A QUALITY ANALYSIS AND UNCERTAINTY MODELING APPROACH FOR CROWD-SOURCING LOCATION CHECK-IN DATA

Meng Zhou [a, *], Qingwu Hu [a], Ming Wang [a]

[a] School of Remote Sensing and Information Engineering, Wuhan University, China -
(zhoumeng,huqw,m_wang)@whu.edu.cn

KEY WORDS: Crowd-sourcing, Location check-in, Quality analysis, Spatial registration, Uncertainty, Error distribution

ABSTRACT:

The location check-in data, developing along with social network, are considered as user-generated crowd-sourcing geospatial data. With massive data volume, abundance in contained information, and high up-to-date status, the check-in data provide a new data source for geographic information service represented by location-based service. However, there is a significant quality issue regarding to crowd-sourcing data, which has a direct influence to data availability. In this paper, a data quality analysis approach is designed for the location check-in data and a check-in data uncertainty model is proposed. First of all, the quality issue of location check-in data is discussed. Then, according to the characteristics of check-in data, a location check-in data quality analysis and data processing approach is proposed, using certain standard dataset as reference to conduct an affine transformation for the check-in dataset, during which the RANSAC algorithm is adopted for outlier elimination. Subsequently, combining GIS data uncertainty theory, an uncertainty model of processed check-in data is set up. At last, using location check-in data obtained from jiepang.com as experimental data and selected navigation data as data standard, multiple location check-in data quality analysis and uncertainty modeling experiments are conducted. By comprehensive analysis of experimental results, the feasibility of proposed location check-in data quality analysis and process approach and the availability of proposed uncertainty model are verified. The novel approach is proved to have a certain practical significance to the study of the quality issue of crowd-sourcing geographic data.

1. INTRODUCTION

With the increasing improvement and popularization of Web2.0 technology, location-based service (LBS) has become among the most rapidly developing geographic information service applications (Liang, 2011). And with the rapid development of LBS, the demand for the volume and timeliness of location information is increasing rapidly. The term “point of interest” (POI) refers to a specific location that certain individuals find interesting or useful, or in a broader sense, all geographical objects that can be represented by points. It is currently among the major and most applied forms of location information. The data volume, timeliness, and reliability of POI have a direct influence to LBS. Since traditional POI update is mainly carried out by professional surveying division through manual field collection and storage, its data products cannot meet the demand of data timeliness (Han, 2008). It is detrimental for the development of LBS.

Location check-in data refer to the spatial location data obtained through check-in operation. As for check-in, it’s defined as the activity of a user to record location information at a certain location using mobile terminal represented by smart phones.

Location check-in data are user generated contents. It is characterized by large data volume, abundant content, and high sensitivity to time. In particular, its data volume and timeliness are incomparable to traditional data forms. Therefore, check-in data can be used for as new POI data source of LBS. Check-in data also have abundant time stamps and check-in frequency information, which are suitable for the analysis of POI development.

Location check-in data are strongly crowd-sourcing characterized. As a data output of the harnessing of tools to create, assemble, and disseminate geographic data provided voluntarily by individuals (Goodchild, 2007), check-in data have natural defects in data quality and accuracy. The data content is asserted rather than authoritative which probably means no quality control in any form (Goodchild, 2010). Also there seems to be no valid method to handle such complex issue. As a result, the quality analysis and data processing of check-in data are crucial in order to meet the request of LBS applications. A quality analysis framework for location check-in data as well as a spatial registration approach based on the RANSAC algorithm with the purpose of improving positional accuracy is put forward in this paper. For the data output of spatial registration, an uncertainty model is established adopting classical uncertainty theory. The feasibility of proposed approaches is tested through an experimental validation.

2. QUALITY ANALYSIS

It is essential to perform quality analysis for location check-in data and other crowd-sourcing characterized data. A suitable quality model and its elements are the basis of an effective quality analysis, while a proper workflow of the analysis approach should be utilized for the analysis.

2.1 Quality Model

During a progress of quality analysis for the OpenStreetMap data, researchers brought up an indicator model for crowd-

[*] Corresponding author.

sourcing data quality. In this quality model, the completeness, thematic accuracy, and positional accuracy are included as the three components of the quality elements. The quality of the road data is assessed by the calculation and the analysis of these elements.

Learning from the quality model mentioned above, meanwhile considering the check-in data features, a quality model for location check-in data is proposed in this paper. In this model, the classification accuracy, degree of matching, and positional accuracy are selected as the quality elements. To be specific, the classification accuracy shows the accuracy of the classification attributes, the degree of matching indicates the overlapping between check-in data and standard data in content coverage, and the positional accuracy is presented by the offsets between check-in data and standard data in spatial locations.

2.2 Quality Analysis Approach

The matching between location check-in data and standard data is the basis of quality analysis. The data participated in the analysis are the successfully matched records in the matching operation. The others are insignificant in quality analysis, but valuable in POI update.

After data matching, the classification accuracy can be calculated by the comparison of classification attributes between both datasets. Simultaneously, the degree of matching can be collected from the statistical counts of successfully matched data records. And similarly, the positional accuracy can be assessed by a statistical analysis of the offsets between check-in data records and corresponding standard data records in spatial locations.

The technical workflow for the proposed quality analysis approach is as shown in Figure 1.

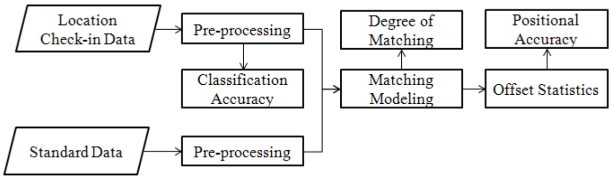

Figure 1. Flow chart of the quality analysis

3. QUALITY CONTROL

After quality analysis, a data processing progress would be achieved in order to perform quality control for the check-in data. A pre-processing operation and a spatial registration operation are included in the quality control procedure proposed in this paper. The pre-processing operation is for the quality control of the attribute information. While the spatial registration, being the emphasis of the data processing, is for the purpose of improving positional accuracy of check-in data. The RANSAC algorithm is adopted in this paper for model establishment. The result model is to be used in the registration operation to achieve positional quality control.

3.1 The RANSAC Algorithm

The Random Sample Consensus (RANSAC) algorithm is designed to estimate a group of mathematical model parameters from observed datasets with abnormal value data. It is put forward originally by Fischler and Bolles in 1981 (Fischler, 1981). The basic assumption of RANSAC is that in the samples, there are inliers which can be expressed in a certain model as well as outliers which are distinctly abnormal and cannot fit in any model. In short, there is noise in the datasets. In addition, the algorithm also suggests that given a group of inlier data, a set of model parameters suitable for the expression of these data is existent and can be obtained by calculation.

The RANSAC algorithm is highly robust with the ability to estimate high-accuracy parameters from a dataset with a considerable amount of outliers. It is suitable for the establishment of a certain optimal model from a dataset with a relatively large deviation.

3.2 Pre-processing

In data pre-processing, the completion of missing attributes and the amalgamation of repeated records are included. The completion of missing attributes needs to rely on certain standard format in order to preserve data value in use. The repeated data records need to be merged to reduce redundancy (Du, 2011). The amalgamation of different aliases, nicknames, and standard names of the same geographic object can be achieved through the comparison between POI data dictionary and check-in data records (Wu, 2012).

3.3 Spatial Registration

The spatial registration process is required to reduce offset error and improve positional accuracy. The RANSAC algorithm is applied in this paper to estimate the affine transformation relational model between location check-in data and corresponding standard data. The basic idea of the algorithm is that, when estimating parameters, by obtaining basic data subset through repeated sampling, model estimation is achieved (Shan, 2006). To acquire optimal model by data fitting, the size of randomly selected sample needs to be limited, meaning the minimum data set size to determine the model needs to be specific. In this paper, the affine transformation formula is used as model, which means at least four point pairs are demanded for the solution of the six parameters.

The spatial registration process is described as follow.

Initialization: Initialize model by selecting four point pair samples randomly from set.

Parameter Estimation: Identify an inner point set that is suitable for current model using threshold. If the size of this set is larger than a pre-defined threshold, re-evaluate model parameters using this set.

Optimal Model Solution: Define suitable iteration count. Then during these iterations, use the maximum inner point set to re-evaluate model parameters and obtain optimal model (Qu, 2010).

4. UNCERTAINTY MODELING

Uncertainty is the reflection of how the various objects and processes in the natural world are short of certainty (Hu, 2004). Due to the broad usage in subtly different ways in various fields and a relatively wide range of implications, there is yet to be an agreed definition to this term. As for in the field of GIS or geoscience, uncertainty can be regarded as a variation the objective entities have, mainly expressed as inaccuracy, randomness, and ambiguity (Wu, 2002). The error of data, the fuzziness and incompleteness of data and concepts can be all deemed as among the area of uncertainty issue.

To study the uncertainty of location check-in data and to establish a suitable uncertainty model have significant influence to both the utility of the data and further data analysis and data mining.

4.1 Uncertainty Theory

The study of uncertainty is substantially important to data usage and sharing and is also fundamental to ensure the usability of geospatial data during whole lifecycle. The uncertainty of geospatial data is presented in many aspects, including positional uncertainty, attribute uncertainty, logical uncertainty, and incompleteness.

Due to the coverage of uncertainty in such wide range, it is naturally unpractical to research or analyze uncertainty issue by a certain method or theory alone. The research on uncertainty is largely interdisciplinary, with plenty of proven methods in mathematics and computer technology adopted and applied. For example, classical error theory, fuzzy set theory, and entropy theory have all been applied in studying positional uncertainty, while rough set theory, target model, and spatial statistics are proven useful in exploring attribute uncertainty.

4.2 Uncertainty of Check-in Data

Almost every aspect of geospatial data can be related to uncertainty. As for location check-in data, the positional accuracy issue is relatively noticeable and is the core issue of the uncertainty modeling for the check-in data.

The uncertainty of the two-dimensional point objects represented by check-in data is mainly performed as the deviation in two-dimensional coordinates, namely the planar positional error. The study of the point positional error distribution of check-in data based on classical error theory is carried out in this paper.

According to error theory, the positional uncertainty of planar points can be expressed by error ellipses. Under the assumption that the error distribution accords with bivariate normal distribution, the bivariate normal distribution function can be used to present the error distribution of the dataset. In this case, the error features of planar points along both coordinate axes are expressed in the density function. The expression of the function is as shown below.

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\{-\frac{1}{2(1-\rho^2)}[\frac{(x-\mu)^2}{\sigma_1^2}-2\rho\frac{(x-\mu)(y-\nu)}{\sigma_1\sigma_2}+\frac{(y-\nu)^2}{\sigma_2^2}]\} \quad (1)$$

Where μ and ν are respectively the mean value of error in X and Y directions, σ_1 and σ_2 are respectively the standard deviation of error in X and Y directions, and ρ is the correlation coefficient.

Given the description above, the expression of the error ellipse formula is as shown below.

$$\begin{aligned}&(x-\mu, y-\nu)B^{-1}(x-\mu, y-\nu)^T = R^2\\&B=\begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}, \sigma_{12}=\rho\sigma_1\sigma_2\end{aligned} \quad (2)$$

5. EXPERIMENTAL ANALYSIS

A series of test experiments is conducted to verify the feasibility and practicality of the proposed approach.

Location check-in data records of two different time nodes in September and October, 2011 from jiepang.com and the 2011 version of navigation data from Navinfo Co. as original POI data records are used in this experiment. The procedure is similar to what's described in the method mentioned before.

5.1 Quality Analysis Experiment

5.1.1 Classification Accuracy: A total of 4584 data records participated in this experiment with 149 records being misclassified, which makes the accuracy 96.75%. High classification accuracy shows relatively high quality in check-in data attribute information.

5.1.2 Degree of Matching: In this test, the degree of matching is further divided into the degree of spatial matching and the degree of attribute matching. The degree of spatial matching is the ratio of the check-in data successfully matched in both spatial location and name attribute. And the degree of attribute matching is the ratio of the check-in data successfully matched in name attribute only, which means larger deviation in spatial location than a certain threshold.

A total of 4584 data records are used in this test, 752 of which are qualified as spatial matching and 145 of which are qualified as attribute matching. Therefore, the degree of spatial matching is 16.40% and the degree of attribute matching is 3.16%. A relatively low ratio of spatial matching indicates a large number of potential new POI data records in the check-in dataset, and a significantly low ratio of attribute matching indicates the number of the check-in data records with large spatial gross error is very small.

5.1.3 Positional Accuracy: In this experiment, the mean value of offsets between spatial matched check-in data records and standard data records is 596.49 meters and the number rises up to 4270.52 for the attribute matched records.

In general, the deviation between check-in data and standard data in location is relatively large, which proves the existence of data accuracy issue. The results also show that the threshold value for the test is proper since the mean value of deviation differs a great deal between spatial matching and attribute matching.

5.2 Spatial Registration Experiment

A RANSAC adopted affine transformation model is used to perform spatial registration of location check-in data in this experiment. A total of 188 data pair records of check-in data and POI data with the exact same name attributes are selected for the experiment. The RANSAC algorithm is capable of deleting points with gross error, which leaves only inner points that have no gross error participating in model estimation. Under the condition of iteration count set to 20, threshold for inner point verdict set to 0.003, and minimum inner point number set to 100, 120 pairs of effective inner points are extracted. An affine transformation of these 120 records is carried out according to the optimal transformation model. A statistical analysis of spatial offsets between check-in data and corresponding original POIs before and after the registration is carried out and the result is as shown in Figure 2. The average offsets and standard deviations before and after the registration are as shown in Table 1. The results show that the accuracy of location check-in data is significantly improved through spatial registration.

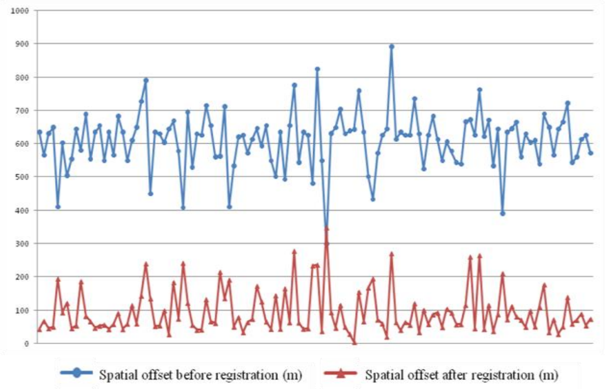

Figure 2. Spatial offsets before and after registration

	Mean value (m)	Standard deviation (m)
Before Registration	602.9069	31.43833
After Registration	57.70259	15.34206

Table 1. Average offsets and standard deviations before and after registration

5.3 Uncertainty Model Validation

The 188 check-in data records through spatial registration and the corresponding standard data records are analyzed in this validation. The results are as shown in Figure 3, where the charts above and below are respectively the demonstrations of the comparison between point coordinate error distribution and normal distribution in X and Y directions. The results show relatively high similarity between location uncertainty distribution of the data and normal distribution, which implies high rationality of applying error ellipse model with bivariate normal distribution model to express the uncertainty of location check-in data.

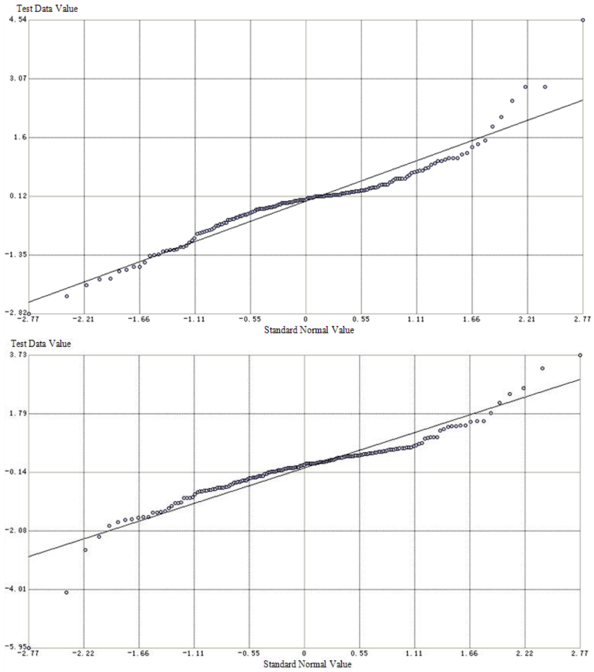

Figure 3. Comparison between check-in data error distribution and normal distribution

6. CONCLUSION

The significance of location check-in data towards POI update and LBS development and the objective existence of data quality issues make it critical to implement quality analysis and data processing. A quality analysis and uncertainty modeling approach for location check-in data is proposed in this paper. The experimental results show high feasibility of the quality analysis approach and excellent ability of quality control method to improve data accuracy. Deep analysis and data mining of the location check-in data will be the focus of future research.

ACKNOWLEDGEMENTS

The authors would like to express appreciations to the Fundamental Research Funds for the Central Universities to support the project “Quality Analysis and Data Mining of Crowd Sourcing Check-in Data ” (Grand No: 2012213020208).

REFERENCES

Du, P. and Liu, Y., 2011. Recognition of Chinese place names based on ontology. *Journal of Northwest Normal University (Natural Science)*, 47(6), pp. 87-93.

Fischler, M. and Bolles, R., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), pp. 381-395.

Goodchild, M., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), pp. 211-221.

Goodchild, M. and Glennon, J., 2010. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3), pp. 231-241.

Han, X. and Lv, Y., 2008. Web GIS data update based on Wiki technique. *Computer Engineering*, 34(11), pp. 283-285.

Hu, S., Pan, Z., Wang, X. and Tao, B., 2004. On the Uncertainty of GIS. *Bulletin of Surveying and Mapping*, (9), pp. 13-16.

Liang, L., Ren, L. and Wan, Y., 2011. ‘LBS-based social network’ of the management and operations in urban public space. *Information Security and Technology*, (7), pp. 56-59.

Qu, T., An, B. and Chen, G., 2010. Application of improved RANSAC algorithm to image registration. *Journal of Computer Applications*, 30(7), pp. 1849-1851.

Shan, X., Wang, Y. and Dong, J., 2006. The matching method based on RANSAC algorithm for estimation of the fundamental matrix. *Journal of Shanghai Dianji University*, 9(4), pp. 66-69.

Wu, L., Yu, H., Gao, Z. and Cheng, J., 2002. The Frame of GIS Uncertainty and Methods of GIS Data Uncertainty. *Geography and Territorial Research*, 18(4), pp. 1-5.

Wu, Y., Lai, J. and Wu, Y., 2012. Preprocessing of LBS based POI data updating. *Computer and Digital Engineering*, 40(8), pp. 87-89.