Abstract: Population census is a major activity to collect population data in China. This information has provided reliable information for the scientific formulation of mid- and long-term programs for national economic and social development, for the integrated arrangement for the material and cultural life of the population, and for the coordinated development of population, economy and resources and environment. Population GIS in national and provincial level have been established based on the population data from the fifth Population Census in 2000 and geographic data from the National Fundamental GIS. From the practice of the systems establishment, the data uncertainty in population GIS is studied and summarized. The uncertainty mainly includes uncertainty of spatial data, uncertainty of population data and uncertainty from the data processing and analysis.

1 INTRODUCTION

Data is a basic and important part of Geographic Information System (GIS) and some studies show that the cost for data collection is about 80% of the total cost in the establishment of GIS. The quality of the data in GIS directly affect its reliability. Because of the complexity of the real world and the limitation of representation ability of human being, uncertainty exists in the Geographic Information System. Uncertainty is a differentiation that the real world or entity has and is represented as no accuracy, randomicity and illegibility. The uncertainty can be seen generally as an extension of error. The source of the uncertainty is from the instability of the real world on one hand, or digitization, edit and representation of geographic information on the other hand.

Population GIS includes large amount of spatial and population data. In the establishment and application of the system, the steps, such as data capture, edit, overlay and buffer analysis must be done. Randomicity, complexity, illegibility and no integrity of data and its concept can be the sources of uncertainty in the Population GIS. Uncertainty in Population GIS can be divided into three groups, which are uncertainty of spatial data, uncertainty of population data and uncertainty induced in the data processing and analysis.

2 UNCERTAINTY OF SPATIAL DATA
Uncertainty of spatial data includes uncertainty of spatial data framework, boundary of enumeration district in the population census, integration of multi-source and multi-scale geographic data, virtual enumeration district.

2.1 Uncertainty of Spatial Data
The spatial data framework of Population GIS is from the Topographic Database of National Fundamental Geographic Information System of China, which include 1:1,000,000, 1:250,000 and 1:50,000 scale databases. These databases include layers of hydrography, boundary, transportation, residence, topography and vegetation cover. In the establishment of these databases, uncertainty is from the source of data collection and process of data capture.

Uncertainty of data source is from the survey process to make the map, compilation, storage condition, currency. Distortion of paper map is greater than that of film map. Because the update of maps is a time-consuming process, the features in the map are not all represent the up-to-date changes. In the map digitization, the manual positioning or scanning, editing such as edge matching, data processing such as geometry collection, coordinate transformation, graphic generalization, input of attribute data, etc. would induce uncertainty. Figure 1 shows edge matching process of the maps.

2.2 Uncertainty of Boundary of Enumeration District in the Population Census
Geographic districts include six levels, which are province, prefecture, county, township, enumeration district and enumeration block. The boundary of county and above level is the same as the administrative boundary respectively, while township, enumeration district and enumeration block are delineated by the corresponding census offices. Enumeration block is the basic geographic unit in the population census.

Uncertainty of boundary of enumeration district is come from no legal boundary in many townships and uncertainty in delineation of enumeration district and enumeration block. There are two ways to determine the boundary of enumeration district and enumeration block. One way is use topographic map or administrative region map. The other way is delineated by the enumerator with no accurate map to support, which means the enumerator map is a sketch map. For the first case, the uncertainty of boundary of enumeration district would come from the accuracy of the reference map, error of enumerator’s interpretation and error in the digitization of enumerator’s map. For the later case, the uncertainty of boundary of enumeration district would come from quality of the sketch map.

2.3 Uncertainty in the Integration of Multi-source and Multi-scale Geographic Data
In the 2000 population census of China, there is no uniform geographic database to support the production of enumerator’s map. Different province or different county my use the maps at different scales or different periods. The geographic database for Population Geographic Information System has the characteristic of multi-source and
multi-scale. When this data is integrated in the one spatial database, uncertainty would occur.

2.4 Uncertainty of Virtual Enumeration District
In the 2000 population census of China, because some persons do not clearly belong to some level of administrative region, such as person served in army, prisoner, and some large enterprises which have some sort of governmental capability, etc., the boundary of these region can not be clearly delineated and virtual enumeration districts are created. These regions may be overlaid with the administrative regions and the uncertainty may occur.

3 UNCERTAINTY OF POPULATION DATA

In the Population Geographic Information System, the uncertainty of population data plays a more important role than the uncertainty of geographic data. Uncertainty of population data includes uncertainty in population enumeration, error in the enumeration map and uncertainty in the population data processing.

3.1 Error in the Enumeration Map
As mentioned above, because of the uncertainty of the population enumerator’s map, the enumeration district may be overlaid or some enumeration district may be lost. When the enumerators visit the person according these maps, the number of person may be overcounted or undercounted.

3.2 Uncertainty in Population Enumeration
In the 2000 population census of China, quality control was introduced in each link of the census process to ensure the quality of census data. Strict specification was prepared in recruiting enumerators. Enumerators had to sit for intensive training and examination before they were recruited. Detailed work flow charts were prepared. Census data collection had to follow the predetermined procedures. The whole census area was divided into enumeration districts and blocks, with each enumerator responsible for one enumeration block. After clarifying the boundary and basic situation of the enumeration block, he/she was to prepare a map for the enumeration block, on which each building with possible inhabitants had to be marked. Then he/she was to make pre-enumeration visit to the households and prepare a list of heads of households. During the enumeration, the enumerator, taking the map and the list of household heads as the reference, asked questions and got answers in person for each household and wrote down the records for each individual in the household. After the completion of the field enumeration, repeated checks and cross verification on the census questionnaires were conducted by enumerators. The requirement was that there should be no mis-recording of items for each individual, no omission of a single individual in the household, no omission of a household in the building, and no omission of a building in the census block. The strict regulations were closely followed in the field enumeration.

As China has the largest population in the world, over 10,000 tons of papers were used for the printing of census questionnaires, another 400 million copies of Letter to Households, 1 million copies of Population Census Publicity Manual and tens of millions of posters were printed and distributed. More than 5 million enumerators and 1 million supervisors were mobilized. Because of the large scale covering a population of about 1.26 billion, the post-enumeration survey conducted after the 2000 population census of China indicated 1.81% undercount of total population. This is one source of the uncertainty of population data.

3.2 Uncertainty in Population Data Processing
In order to obtained as much as information as possible while keep the work load to the manageable limit long-form and short-form approach, which had been popular in the world, was adopted in the design of the census questionnaires. The
short-form was completed 90 percent of the households and the long-form was used by the remaining 10 percent of the households. Items covered in the short-form were reduced from 21 in the 1990 census to 19 in the current census, and the content of the long-form were enriched in order to reflect the requirements of the economic and social development. Information collected from the census included natural characteristics of the population, migration, education attainment, economic participation, marriage and fertility, living conditions and environment, etc. Information on housing was collected for the first time in the Chinese population census.

The OCR (optical character reading) technology was used for the first time in the 2000 population census, contributed to saving human resource, reducing possible recognition errors introduced by manual data entry shortening the time required for data processing and improving the quality of data. Because of the complicated environment, such as machine, paper, writing, software, Uncertainty in population data processing is still exist.

4 UNCERTAINTY INDUCED IN THE DATA PROCESSING AND ANALYSIS

Uncertainty induced in the data processing and analysis includes uncertainty induced by population distribution unevenly, disagreement between boundary of query and boundary of enumeration, and uncertainty in the population data visualization.

4.1 Uncertainty Induced by Population Distribution Unevenly

Because of the unbalance of social and economic development and limitation of the geographic environment, the population in China is distributed unevenly. In general, east part is densely populated and west part has less population. More population live in plain and basin area and less population live in mountainous region and plateau area. More people live in agriculture area and less people live in forest area. In these circumstances, when the spatial analysis, such as union, combination, clip, and overlay, are performed cross the basic enumeration blocks in the GIS, the accurate population number can not be obtained. The uncertainty occurs.

4.2 Disagreement Between Boundary of Query and Boundary of Enumeration

When we query the population data, the query boundary is not always the same as the enumeration boundary. Three kinds of spatial relationships between the query boundary and enumeration boundary exist, which are separate, intersect and contain. Sometimes final data must interpolated from the original enumeration district (Figure 2).

Figure 2: Spatial Relationships Between the Query Boundary and Enumeration Boundary
4.3 Uncertainty in the Population Data Visualization

When the population data is represented in GIS, such as make the thematic map, normally the population data must be classified. In that case, the difference in the same class disappears. The thematic mapping model can induce the uncertainty.

5 CONCLUSIONS

Population census is the basic way to obtain information on the status of the population. China conducted 4 population censuses during 1953-1990, and the results of the censuses played important role in the country’s decision-making at the macro level. The 2000 population census, as a key survey of the national conditions and strength conducted at the turn of the century, is of great significance. Rich information is available on the size of Chinese population, its sex and age composition, industrial and occupational distribution, composition of population with various education attainments, employment and unemployment, migration, marriage and fertility, urbanization process, living environment, etc. Cross tabulation of census data will provide even richer information for in-depth studies on population reproduction and on a series of social end economic issues.

In China, population is always an issue of overall importance. The large size of population is always an influential factor that has to be taken into consideration in all major decision-making process. With the valuable information from the 2000 population, we can better understand the trend of change in the Chinese population in order to formulate sound population policies and programs for social and economic development by further control of the growth rate and improvement in the quality of the population in the 21st century. It also helps us in dealing with such issues as employment, urban development and management, development of education, environment protection, improvement of the social security system, so as to promote the coordinated social and economic development and the implementation of the strategy of sustainable development.

Geographic Information System has a great capability for decision-making. The focus is on the complicated spatial analysis and modeling. The model can be analytic, simulative, logistic and mathematic. The model and data in the Geographic Information System has the nature of uncertainty. When we use the decision made by Geographic Information System, we must fully understand the uncertainty in the GIS model, data and its propagation.

6 REFERENCES

Biography of the Presenting Author

LIU Deqin, Professor in the Chinese Academy of Surveying and Mapping (CASM) and the deputy director of Institute of Cartography and GIS, CASM. His research interest is the theory and application of Geographic Information System. In the 2000 Population Census of China, he is chief of the project to establish the national and provincial level Population Geographic Information System in China.