CLASSIFICATION ACCURACY ASSESSMENT FOR REGIONAL VECTOR DATA PRODUCT BASED ON SPATIAL SAMPLING: A CASE STUDY OF JAPAN

Yao Lu [1,2,*], Jixian Zhang [2], Xiaohua Tong [1], Wenli Han [2], Haitao Zhao [2]

[1] School of Surveying and Geo-Informatics, Tongji University, Shanghai, China. 200092 - xhtong@tongji.edu.cn
[2] National Quality Inspection and Testing Center For Surveying and Mapping Products, Beijing, China 100830 - 156291029@qq.com, zhangjx@casm.ac.cn, 88212479@qq.com, 85623202@qq.com

KEY WORDS: Spatial sampling, Spatial Vector Data, Stratification Strategy, Classification Accuracy Assessment, Spatial Correlation

ABSTRACT:

Spatial vector data is a kind of data that represents real spatial information through points, lines and polygons. Spatial data quality is one of the basic theoretical research in geographic information science. Accurate and reliable data quality assessment is very important for its theoretical significance and practical value. This paper proposes an improved method for the traditional classification accuracy evaluation of spatial vector data :(1) Quantitative estimation of sample size. According to the statistical principle of probability theory, the overall quantity is estimated by controlling the sampling error and the acceptance quality level. The sample quality is the unbiased estimate of the overall quality. (2) Stratification strategy: the overall objects are divided into three layers according to the three basic geometric structures -- points, lines and polygons. The difference within the layer is small and the difference between layers is large, which conforms to the basic principle of stratification. Then, the proportion of the total number of elements in each layer is taken as the weight to distribute layer by layer, and the sample size of each layer is obtained. (3) Allocation of samples. The spatial property of spatial sampling is mainly reflected in the allocation of samples. Considering the spatial correlation of elements in same layer, Local Moran's I index was used to calculate the correlation degree of a certain attribute between each spatial element and its neighbouring elements. After cluster analysis of elements in each layer, samples were screened by setting a reasonable threshold value. (4) Sample inspection. Each sample was examined against reference information, including images and data. The classification of each sample is judged by the principle of majority judgment. (5) Classification accuracy assessment. The classification accuracy information of samples was obtained by making the confusion matrix of the classification result of samples and the real results. The classification accuracy of experimental data is evaluated according to Kappa index. A case study of Global Core Vector Data of Japan shows the improved method in this paper and process of classification accuracy assessment for regional spatial vector data product. Global Core Vector Data are organized according to the country or region, including three categories of transportation, river system, place names, which are divided into 8 middle categories and 52 small categories. In this paper, 1405 samples of Global Core Vector Data in the experimental area of Japan are selected by spatial stratified sampling in 3 strata. The experimental results show that the proposed improved method is applicable to classification accuracy assessment of regional spatial vector data product and overcomes the disadvantages of type-based spatial stratified sampling that relies on the classification information of all elements. The Kappa coefficient is 0.831, which reflects the result of classification accuracy assessment in the experimental area is good. The proposed improved method provides a reference for the method of classification accuracy assessment classification of following global spatial vector data product.

1. INTRODUCTION

Geospatial database is a database of geographic data and information, such as countries, cities, natural landscape, cultural landscape and related information (Donath, M., et al., 2006). It is divided into raster data and vector data. Vector data is a kind of spatial data represented by three simple geometric structures of points, lines and planes (Xiaohua, T., et al., 2011). There are a lot of geospatial vector database around the world, such as The National Register Information System (NRIS), OpenStreetMap (OSM), The European Soil Database, Geo-Names Data and so on. These databases are only specific to a certain application field or an area, but lack of coverage and universality. China has produced an annual updated Global Core Vector Data (GCVD), which includes road network, water systems and geographical names.

In the whole process of vector data produced, it faces various quality problems, such as the uncertainty of satellite image, data sources and reference data, production technology, etc. These uncertainties can lead to quality problems in vector data products, causing serious losses in related applications. Therefore, the development of a reasonable accuracy assessment method for data quality control is the key to ensuring the accuracy of vector data.

Sampling inspection is an important means of quality control and provides reliable information on the quality of a product (Wetherill, G. B., 2013). Traditional sampling methods include simple random sampling, stratified sampling, systematic sampling, and cluster sampling.

The sampling methods of existing geospatial vector data products are mainly based on random sampling, such as OpenStreetMap (Ciepłuch, B., et al., 2010), GeoNames Data (Ahlers, D., 2013) and so on. Geospatial vector data has the characteristics of massive, multi-dimensional and non-homogeneous. Objects with different geometric structures have different natural distributions, and there is spatial autocorrelation between objects of the same structure. Therefore, the traditional sampling method can't describe the spatial information of the sample. it is necessary to adopt a more rational sampling method for determining the

* Corresponding author.
E-mail addresses: 156291029@qq.com (Yao Lu)

location of samples. In terms of sample size, the existing sampling method mainly uses expert experience or fixed value, which lacks reasonable and effective estimation model. It makes the sample size generally low.

In view of the characteristics of geospatial vector data and the defects of the existing sampling methods, considering the spatial correlation of samples, this paper proposes a spatial stratified sampling method based on geometric structure. At the same time, taking some areas of GCVD in Japan as an example, Kappa coefficient were calculated by using the confusion matrix. Finally, a discussion on the proposed method is given in Section 4.

2. METHODOLOGY

2.1 Sampling principles

Geospatial vector data is represented by the geometry and attribute information of the object. Its characteristics are obvious, that is, diverse geometric structures, each structural category has its own distribution pattern in space, and there is correlation between the same structure data. Therefore, to evaluate the classification accuracy of space vector feature data, the following principles must be considered:

(1) Representativeness of geometric structures: for the classification accuracy assessment of geospatial vector data, the three geometric structures----points, lines and polygons should be sampled and participate in the final accuracy assessment. Only by synthesizing the samples of all geometric structure types can the overall accuracy be unbiased.

(2) Rationality of sample size: the calculation of sample size needs to be based on science, and the relationship between the number and the total amount of each geometric structure needs to be considered. Too much sample size will lead to higher costs such as time and funds, while too small sample size will lead to deviation of precision evaluation results.

(3) Representativeness of samples: according to the first law of geography (Tobler, W. R., 1970), every object in space has a spatial correlation. The spatial distribution of samples should avoid clustering distribution as much as possible. If the spatial correlation of clustered samples is obvious, the final accuracy will be overestimated or underestimated (Jinfeng, W., A, Stein, Gao, B. B., Ge, Y., 2012).

According to the above three principles, the sampling of accuracy assessment needs to be considered from three aspects: sampling method, determination of sample size, and allocation of samples.

2.2 Sampling Method

Geospatial vector data has geometry attribute, space attribute and information attribute. Stratified sampling by region will lead to undefined subdivision scale. Stratified sampling based on classification can avoid the problem of scale. The difference within stratification is small, and the difference between stratifications is large. However, this method needs to calculate the weight of each classification in the population, and usually the number of each classification is uncertain. Therefore, stratified sampling based on geometric structures can not only cover different types of geometric structure, but also avoid the uncertainty of the number of different classifications. This method can achieve the effect of stratification and accord with the principle of sample representativeness.

2.3 Determination of sample size

In accuracy assessment, the previous determination of sample size is mainly based on empirical value or fixed value, and there is a lack of reasonable sample size estimation model. According to the theory of probability and statistics, the sample size n can be calculated as

$$n = \frac{\frac{\mu^2_{1-a/2}p}{r^2(1-p)}}{1+\frac{1}{N}\left(\frac{\mu^2_{1-a/2}p}{r^2(1-p)}-1\right)} \quad (1)$$

Here, $\mu^2_{1-\frac{a}{2}}$ denotes the critical value of the standard normal distribution at the confidence level of $(1-\frac{a}{2})$, r is the limit value of the relative difference, p is expected classification accuracy (Xiaohua, T., et al., 2011).

The sample size of each stratum is proportionate to the population size of the stratum, the sample size of each stratum can be calculated as

$$n_h = nw_h, w_h = W_h = \frac{N_h}{N}. \quad (2)$$

Here, n_h is the sample size of stratum h (h=1 to m, m is the number of stratum), W_h is the weight of stratum h (Wang, J., Haining, R., Cao, Z., 2010).

2.4 Allocation of samples

The spatial property of spatial sampling is mainly reflected in the allocation of samples. Considering the spatial correlation of elements in same stratum, Local Moran's I index was used to calculate the correlation degree of a certain attribute between each object and its neighbouring object (Anselin, L., 1995; Moran, P. A., 1950; Li, H., Calder, C. A., Cressie, N., 2007). After cluster analysis of objects in each stratum, samples were screened by setting a reasonable threshold value. The screening principle is to minimize the spatial correlation of objects and avoid the aggregation of objects (Balaguer-Beser, A., Ruiz, L. A., Hermosilla, T., Recio, J. A., 2013). After that, the samples of each stratum were randomly allocated according to the calculated sample size.

3. EMPIRICAL CASE STUDIES

The accuracy assessment method mainly obtains the accuracy information through the cross tabulation of the sample classification result and the real result, namely the confusion matrix. By means of confusion matrix, accuracy indexes——Kappa coefficient can be calculated (Cohen, J., 1960). Kappa coefficient will be used in the classification accuracy assessment of regional geospatial vector data to describe the quality of data from another perspective.

Kappa coefficient is a special precision evaluation index, which is used to express the degree of consistency between two maps. It can be calculated as

$$k = \frac{p_0 - p_c}{1 - p_c}. \quad (3)$$

Here, p_0 is the relative observed agreement among raters, and p_c is the hypothetical probability of chance agreement. The value range of Kappa is [-1, 1], and the closer to 1, the better the consistency (Sim, J., Wright, C. C., 2005).

3.1 Experimental data for accuracy assessment

Part of Japan in 2018 GCVD was selected for experimental data (figure 1). The classification of objects in each geometric stratum are derived from the GCVD production manual. Expected classification accuracy in table 1 is a reference value obtained by combining previous experiments and expert experience for calculating the sample size.

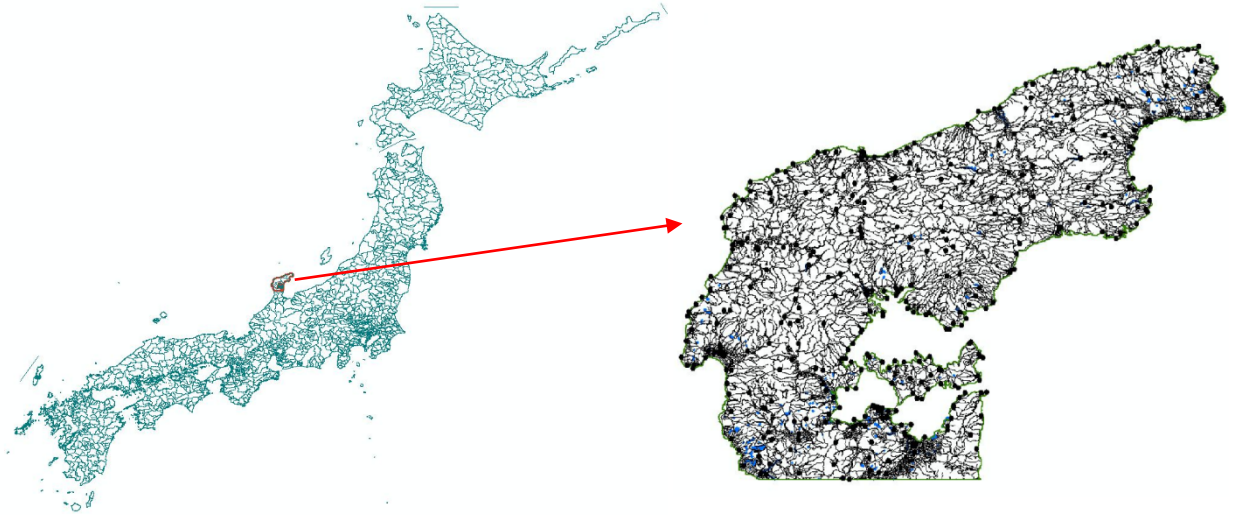

Figure 1. Geographical location and objects of experimental data

3.2 Experimental procedure for precision evaluation

The purpose of the experiment is to use the method proposed in this paper to evaluate the classification accuracy of vector data of three geometric structures in the experimental area, including process of sampling method, determination of sample size and allocation of samples.

The specific experimental steps are as follows:

(1) Determine the overall sample size of the region. The expected classification accuracy p of point, line and plane was calculated through Eq. (4), and then p was substituted into Eq. (1) to calculate the total sample size of the experimental area. p can be calculated as (Sim, J., Wright, C. C., 2005)

$$p = \sum_{h=1}^{m} p_h \times \omega_h. \quad (4)$$

Substitute the calculated p value (p=80%) into Eq. (1), where N is the total number of objects in the experimental area, r=0.1, and the calculated total sample size in the experimental area is 1405 (table 1).

(2) Stratified sampling. According to the above discussion results, stratified sampling was carried out according to the geometric structure. The total sample size of the experimental area was 1405, and the samples were distributed to each stratum through the quantitative weight of each geometric structure.

According to the distribution method of quantity weight, it solves the quantitative problem of sample size, avoids the problem that sampling strength is not universal due to empirical value or fixed value, and ensures that every geometric type has samples, making the accuracy assessment result more scientific and reasonable (Levy, P. S., Lemeshow, S., 1999).

(3) Spatial correlation analysis. The similarity of samples is inversely proportional to the representativeness of samples. Samples with high similarity will reduce the representativeness and sampling efficiency, which obviously cause the bias of accuracy assessment results. Therefore, spatial correlation analysis is carried out stratum by stratum, that is, Local Moran's I index is calculated for each stratum of data, which can quantitatively represent the clustering degree of the same geometric structure in space. The samples were then filtered by setting a reasonable threshold.

Stratum	Geometric structure	Classification of object	Number of object	Expected classification accuracy (%)	Quantity ratio (%)	Sample size
1	Point	Geographical name	377	82	1.3	18
2	Line	Road network and water systems	29012	76	98	1377
3	Polygon	Water systems	214	80	0.7	10
total			29603			1405

Table 1. Calculation of sample size of experimental area

Taking linear objects in the experimental area as an example, spatial correlation was calculated by ArcGIS software spatial analysis module. Figure 2 represents the linear sample after spatial analysis, in which the black object represents the sample with low spatial correlation and the red object represents the sample with high spatial correlation. The linear objects are mainly composed of natural landscape and human activities.

Figure 2. Spatial correlation analysis of linear objects

After spatial analysis, samples with low spatial correlation were screened out according to Moran's I index. On this basis, 1370 samples were randomly selected according to the linear samples calculated in step (2). The distribution results are shown in figure 3. After spatial analysis stratum by stratum, the sample distribution results of the other two stratum are shown in figure 4 and 5.

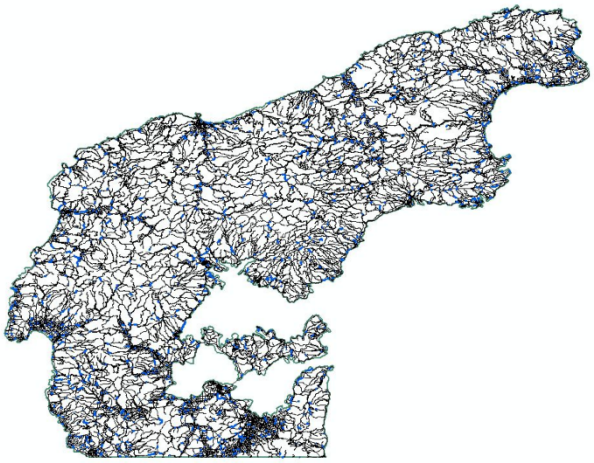

Figure 3. Samples in stratum of lines (blue lines represent samples).

(4) Sample inspection. The sample unit is a single representational object. The reference data adopts OSM, Google Earth and Google Maps with the same current situation. The majority rule is adopted in the sample inspection: that is, each sample is judged by 3 experts independently, and 2 or more consistent classification results are deemed to be correct, otherwise it is wrong.

(5) Classification accuracy assessment. The confusion matrix is established according to the results of sample inspection. The classification accuracy is evaluated by calculating a series of precision indexes.

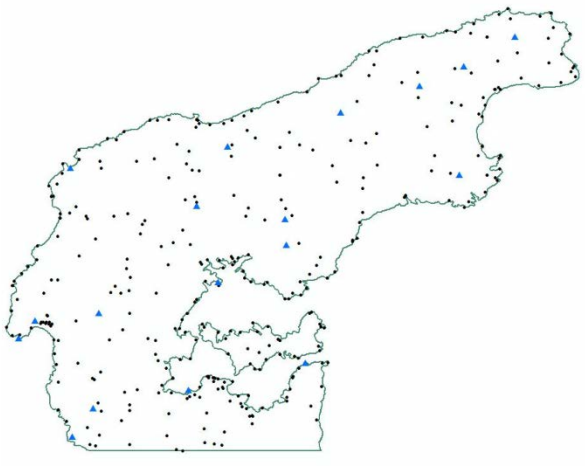

Figure 4. Samples in stratum of points (blue points represent samples).

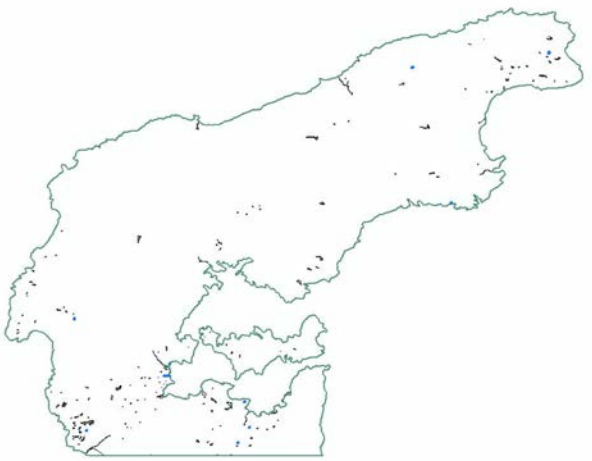

Figure 4. Samples in stratum of polygons (blue polygons represent samples).

	Points	Lines	Polygons
Points	16	4	0
Lines	2	1370	1
Polygons	0	3	9

Table 2. Confusion matrix of samples in the experimental area

3.3 Analysis of experimental results

The confusion matrix of samples was established through the above steps (table 2). The Kappa index is calculated based on the confusion matrix.

In the sample allocation, the proportion of sample size per stratum in the population is considered. Therefore, the overall accuracy obtained is not only the overall accuracy of the finite sample, but also reflects the entire area, which can be confirmed

from the overall accuracy value: the number of linear objects in the experimental area accounted for 98% of the total number of objects, and the correct rate of this type is 99%, which is consistent with the overall accuracy. The Kappa coefficient is 0.831, which is close to 1, indicating that the consistency between the experimental area data and the real data is very high.

4. CONCLUSIONS

In this paper, the GCVD produced in China is analysed through the characteristics of vector data, and some regions in Japan are taken as examples for experiments. In consideration of the characteristics of spatial data and the scientific probability and statistics method, the sampling method, sample layout method and determination of sample size are improved, and a classification accuracy assessment method of regional spatial sampling based on geometric structure is proposed.

The results of GCVD data from the experimental area in southwest Japan are as follows: 1405 samples are selected by three types of geometric structure in the experimental area in 2018. By calculation, Kappa coefficient is 0.831 and the overall accuracy is 99.29%. The results show that the method proposed in this paper is suitable for the classification accuracy assessment of regional or small-range spatial vector data, which reflects the good quality of data in some areas of GCVD produced in China. To further expand the classification accuracy assessment on a global scale, it is necessary to consider the relationship between the accuracy evaluation object and spatial scale, and the method to calculate the spatial relationship at different scales. The method proposed in this paper can provide reference. The accuracy assessment for global spatial vector data based on spatial sampling is the main research direction in the future.

REFERENCES

Ahlers, D., 2013. Assessment of the accuracy of GeoNames gazetteer data. In Proceedings of the 7th workshop on geographic information retrieval (pp. 74-81). ACM.

Anselin, L., 1995. Local indicators of spatial association—LISA. Geographical analysis, 27(2), 93-115.

Balaguer-Beser, A., Ruiz, L. A., Hermosilla, T., Recio, J. A., 2013. Using semivariogram indices to analyse heterogeneity in spatial patterns in remotely sensed images. Computers & geosciences, 50, 115-127.

Ciepłuch, B., Jacob, R., Mooney, P., Winstanley, A. C., 2010. Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. In Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010 (p. 337). University of Leicester.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational & Psychological Measurement, 20(1), 37-46.

Donath, M., Newstrom, B., Shankwitz, C.R., Gorjestani, A., Lim, H., Alexander, L., 2006. Real time high accuracy geospatial database for onboard intelligent vehicle applications.

Jinfeng, W., Haining, R., Cao, Z., 2010. Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. International Journal of Geographical Information Science, 24(4), 523-543.

Jinfeng, W., A, Stein, Gao, B. B., & Ge, Y., 2012. A review of spatial sampling. Spatial Statistics, 2, 1-14.

Levy, P. S., Lemeshow, S., Ferketich, A., 1999. Sampling of populations : methods and applications. Technometrics, 34(3), 372-372.

Li, H., Calder, C. A., Cressie, N., 2007. Beyond Moran's I: testing for spatial dependence based on the spatial autoregressive model. Geographical Analysis, 39(4), 357-375.

Moran, P. A., 1950. Notes on continuous stochastic phenomena. Biometrika, 37(1/2), 17-23.

Sim, J., Wright, C. C., 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical therapy, 85(3), 257-268.

Tobler, W. R., 1970. A computer movie simulating urban growth in the Detroit region. Economic geography, 46(sup1), 234-240.

Wetherill, G. B., 2013. Sampling inspection and quality control (Vol. 129). Springer

Xiaohua, T., Wang, Z., Xie, H., Liang, D., Jiang, Z., Li, J., Li, J., 2011. Designing a two-rank acceptance sampling plan for quality inspection of geospatial data products. Computers & geosciences, 37(10), 1570-1583.