or an average value.
Use the following formulas (1) and (2) to estimate the length of the two edges of a grid and it assumes the earth is completely spherical. We use G_{NS} to represent the length of north-south edge and G_{EW} to represent the length of east-west edge as they are stroked in Figure 9. The index *geobit* is the length of total GeoHash bits which is used to represent a grid. The length of a meridian which is $L_{meridian}$ in Formula (1) has been estimated at 20,003.93 km (12,429.9 miles) on a modern ellipsoid model of the earth (WGS 84)[**?**]. The length of the equator which is $L_{equator}$ in Formula (2) is about 40,075 km (24,901 miles) long[**?**]. Formula (1) is used to estimate

[4] USI is "User Safety Index", a value of user's health in our system

Table 1. Comparison table of GeoHash bits and estimated geographical length of one grid (around 30° latitude)

East-West Length(m)	South-North Length(m)	Bits
32.67	19.05	20*2
65.34	38.1	19*2
130.68	76.2	18*2
261.36	152.4	17*2
522.72	304.8	16*2
1045.44	609.6	15*2
2090.88	1219.2	14*2
4181.76	2438.4	13*2
8363.52	4876.8	12*2
16727.04	9753.6	11*2
33454.08	19507.2	10*2
66908.16	39014.4	9*2
133816.32	78028.8	8*2
267632.64	156057.6	7*2
535265.28	312115.2	6*2
1070530.56	624230.4	5*2
2141061.12	1248460.8	4*2
4282122.24	2496921.6	3*2
8564244.48	4993843.2	2*2
17128488.96	9987686.4	1*2

the length of the north-south edge of a grid. Formula (2) is used to estimate the length of the east-west edge of a grid at latitude φ.

$$G_{NS} \approx \frac{L_{meridian}}{2^{geobit/2}} \tag{1}$$

$$G_{EW}(\varphi) \approx \frac{L_{equator} \times \cos(\varphi)}{2^{geobit/2}} \tag{2}$$

The reason why the word “estimate” is used when caculating G_{NS} and G_{EW} here is that the earth is actually elliptical. Earth ellipsoid will cause the following two issues:

- The meridian and equator are different in length, grids at different latitudes own different G_{NS} values.
- If the precision of the grid is not high enough, the grid is not a rectange, two different G_{EW} values will appear in one grid.

In fact, under the same GeoHash bits, whether it is “Earth Ellipsoid” or “Earth Spheroid”, the length of the two edges of a grid depends only on latitude. If we need the accurate length of the two edges, we should start a deep discussion about “Length of a degree of longitude” and “Length of a degree of latitude”, which are beyond the research scope of this paper. Here we list our final calculation formulas when the earth is modelled by an ellipsoid, this is much more complicated but accurate than the spherical earth.

$$a = \frac{L_{equator}}{2\pi} \qquad b = \frac{L_{meridian}}{2\pi} \tag{3}$$

$$f = \frac{a-b}{a} \qquad e^2 = f(2-f) \tag{4}$$

$$\triangle\,\varphi = \frac{180}{2^{geobit/2}} \tag{5}$$

$$G_{EW}(k) = \frac{2\pi a \cos(k\Delta\varphi)}{2^{geobit/2}\sqrt{1-e^2\sin^2 k\Delta\varphi}} \tag{6}$$

$$G_{NS}(k) = a(1-e^2)\int_{k\Delta\varphi}^{(k+1)\Delta\varphi} \frac{\mathrm{d}\,\phi}{(1-e^2\sin^2\phi)^{\frac{3}{2}}} \tag{7}$$

$$k \in \{-2^{geobit/2-1}, -2^{geobit/2-1}+1, \ldots, 2^{geobit/2-1}-1\} \tag{8}$$

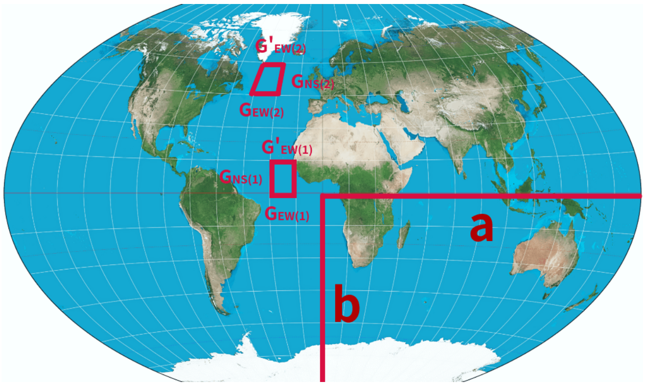


Figure 9. Grid of elliptical earth

Formula (3), (4), (5), (6), (7) and (8), give the methods to estimate the grid size of eplliptical earth. They can be used to calculate the length of each side of a grid. The formulas are derived from *The Mercator Projections*[?]. Any GeoHash grid in our GIS has 3 different side lengths, which are represented by $G_{NS}(k)$, $G_{EW}(k)$ and $G_{EW}(k+1)$ respectively.

4.3 Block Storage

We store all these blocks and user data in common databases on the server. In the real engineering environment, data storage includes two parts:

1. user geographic location data
2. block data

One block contains many users, and one user has many position logs. Block and user information are bidirectional associated, which is called “many to many” in the field of computer storage. This section will discuss the stored data models and algorithms.

4.3.1 Position to GeoHash Bit. Before storing the block, we have to first encode the position by GeoHash. Referring to the discussion about grid division in Section 4.1, we adopt the following algorithm 1.

Algorithm 1 Transform position to GeoHash bit

```
    latitude, longitude, bit geohash
```

4.3.2 GeoHash Bit to Hex.

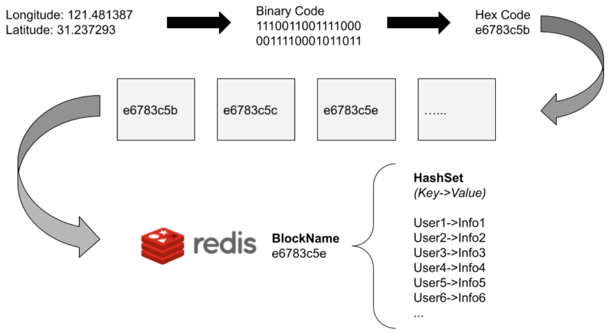


Figure 10. The process of block storage

4.3.3 Block Storage.

4.4 Grid Quantification

5 Test

6 Conclusions