Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography
Vol. 37, No. 2, 91-97, 2019
https://doi.org/10.7848/ksgpc.2019.37.2.91

ISSN 1598-4850(Print)
ISSN 2288-260X(Online)
Original article

Generalization of Road Network using Logistic Regression

Park, Woojin [1] · Huh, Yong [2]

Abstract

In automatic map generalization, the formalization of cartographic principles is important. This study proposes and evaluates the selection method for road network generalization that analyzes existing maps using reverse engineering and formalizes the selection rules for the road network. Existing maps with a 1:5,000 scale and a 1:25,000 scale are compared, and the criteria for selection of the road network data and the relative importance of each network object are determined and analyzed using Töpfer's Radical Law as well as the logistic regression model. The selection model derived from the analysis result is applied to the test data, and road network data for the 1:25,000 scale map are generated from the digital topographic map on a 1:5,000 scale. The selected road network is compared with the existing road network data on the 1:25,000 scale for a qualitative and quantitative evaluation. The result indicates that more than 80% of road objects are matched to existing data.

Keywords : Road Network, Map Generalization, Selection and Elimination, Radical Law, Logistic Regression Model

1. Introduction

The development of network data at various scales is useful in providing a web map service or a search route for network data. In particular, a multi-scale model of network data is required for maintenance of digital topographic maps in national mapping agencies and for the generation of a map cache for various zoom levels in a web map service. Although several mapping rules for topographic maps exist, in many cases, the rules are ambiguous, and the subjective judgment of the cartographer can be interrupted during mapping processes. Because of the ambiguity of the production rules and the dependence on manual editing, production of homogeneous mapping results is a complicated problem. Thus, to generate a map dataset with various scales, automated map generalization methodologies should be applied to the original map data.

For map generalization of linear data such as road network, selection operation is important (Choe and Kim, 2007). Because the road network database is composed of several intersection nodes and linear links, the selection is implemented by measuring the relative importance of each node or link (Thomson and Richardson, 1995; Mackaness and Machechnie, 1999). A measurement of the relative importance of the network object utilizes various characteristics. Li and Choi (2002) analyzed the six attribute values of roads, including road type (class), length, number of lanes, number of directions, width, and connectivity. Zhang (2005) built the stroke using the road name and built the ordered strokes using three topographical centralities: degree, closeness and betweenness. Chen *et al.* (2009) measured importance by considering the road class, the length of the stroke, and the

Received 2019. 04. 09, Revised 2019. 04. 19, Accepted 2019. 04. 26
1) Member, Seoul Institute of Technology (E-mail: wjpark@sit.re.kr)
2) Corresponding Author, Member, Korea Research Institute for Human Settlements (E-mail: yhuh@krihs.re.kr)

connectivity for mesh-based elimination of road segments. Liu *et al.* (2010) measured four attributes for the road (i.e., statistical, quantitative, topological, and thematic) based on the stroke and the Voronoi diagram for road selection. Zhou and Li (2011) undertook a comparative analysis of different properties (e.g., length, degree, closeness, betweenness, and road class) to determine the importance of individual roads for road network selection methods. Recently, pattern recognition methodologies are having been applied to map object selection model. Zhou and Li (2014) applied a SOM (Self Organizing Map) and BPNN to road network data to update them. The SOM worked as an unsupervised, clustering approach and the BPNN worked as a supervised, classification approach. Zhou and Li (2017) compared nine machine learning methods (ID3, C4.5, CRT, Random Forest, SVM (Support Vector Machine), Naïve Bayes, kNN (k Nearest Neighbor), Multilayer Perception, and Binary Logistic Regression) for road network selection. Lee (2017) tested four machine learning methods such as Naïve Bayes, Decision Tree, kNN, SVM for generalization of building objects.

Most of road network selection methods consider road object attribute information and select road objects for a target scale level considering the attributes. In this process, objects that should be eliminated are determined by importance of each object. However, there are few studies which consider two questions of 'which objects should be selected' and 'how many objects should be selected' simultaneously. In this study, methodology using Töpfer's radical law and a regression analysis to estimate selection model with relative weight for attributes of road network is proposed and evaluated. In the next section, the details of the proposed method are presented and a qualitative evaluation is analyzed in section 3. And section 4 concludes this study.

2. Proposed method

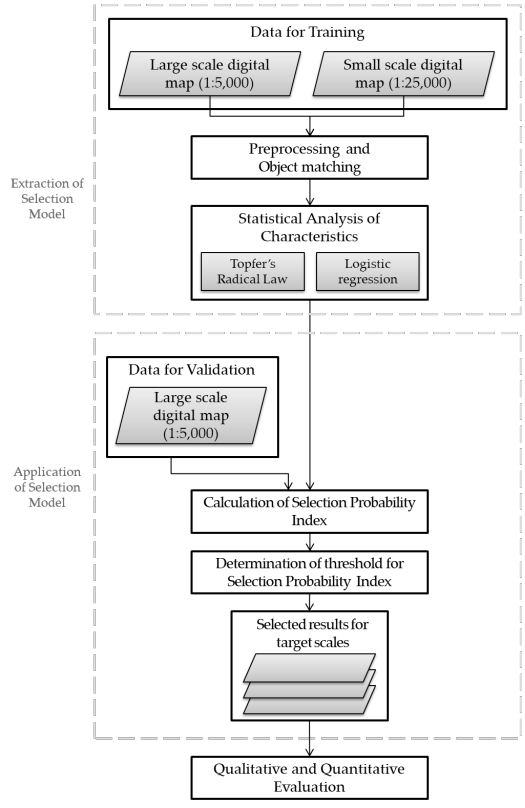

Fig. 1. Workflow of this study

This study uses the workflow as shown in Fig. 1. First, the road centerline data from existing map series with a 1:5,000 scale and a 1:25,000 scale are superimposed, and the matching pairs within the road objects for the two scales are detected. The attribute information for the matched object group and the unmatched object group is measured and analyzed using the regression analysis for calculating the influence of the attributes. Based on the analytical results obtained, the selection model is derived from a form of the probability function with the attribute values and is used as the criteria for the selection of the network objects. The flowchart for the overall processes of knowledge acquisition and the formalization of the road selection is presented in the upper part (extraction of selection model) of Fig. 1.

2.1 Preprocessing and object matching

Automated map generalization for road network needs rich semantics for each road centerline data, thus it is necessary to enrich the source road network with additional information. Table 1 shows attribute schema for the above centerline data.

Table 1. Attribute schema of reconstructed road centerline data

Attribute name	Attribute information	Attribute type
Road class	National expressway, national road, local road, road between towns, narrow path, unclassified road	STRING
Width of road	Length between outer ends of the curb	DOUBLE
Length of road	Length of road object	DOUBLE

Then, matching pairs of road centerlines in the two maps at different scales are detected. From the matching pairs, the road centerlines on a large scale map are divided into two groups: centerlines that exist on both maps for both scales and centerlines that exist only on the larger scale map. Assuming that the generalized result for the large-scale map serves as the road data for the small-scale map, the centerlines selected and eliminated during the generalization process are therefore distinguished. The detection process for the matching pair of road centerlines for the two scale levels is shown in Fig. 2.

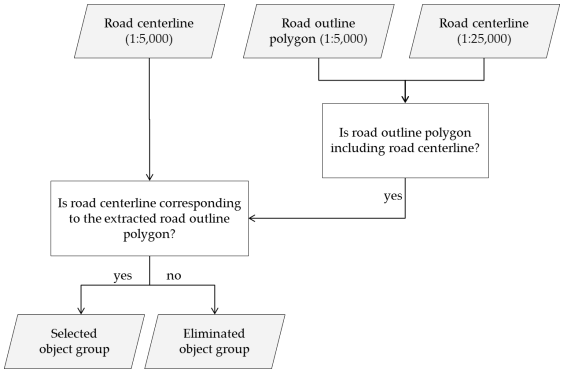

Fig. 2. Extraction process for matching pair of road centerlines between different scales

The network matching methodology at this stage is buffer polygon-based matching, which is generally applied. At first, using buffer polygons, road outline polygon objects on a 1:5,000 scale and road centerlines on a 1:25,000 scale are super imposed. The road outline polygons on the 1:5,000 scale that include the centerline of the 1:25,000 scale are extracted. The road centerlines at the 1:5,000 scale, which correspond to the extracted outline polygons, are classified by the selected object group, and the other centerlines are classified by the eliminated object group.

2.2 Statistical analysis of characteristics

As the first statistical analysis process for determining the criterion for the appropriate number of segment in road networks, the Radical Law, modified to use the length of the network instead of the number of objects (Park *et al.*, 2012), is applied to the network, and the constant for symbolic exaggeration is calculated (Park *et al.*, 2012; Töpfer and Pillewizer, 1966; Wilmer, 2010). The constant of symbolic exaggeration is an additional factor for the Radical Law used to address the difference between the appropriate number of features and the actual features on a map (Joao, 1998; Wilmer, 2010). The equation for the modified Radical Law for the road network selection is as follows (Eq. (1)):

$$n_f = n_a \times C \times \sqrt{M_a / M_f} \quad (1)$$

Where n_f is the length of the road objects shown at the target scale, n_a is the length of road objects shown on the source map, M_a is the scale denominator for the source map, M_f is the scale denominator for the derived map, and C is the constant for symbolic exaggeration.

To determine the influence of each road attribute on the selection, statistical analysis of the relationship between the selected/eliminated groups and the attribute values is performed. A multiple regression model, multi-criteria decision analysis, classification of machine learning, and data mining can be used as methodologies to analyze the characteristics and to group the objects. In this study, a qualitative response regression model in econometrics is used as the analytical methodology. The qualitative response regression model is a type of probability model that calculates

the probability that a dependent variable will be '1' according to the independent variables when the dependent variable is a binary or dichotomous variable (e.g., Y = 1 if something happens and Y = 0 if nothing happens). For this model, the LPM (Linear Probability Model), Logit model, and Probit model were developed (Gujarati and Porter, 2009).

Among these models, the Logit model is adapted in this study to analyze the relationship between the object groups and the attribute information for the objects. The Logit model is a statistical methodology for measuring the probability of an event using a linear combination of the independent variables and is called a logistic regression. By calculating the probability of the dependent variable to equal '1', while LPM utilizes the probability function of the linear form, the Logit model applies a function of a nonlinear form, which is similar to a cumulative distribution function and enhances the power of explanation.

The expressions in the Logit model are as follows (Eqs. (2) and (3)):

$$P_i = E(Y_i = 1 \mid X_i) = \frac{1}{\{1+e^{-(\beta_1+\beta_2 X_i)}\}} \tag{2}$$

$$L_i = \ln\left(\frac{P_i}{1-P_i}\right) = \beta_1+\beta_2 X_i \tag{3}$$

where, Y_i is the dependent variable, X_i is the independent variable, P_i is the probability of Y_i to equal '1', β is the regression coefficient, and L_i is the odds ratio.

3. Experiment

3.1 Data training

Fig. 3. Road centerline of digital topographic map at 1:5,000 (left) and 1:25,000 (right) scale for target area of training data (all over Giheung-gu, Yongin-si)

The dataset for the experiment is the road centerline data of the digital topographic maps at a 1:5,000, 1:25,000 scale of two neighboring area in Suwon-si and Yongin-si. The maps of Yongin-si were used for training data as shown in Fig. 3. The relationship between the selected object group and the attributes such as road length, width and class is derived as Eq. (4).

$$P_i = \frac{1}{\left\{1+e^{-(\beta_1+\beta_{width}Width_i+\beta_{length}Length_i+\beta_{class}Class_i)}\right\}} = \frac{1}{\left\{1+e^{-(-7.5165+0.029\times Width_i+0.007\times Length_i+0.2182\times Class_i)}\right\}} \tag{4}$$

The *t*-test was performed on the analysis result of the Logit model as follows. In this model, as the number of samples approaches 917, the degree of freedom becomes infinite, and the threshold of the *t* value is 1.96 at a 95% confidence interval. Because the t values for each regression coefficient are -13.0401, 3.7460, 3.1291, and 10.1811, all absolute values are higher than the threshold of the *t* value. Therefore, all variables are statistically significant. With the P_i value, the SPF (Selection Probability Function) for all road centerlines in the test dataset is calculated with Eq. (4) and Fig. 3 shows the histogram of the calculated SPF values.

Based on the SPF values, the objects that have a high probability should be selected, and the objects with a low probability should be deleted according to the target scale level. To determine whether an object should be selected or eliminated, establishing a threshold SPF value is necessary. In this study, a method to determine the threshold SPF value according to the target scale level using the modified Radical Law is developed. Namely, the total length of a line from the test map at a 1:5,000 scale, the constant for symbolic exaggeration derived in Section 2, the scale denominator for the source map and the target scale are applied to the Radical Law. The total length of the line for the result map at the target scale is calculated.

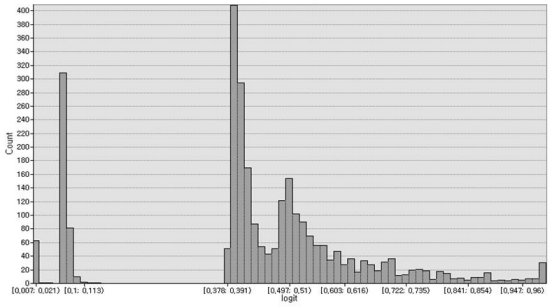

Fig. 4. Histogram of SPF value for road network of test data

After developing a cumulative function around the SPF value and the length of the centerline objects, the threshold for the SPF value is calculated by applying the total length of the target line to the cumulative function graph. Fig. 5 describes the process of calculating the threshold for the SFP value using the total length of the target line and the cumulative function for the SPF-length of the line.

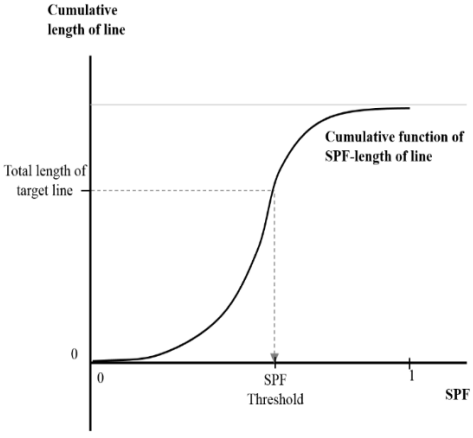

Fig. 5. Threshold estimation using total length of target line, the cumulative function of SPF and length of line

3.2 Evaluation

For quantitative evaluation of the selected result, the matching ratio is measured using the matching results between the road centerlines for the selected results and those in the NGII (National Geographical Information Institute) map. Table 2 shows the total length of the road network for the matching results between the selected road centerlines from the two methods and the NGII map data.

Table 2. Total length of the road network for the matching results between the selected road centreline from the two methods and the NGII map data

	Total Length of Network (m)	Total Length of Matched Objects (m)	Matching Ratio (%)
Selected result of the proposed method	295,389.3	245,207.8	81.66

From the above table, the performance of the selected result for the road centerline is calculated using the matching ratio (precision) (Eq (5)).

$$\text{matching ratio} = \frac{\text{Total length of matched objects in the selected road segments}}{\text{Total length of selected road segments}} \times 100\% \quad (5)$$

The matching ratio was calculated as 81.66% for the selected result from the proposed method. This means that the overall network structure is largely matched, and some unmatched portions are observed. These unmatched portion may be categorized into the following three cases: (a) discrepancy due to the gap in the update period, (b) discrepancy due to the gap in the method which expresses a dual line as a single line, and (c) discrepancy due to the excessive expression of road centerline data. The figures below (Figs. 6, 7 and 8) are enlargements illustrating examples of the unmatched cases discussed above. The problem in case (a) can be resolved if the latest renewal is performed. The problem in case (b) can be resolved if the dual line is converted to a single line. In case (c), the road objects are not eliminated due to the simple selection rules using length, width and class.

Fig. 6. An example of the discrepancy caused by gaps in the update period (the matched line (black) and commission error (gray) of the proposed method, The vertical road in the middle: a road object that was inserted after renewal of the NGII digital map)

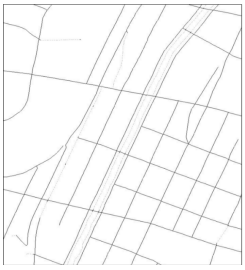

Fig. 7. An example of the discrepancy caused by gaps in the method that expresses a dual line as a single line (the matched line (black) and commission error (gray) of the proposed method)

Fig. 8. An example of the discrepancy caused by excessively expressed small road centerlines (the matched line (black) and commission error (gray) of the proposed method)

4. Conclusion

In this study, a methodology that selects or eliminates road network data according to the target scale level is proposed and examined as a step toward the generalization of digital map data. The criteria for the selection of the road network object are analyzed by comparing an existing digital map at the 1:5,000 scale and the 1:25,000 scale and by applying Töpfer's Radical Law and the Logit model. The selection model, which is drawn from the results of the analysis, is applied to the test map data, and the generalized 1:25,000 scale network dataset is derived from the 1:5,000 road centreline layer.

For the evaluation of the proposed method, the selection result obtained using the rule-based road selection method is compared with the result obtained using the proposed method. The 1:25,000 digital map produced by NGII is used as a benchmark for the qualitative and quantitative evaluation of the two results from the road network selection methods. The result of the quantitative and qualitative evaluation for the generalized output showed that 81.66% of road centerlines in the proposed method were matched with the digital map produced by NGII and this value is higher than that of the rule-based method.

This study has some limitations. Preserving network topology is not discussed in this study, although it an important factor in network generalization. In addition, a greater variety of test sites may be considered to validate the methodology of this study. Moreover, an analytic comparison with other methods for network thinning is not conducted in this study. These limitations should be addressed in future work.

References

Chen, J., Hu, Y., Li, Z., Zhao, R., and Meng, L. (2009), Selective omission of road features based on mesh density for automatic map generalization, *International Journal of Geographical Information Science*, Vol. 23, No. 8, pp. 1013–1032.

Choe, B. and Kim, Y. (2007), Framework and workflows for spatial database generalization, *Transactions in GIS*, Vol. 11, No. 1, pp. 101–114.

Gujarati, D.N. and Porter, D.C. (2008), *Basic Econometrics*, *5th Edition*, McGraw-Hill, New York.

Joao, E.M. (1998), *Causes and Consequences of Map Generalization*, Taylor & Francis, London.

Lee, J., Jang, H., Yang, J., and Yu, K. (2017), Machine learning classification of buildings for map generalization, *ISPRS International Journal of Geo-Information*, Vol. 6, No. 10, pp. 309-324.

Li, Z. and Choi, Y. (2002), Topographic map generalization: association of road elimination with thematic attributes, *The Cartographic Journal*, Vol. 39, pp. 153-166.

Liu, X., Zhan, F., and Ai, T. (2010), Road selection based on Voronoi diagrams and "strokes" in map generalization, *International Journal of Applied Earth Observation and Geoinformation*, Vol. 12, pp. 194-202.

Mackaness, W.A. and Mackechnie, G.A. (1999), Automating

the detection and simplification of junctions in road networks, *GeoInformatica*, Vol. 3, No. 2, pp. 185-200.

Park, W., Ji, B., and Yu, K. (2012), Control methodology of map generalization scale using Töpfer's Radical Law, *Proceedings on 2012 Spring Conference of Korea Society of Surveying, Geodesy, Photogrammetry, and Cartography*, KSGPC, 26-27 April, Gangneung, South Korea, pp. 309–311.

Thomson, R.C., and Richardson, D.E. (1995), A graph theory approach to road network generalisation, *Proceeding of the 17th International Cartographic Conference*, ICC, 3–9 September, Barcelona, Spain, pp. 1871–1880.

Töpfer, F. and Pillewizer, W. (1966), The principles of selection. *The Cartographic Journal*, Vol. 3, No. 1, pp. 10–16.

Wilmer, J.M. (2010), *Application of the Radical Law in Generalization of National Hydrography Data for Multiscale Mapping*. Ph.D. dissertation. The Pennsylvania State University, Pennsylvania, USA.

Zhang, Q. (2005), Road network generalization based on connection analysis. In: Fisher, P. (ed.), *Developments in Spatial Data Handling*, Springer Science & Business Media, Berlin, Germany. pp. 343-353.

Zhou, Q. and Li, Z. (2011), Evaluation of properties to determine the importance of individual roads for map generalization. *Advances in Cartography and GIScience*, Vol. 1, pp. 459-475.

Zhou, Q. and Li, Z. (2014), Use of artificial neural networks for selective omission in updating road networks, *The Cartographic Journal*, Vol. 51, pp. 38–51.

Zhou, Q. and Li, Z. (2017), A Comparative Study of Various Supervised Learning Approaches to Selective Omission in a Road Network, *The Cartographic Journal*, Vol. 54, No. 3, pp. 254-264.