The Cartographic Journal *Vol. 53 No. 1 pp. 31–42 February 2016*

REFEREED PAPER

Identification and Utilization of Land-use Type Importance for Land-use Data Generalization

Wenxiu Gao[1], Alfred Stein[2], Li Yang[3], Jianguang Hou[4], Xiaojing Wu[2] and Xiangchuan Jiang[5]

[1] State Key Laboratory of Information Engineering in Survey, Mapping and Remote Sensing, Wuhan University, No. 129, Luoyu Road, Wuhan 430079, China. [2]Earth Observation Science, International Institute for Geo-Information Science and Earth Observation, PO Box 6, Hengelosestraat 99, 7500 AA Enschede, The Netherlands. [3]Huawei Technologies Co., Ltd, China. [4]East China Mineral Exploration and Development Bureau, Huadong Mansion, No. 26, Daguang Road, Nanjing 210007, China. [5]Changsha Land & Resource Bureau, No. 238, Eastlaodong Road, Changsha 410001, China
Email: wxgao@lmars.whu.edu.cn

A proper characterization of land-use types is critical for constructing generalization constraints to guide and control land-use data generalization. This paper focused on identification and utilization of their importance based upon land-use distributions and application themes. First, this importance was identified using a three-step method that links a diversity index, a multiple attribute decision model and a spatial association analysis. Second, with the importance, a mathematical function was designed to determine minimum area thresholds of land-use polygons as an example of generalization constraints. Third, the importance was used to assist in the selection of generalization operators and evaluation of generalization outcomes. Fourth, a land-use dataset at 1:10 000, describing the land use of a typical rural area in Hubei province of China, was generalized towards a 1:50 000 dataset to verify the effects of the presented method and function. Three additional tests were implemented to analyze the sensitivity of the importance of land-use types on setting the minimum area threshold and generalization operations. The outcome showed that the proposed methods and functions make land-use data generalization more adaptable for in-use datasets and applications.

Keywords: land-use type importance, land-use data generalization, generalization constraint, multi-attribute decision model

INTRODUCTION

Multi-scale or multi-theme land-use maps are widely used as a primary information source in agricultural planning, disaster monitoring and environment protection. These maps can be produced from generalizing high-resolution land-use datasets. Typically, a land-use map represents the spatial variation of land-use types with space-exhaustive polygons. Polygons that share one or more boundaries represent contrasting land-use types (Castilla and Hay, 2007; Gao *et al.*, 2004b). Land-use data generalization, therefore, should follow similar principles and constraints as polygon generalization (Oosterom, 1995; Bader and Weibel, 1997; Cheng and Li, 2006; Zhang *et al.*, 2011). In addition, land-use data generalization has to consider established land use principles in order to make the generalized results adaptive to specific land-use characteristics and application requirements.

Polygon generalization involves the transformations of a polygonal subdivision obeying semantic and geometric aspects (McMaster and Shea, 1992; Muller and Wang, 1992). During land-use data generalization, semantic transformation occurs in two cases. The first case is to shift land-use types from a higher to a lower classification level in order to produce a finer scale land-use map that is of an insufficient space to express all land-use types at the same levels as the original map. In such cases, only significant land-use types are retained, while land-use types of less relevance are shifted to general levels. It secondly occurs when deriving an application-specific map. In doing so, the characteristics of some land-use types closely related to the specific application theme are highlighted and emphasized, while others are constricted. In both cases, however, semantic changes should be be minor, in order to display the essential characteristics of land use (Haunert and Wolff,

DOI: 10.1179/1743277413Y.0000000054

2010), especially if such land-use types are considered to be important. Next, a geometric transformation can reduce the geometric complexity of the emerging land-use polygons, especially to remove small polygons which are not readable in a target map (Gao *et al.*, 2012b). Such a geometric transformation varies according to different land-use types according to their importance. The spatial structures and patterns of polygons belonging to important land-use types should be preserved, while polygons of the non-important land-use types are to be largely simplified. For this reason, generalization constraints must be constructed on both semantic and geometric aspects based upon the importance of land-use types for data generalization.

The importance of land-use types is determined synthetically by various factors related to natural and socio-economic characteristics of land-use patches. The dominant land-use types are able to be identified by using a dominance index or diversity index based on the ratio of land-use area of each type (O'Neill *et al.*, 1988; Riitters *et al.*, 1995). This paper aims to employ a multiple attribute decision model (MADM) to determine the importance of land-use types under the joint consideration of natural and socio-economic factors related to land use.

The purpose of identifying the importance of land-use types is to construct generalization constraints used to control the generalization process. Generalization constraints reflect cartographic principles and related professional applications (Beard, 1991; Muller and Wang, 1992). Such generalization constraints can be acquired from the analysis of existing map series and texts, or interviews with experts (Edwardes and Mackaness, 2000; Kilpelainen, 2000). For instance, a minimum area threshold is widely used to constrain polygon objects to remain visible by the human eye (Muller and Wang, 1992). Minimum area thresholds are set based upon empirical values or examples from the literature. Some authors propose the use of a single minimum area threshold for identifying all invalid small objects (Muller and Wang, 1992; Cheng and Li, 2006), while others adopt different thresholds for different types of objects (Mackaness *et al.*, 2008). For land-use data generalization, however, such thresholds should differ for different land-use types as some are considered to be more important than others (Haunert and Wolff, 2006). An ideal routine is to construct generalization constraints directly based upon an in-use dataset. Such routines can make generalization constraints adaptive to land-use characteristics described in the dataset.

The aim of this paper is to present a three-step method to identify the importance of land-use types. First, a dominance index initially investigates the general distribution of the areas covered by different land-use types. Second, the importance of each land-use type is quantified according to MADM with specific factors. Third, the preliminary importance of land-use types is adjusted based upon the spatial association between land-use types. The adjusted importance values are further partitioned into several grades that are expressed as ordinal variables, called the importance ranks. With the importance ranks, a mathematical function is designed to determine the minimum area thresholds of land-use types at each rank. Fourth, the importance and the minimum area thresholds are utilized during generalization of land-use data. One experiment is presented to demonstrate the functions of the land-use type importance, and three additional experiments are carried out to investigate the influences of such importance. The generalized outcomes are evaluated according to the semantic and geometric aspects of land-use polygons.

METHODOLOGY

Initial identification of dominant land-use types

In this study, we consider the dominance index and Shannon's Evenness Index. They can both measure the extent to which one or a few categories dominate the landscape in a categorical dataset (O'Neill *et al.*, 1988; Riitters *et al.*, 1995). They have been widely used in the research of landscape diversity and biodiversity (Martinez *et al.*, 2010; Hietala-Koivu *et al.*, 2004). This study adopts the dominance index developed by O'Neill *et al.* (1988) since it describes characteristics of land-use distribution

$$D = \ln(m) + \sum_{i=1}^{m} P_i \ln(P_i) \quad (1)$$

where m is the total number of land-use types and P_i is the area proportion of land-use type i. Since $P_i<1$ and hence $\ln(P_i)<0$, the summation yields negative values. When all land use types are present in equal proportions, the term $\ln(m)$, represents the maximum of D. If the landscape is dominated by one or a few land-use types then the value of D is close to 0.

In our study, D is used to provide the first overview of land-use characteristics at the landscape level, that is, the landscape is dominated by a few land-use types or the landscape is divided into approximately equal proportions belonging to many land-use types. Dominant land-use types may be of high importance for some studies. In addition, D is used to compare the difference between the original land-use map and the generalized land-use maps at the evaluation stage.

Quantify importance of land-use types using MADM

A single index provides inadequate information about spatial patterns of a landscape (Saura and Martinez-Mlian, 2001). The MADM is widely used as an aid in decision-making, which seeks to integrate objective measurements with subjective judgment (Belton and Stewart, 2002). It is employed to integrate natural and cultural attributes of land-use for identifying the importance of land-use types. There are three basic steps through this model, (1) selecting attributes, (2) setting weights of attributes, and (3) calculating importance values of land-use types.

Step 1: Selection of attributes

Considering land-use characteristics and the purpose of land-use data generalization, we chose four attributes to build the MADM model: area ratio (AR), patch number ratio (NR), theme-relevant degree (TD) and economic

value (*EV*) of each land-use type. The first two attributes present the spatial configuration of land-use types in a region. The third attribute reflects the degree of association between a land-use type and the target application theme. If the motivation of land-use data generalization is to derive a dataset for a specific-theme application, land-use types closely related to that target theme are relatively important. The last attribute is about socioeconomic characteristics reflecting the status of a land-use type on human life. The four attributes jointly describe the characteristics of land-use from the natural, socioeconomic, and application-oriented points of view, respectively.

The attributes *AR* and *NR* can be directly calculated from the area and number of land-use patches of each land-use type. *TD* is determined by a subjective judgment depending upon people's understanding of the target application and related professional knowledge. A two-step method can assist in determining its value. First, experts select the land-use types which have direct and close relationships with the target theme. For example, if land-use data generalization is used to extract general land-use information for agriculture in a region, then the typical agricultural land-use types such as irrigated paddy and orchard will have the highest priority to be the important types. Second, the importance of other land-use types is determined according to their own characteristics and spatial association with the important land-use types identified in the first step. *EV* of a land-use type largely depends upon the specific application. For an agricultural example *EV* can be derived as an economic statistic, e.g. the percentage in the economic revenue of the benefits yielded by a particular land-use type. If an application, however, requires urbanized characteristics, then cultural value, e.g. the historical value of buildings or gardens, can reflect the socioeconomic status of a land-use type.

Step 2: Weights of attributes

Mostly, the impacts of selected attributes are unequal and vary with different cases. Therefore, a weight is set for each selected attribute according to the dataset in use, the target application and land-use characteristics of an area. If the purpose of data generalization is to produce a smaller scale map without further requirements, area ratio and patch number ratio can share the weights without considering the other two attributes. If data generalization aims to extract general land-use information for analysis of agricultural economy, then the theme-relevant degree should have the highest weight and the socioeconomic value should be considered.

The way we choose in this paper for determining the weights is to conduct tests with different weight values and then choose appropriate values. This approach is a so-called return task (Podolskaya *et al.*, 2007). This results into a weight vector $\mathbf{w}=(\omega_1, \omega_2, \omega_3, \omega_4)^{\mathrm{T}}$, $(0\leq\omega\leq1)$, that is introduced to MADM.

Step 3: Importance value of land-use types

Let m be the number of land-use types. Then the MADM decision matrix Υ is constructed on the basis of the four attributes

$$\Upsilon=\begin{bmatrix} AR_{11} & AR_{12} & \cdots & AR_{1\mathrm{m}} \\ NR_{21} & NR_{22} & \cdots & NR_{2\mathrm{m}} \\ TD_{31} & TD_{32} & \cdots & TD_{3\mathrm{m}} \\ EV_{41} & EV_{42} & \cdots & EV_{4\mathrm{m}} \end{bmatrix} \tag{2}$$

Since the units of the four attributes are different, a min–max normalization is employed to perform a linear transformation on this decision matrix to the range [0, 1] (Han and Kamber, 2006a). Taking *AR* of the *i*th land-use type as an example, we then obtain

$$r_{1\mathrm{i}}=\frac{AR_{1\mathrm{i}}-\mathrm{Min}AR}{\mathrm{Max}AR-\mathrm{Min}AR} \tag{3}$$

Here, Min*AR* and Max*AR* are the minimum and maximum *AR*, respectively. The normalized decision matrix denoted as **R** equals

$$\mathbf{R}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1\mathrm{m}} \\ r_{21} & r_{22} & \cdots & r_{2\mathrm{m}} \\ r_{31} & r_{32} & \cdots & r_{3\mathrm{m}} \\ r_{41} & r_{42} & \cdots & r_{4\mathrm{m}} \end{bmatrix} \tag{4}$$

The final decision vector is the product of **w** and the matrix **R**

$$\mathbf{P}=\mathbf{w}^{\mathrm{T}}\mathbf{R}=(I_1,I_2,\cdots,I_{\mathrm{m}}) \tag{5}$$

where I_{i} is the importance value of the *i*th land-use type. A large I_{i} corresponds to a high importance of the *i*th land-use type. We will consider the importance values sorted from the largest to the smallest and denote them as a set of variables (P_1, P_2,..., P_{m}).

Adjustment of preliminary importance ranks

Commonly, the distribution of land-use types is spatially associated with other land-use types (Walsh *et al.*, 2003; Han and Kamber, 2006b). For instance, the distribution of irrigated ploughs depends upon water bodies as indispensible water resources; naturally grasslands distribute along the ridge of forestlands. Such spatially-associated relationships must be taken into account when determining the importance of land-use types. If a land-use type has a low importance according to the MADM model but it has a very strong spatially-associated relationship with the land-use types at the top of the importance values in the vector **P**, then the importance of this land-use type should be adjusted according to such relationship.

In this study, a spatial-association relationship is defined as the spatially touching relationship between land patches of different land-use types. It is measured with the following equation according to association rules analysis

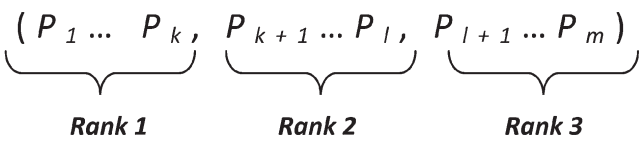

Figure 1. The importance ranks of the different land-use types

$$c_{ij} = \frac{\sigma(\{T_i \text{ touching with } T_j\})}{\sigma(\{T_i\})} \quad (6)$$

here, c_{ij} is the confidence of the patches of one land-use type (T_i) touching the patches of land-use type (T_j). In particular, T_j corresponds to the land-use types at the top of the importance values. $\sigma(\{T_i$ touching with $T_j\})$ is the support of the number of the patches of T_i which touch with the patches of T_j. If c_{ij} is larger than a specific minimum support, then land-use type T_i is entitled to analyze the spatial association with the important types.

Considering the spatial association, the importance value P_i of the ith land-use type is adjusted as P_i' equals

$$P_i' = P_i + \sum_{j=1}^{k} P_j \cdot c_{ij} \quad (7)$$

where k is the number of the land-use types at the top of the importance values in P and P_j is the importance value of the jth important land-use. Adjustment is based upon the summed relationship between the ith land-use type and all important land-use types. If the distribution of the patches of the ith land-use type is highly associated with that of the important land-use types, its importance value may be improved greatly. After the adjustment, the importance values in P are graded into ranks, called importance ranks (Figure 1).

UTILIZATIONS OF LAND-USE TYPE IMPORTANCE

This section presents three utilizations of the land-use type importance in land-use data generalization: (1) determination of the minimum area thresholds; (2) assistance in selecting generalization operators for dealing with the patches smaller than the area thresholds (denoted as small-area patches) and (3) evaluation of the generalized outcomes of the small-area patches.

Determination of minimum area thresholds

A minimum area threshold is set for the land-use types of each importance rank (shown in Figure 1) with the following equation

$$MA_i = A_{min} + \frac{Rank_i - 1}{n-1} \cdot (A_{max} - A_{min}) \quad (8)$$

where, MA_i is the minimum area threshold of the land-use types at rank i and $Rank_i$ is the ordinal value of the importance ranks (e.g. $Rank_1=1$, $Rank_2=2$ and $Rank_3=3$ in Figure 1), and n is the count of the importance ranks (e.g. 3 in Figure 1). A_{min} and A_{max}, are constants which limit the thresholds in a specific range. That is to say, the threshold of the most important land-use types (i.e. $Rank_1=1$) is larger than A_{min} and that of the least important land-use types (e.g. $Rank_1=3$) is smaller than A_{max}. Such limitations will avoid leading too dense patches due to preserve too many patches or losing too much information due to elimination of too many patches on a target map.

Selection of generalization operators

In polygon maps, small-area polygons are one of the geometric conflicts which must be resolved in map generalization (Gao *et al.*, 2012a; Muller and Wang, 1992; Peter, 2001; Bader and Weibel, 1997). In this study, a procedure was designed to implement the generalization of small-area land-use patches (Table 1). This procedure integrates some ideas from (Bader and Weibel, 1997; Mackaness *et al.*, 2008; Muller and Wang, 1992; Gao *et al.*, 2012a; Haunert and Wolff, 2010). It provides a routine to introduce the minimum area thresholds and the importance ranks of land-use types for dissolving small-area geometric conflicts. The purpose is to demonstrate the functions and validities of the land-use type importance determined with the proposed methods.

The basic objective of the procedure is to preserve the characteristics of the most important land-use types (i.e. those in the 1st rank) as much as possible by limiting the changes of the semantic characteristics (i.e. land-use types) and spatial contexts around the important focal patches. For this reason, the processes implemented on the land-use patches in the 1st rank are different from those on the patches in the other ranks.

In this procedure, spatial context is the key indicator for selecting generalization operators.

- *Spatial context 1* indicates that a cluster pattern may exist around the focal patch. The aggregation with the nearest adjacent patch is intended to preserve the cluster patterns and limit changes of the spatial context and semantic characteristic.
- *Spatial context 2* considers for preserving the semantic characteristics, i.e. the land area of the same super-types remains unchanged as much as possible.
- For *Spatial context 3*, an isolated patch from its congeneric patches is merged with the touching patch having the longest common boundary. This may avoid shifting of a small-area conflict to another geometric conflict, i.e. those leading to an unreadable narrow gap between boundaries of a patch (Gao *et al.*, 2012a; Bader and Weibel, 1997).

Operations on *Spatial context 1* are more complicated than those on the other two contexts. An approach was developed in the past to combine an outward buffering with a proximity index to search the adjacent patch around a focal patch and further combine an outward buffering with an inward buffering to construct the aggregated region for the aggregating operation (Gao *et al.*, 2012a). To save computing time, therefore, the minor important

Table 1. A procedure for generalizing small-area land-use patches

Generalization processing on small-area patches
Step 1: Search all patches smaller than the corresponding *MA* of a concrete land-use type *(T)* and put them into a patch collection with the order of the size from the smallest to the largest.
Step 2: Check the spatial context of each patch (called as the focal patch) sequentially in the collection.
Search touching patches which have common boundaries with the focal patch.
Search adjacent patches which have no common boundary with the focal patch but within a certain searching radius.
Step 3: Select an appropriate operator to generalize the focal patch based on its spatial context.
(1) When *T* is in the 1st rank, the following three spatial contexts around the focal patch are considered in sequence:
Spatial context 1:
IF: the focal patch has one or more adjacent patches belonging to the same land-use type
THEN: aggregate the focal patch with the single or the nearest adjacent patch
Spatial context 2:
IF: the focal patch has one or more touching patches belonging to the same super-type
THEN: merge the focal patch with the single or the smallest touching patch and the new patch inherits the land-use type from the touching patch.
Spatial context 3:
IF: the above two spatial contexts are not satisfied
THEN: merge the focal patch with its touching patch with the longest common boundary and the new patch inherits the land-use type from the touching patch.
(2) When *T* is not in the 1st rank, the following three spatial contexts around the focal patch are considered in sequence:
Spatial context 2:
IF: the focal patch has one or more touching patches belonging to the same super-type
THEN: merge the focal patch with the single or the smallest touching patch and the new patch inherits the land-use type from the touching patch.
Spatial context 1:
IF: the focal patch has one or more adjacent patches belonging to the same land-use type
THEN: aggregate the focal patch with the single or the nearest adjacent patch
Spatial context 3:
IF: the above two spatial contexts are not satisfied
THEN: merge the focal patch with its touching patch with the longest common boundary and the new patch inherits the land-use type from the touching patch.
Step 4: Check whether there are any touching patches belonging to the same land-use types after the above steps. If it is true, merge them to satisfy the basic principle of polygon maps.
Step 5: Repeat the above steps for each land-use type.

land-use types in the 2nd and 3rd ranks, are therefore generalized starting from *Spatial context 2*.

For an efficient implementation, the generalization of small-area land-use patches starts from the least important land-use types, because the spaces released from the least important land-use patches may solve the geometric conflicts of the more important land-use patches synchronously. It can save time for generalization and also preserve the important characteristics changed as little as possible.

EVALUATION OF GENERALIZATION RESULTS

Since land-use data generalization includes both geometric and semantic transformations, the generalized outcomes need to be evaluated on both geometric and semantic aspects. For the geometric aspects, the comparison of the dominance index before and after generalization indicates the changes of the overall land-use structure on the landscape. With the elimination of small-area patches, the index value should be decreased because the differences in land-use patch areas are reduced. In addition, the changes of in area ratio and patch number ratio of land-use types in each importance rank also present the effects of geometric transformations and influences of land-use type importance.

For semantic aspects, we followed the basic idea of semantic similarity based on a classification hierarchy (Liu *et al.*, 2002) and combined with the method of semantic accuracy proposed by Cheng & Li (2006) to evaluate the semantic changes during data generalization. Suppose that a small patch O_S releases its space to patch O_A in line with the above procedure. The semantic accuracy of the resulting patch $O_{A'}$ belonging to type C_A is determined as

$$\mu_{O_{A'}}^{C_A} = \frac{\mathrm{Area}(O_A)\cdot\mu_{O_A}^{C_A} + \mathrm{Area}(O_S)\cdot\mu_{O_S}^{C_A}}{\mathrm{Area}(O_A) + \mathrm{Area}(O_s)} \tag{9}$$

where, $\mu_{O_A}^{C_A}$ is the membership value of patch O_A belonging to type C_A and $\mu_{O_S}^{C_A}$ is the membership value of patch O_S belonging to type C_A. If $\mu_{O_A}^{C_A}$ equals 1, then this indicates that $O_{A'}$ has the same semantics as O_A. If, however, $\mu_{O_A}^{C_A}$ is close to 0, then the change becomes larger while the quality of the generalization is relatively low.

According this equation, the value of $\mu_{O_{A'}}^{C_A}$ is mainly determined by $\mu_{O_s}^{C_A}$, while $\mu_{O_s}^{C_A}$ depends upon the relationship between the land-use types of O_S and O_A. If both belong to the same land-use type, then $\mu_{O_s}^{C_A}=1$ and $\mu_{O_{A'}}^{C_A}=1$. If O_S and O_A belong to the same super-type described as *Spatial context 2*, then $\mu_{O_s}^{C_A}$ should be larger than in the case of *Spatial context 3*.

After dealing with small patches, the semantic accuracies at the map level (μ_r^M) and the class level ($\mu_r^{C_j}$) are calculated as equations (10) and (11), respectively.

$$\mu_r^m = \frac{\sum_{j=1}^{m}\sum_{i=1}^{N_j}\mathrm{Area}(O_i)\cdot\mu_{O_i}^{C_j}}{\sum_{j=1}^{m}\sum_{i=1}^{N_j}\mathrm{Area}(O_i)} \tag{10}$$

$$\mu_r^{C_j} = \frac{\sum_{i=1}^{N_j}\mathrm{Area}(O_i)\cdot\mu_{O_i}^{C_j}}{\sum_{i=1}^{N_j}\mathrm{Area}(O_i)} \tag{11}$$

Here, N_j is the number of land-use patches of the land-use type *j*, while *m* is the number of the land-use types in a land-use dataset.

CASE STUDY

Data sources

The methods and procedures were applied to a land-use dataset describing a typical rural area at a scale of 1:10 000, located in the Hubei province, China (Figure 2). The area covers a landscape of approximately 52,000 km^2 with 36 land-use types at the tertiary class of Chinese land-use classification hierarchy issued on January 1, 2002 (Table 2)

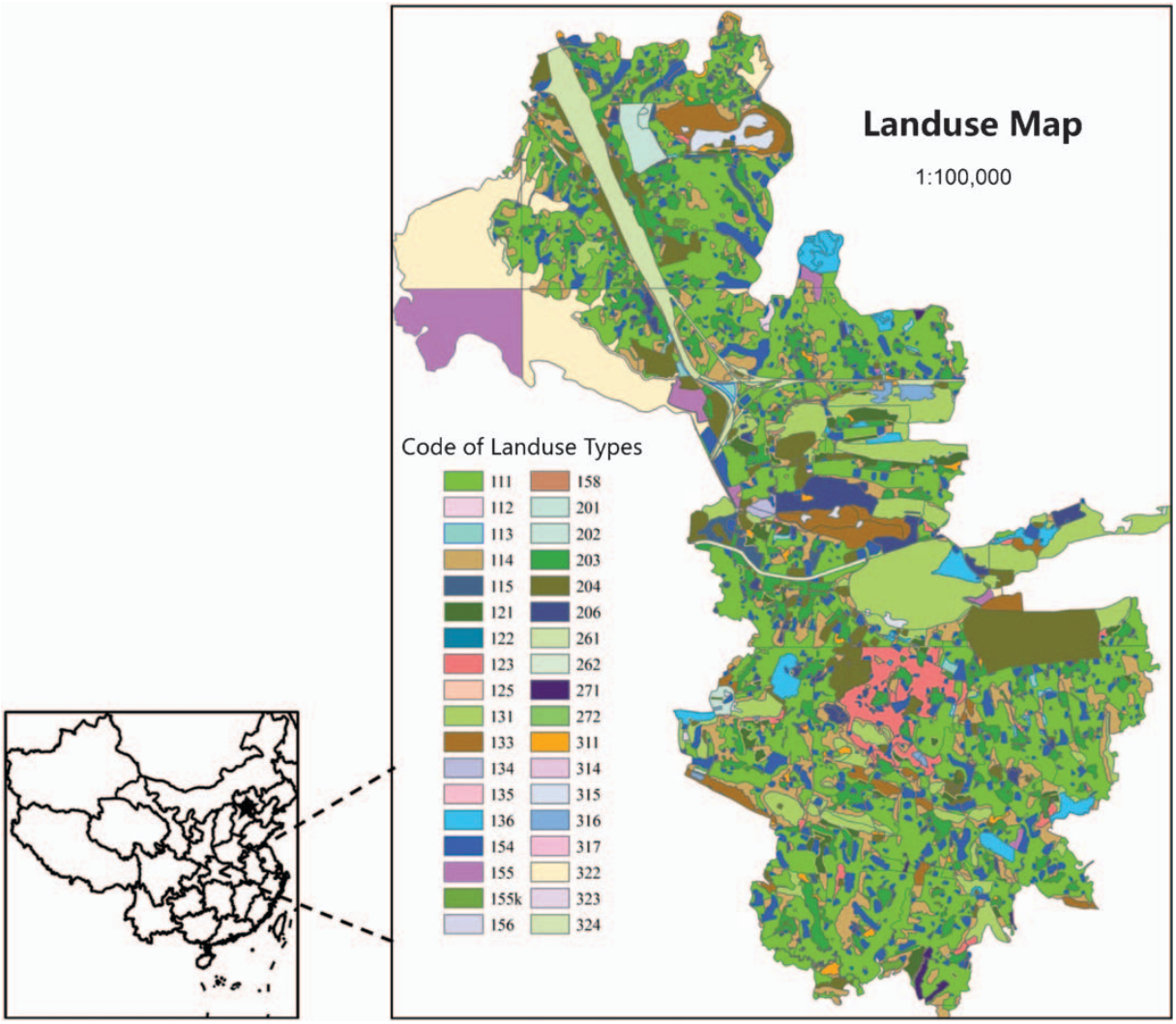

Figure 2. Study area and land-use map with 36 land-use types (the corresponding names of the codes are listed in Table 2)

The aim of this case study was to produce a land-use dataset at a scale of 1:50 000, using the generalization of the 1:10 000 test map. This generalization was intended to display the general land-use characteristics on a scale-reduced map without additional requirements for any land-use themes or applications. The test area is of particular interest, because the land-use types are diverse, including both natural and developed ones. According to the objectives of the case study and according to the principles for selecting attributes in MADM, the degree of theme-relevance and the socioeconomic values of land-use types were not involved in this test. Therefore, the two factors, 'area ratio' and 'patch number ratio', constitute the final MADM decision matrix.

Identification of dominant land-use types

Following equation (1), the dominance index D equals 1.13 on the landscape, while it ranges from 0.0 to 3.58 for the 36 land-use types. The value of 1.13 is relatively close to 0, indicating that the land-use types have a wide distribution on the landscape. To further investigate the distribution of D over the landscape, we ranked the areal proportions of the different land-use types into three classes using a natural break classification. Approximately 68% of the landscape is distributed over 13 land-use types, while the area of irrigated paddy (111) occupies 30% of the total area. The remaining 32% of the landscape is shared by the other 22 land-use types.

After various empirical tests, we set the weights equal to 0.7 and 0.3 for the area ratio and the patch number ratio, respectively. This resulted into

$$\mathbf{P}=(W_{\mathrm{AR}}\ W_{\mathrm{NR}})\cdot\begin{bmatrix} AR_{11} & AR_{12} & AR_{13} & AR_{14} & AR_{15} & \cdots & AR_{1n} \\ NR_{21} & NR_{22} & NR_{23} & NR_{24} & NR_{25} & \cdots & NR_{2n} \end{bmatrix}$$

$$=(0{\cdot}7\ 0{\cdot}3)\cdot\begin{bmatrix} 1 & 0{\cdot}101 & 0{\cdot}303 & 0{\cdot}034 & 0{\cdot}055 & \cdots & 0{\cdot}221 \\ 0{\cdot}305 & 0{\cdot}025 & 0{\cdot}508 & 0{\cdot}112 & 0{\cdot}065 & \cdots & 0{\cdot}025 \end{bmatrix}$$

$$=(0{\cdot}79,\ 0{\cdot}02,\ 0{\cdot}36,\ 0{\cdot}06,\ 0{\cdot}06,\ \ldots,\ 0{\cdot}16)$$

In a descending order, the vector $\mathbf{P}$ equals $\mathbf{P}=(0.79, 0.47, 0.37, 0.36, 0.25, \ldots, 0.0)$. From this we observe that the most important three land-use types are irrigated paddy

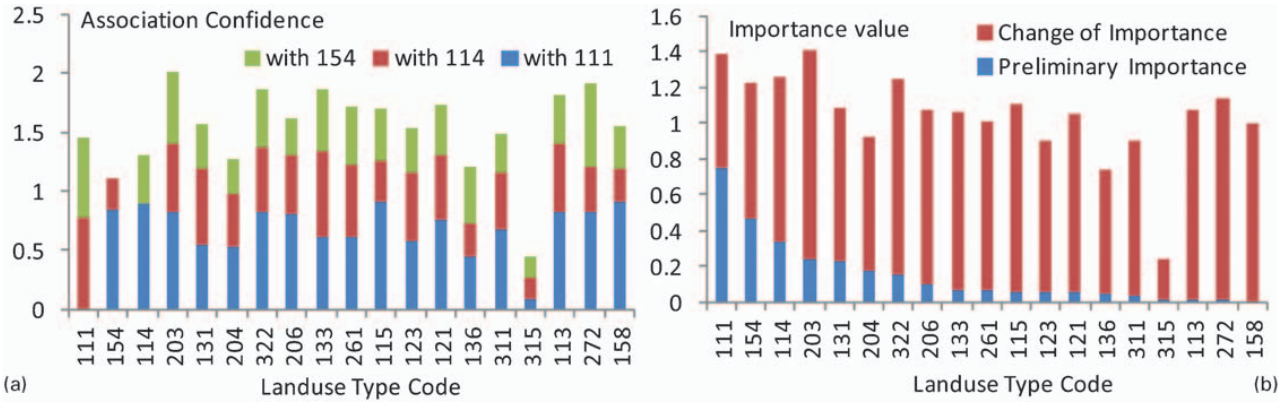

Figure 3. Association confidence with irrigable land (111), dry land (114) and pond for irrigation (154) (a) and importance values (b)

(111) with a P value of 0.79, non-irrigated farmland (114) with a P value of 0.47 and pond for irrigation (154) with a P value of 0.37, respectively. The importance of the other land-use types is then adjusted in terms of the spatial association with the three most important land-use types during the next step.

In the analysis of such spatial association, not all land-use types are necessary. Those land-use types with a low number of patches are directly grouped as the least important class. Here we took $\sigma(\{T_1\})<0.4\%$ as a criterion to identify a land-use type (T_i) belonging to the least important class. Next, association confidence of the other land-use types with the three most important land-use types is calculated with equation (6). Finally, the importance values of these land-use types are adjusted with equation (7).

Figure 3a illustrates the association confidence, and Figure 3b shows the preliminary importance of land-use types and their corresponding adjustments. The height of the blue, red and green blocks of each pillar in Figure 3a

Table 2. Classification hierarchy of land-use types with their corresponding codes (in part). The third column is the Tertiary Class with their corresponding codes which are used to describe land-use types of the test dataset in Figure 2

Primary class	Secondary class	Tertiary class (code)
Agriculture land	Arable land	Irrigated paddy (111)
		Natural paddy (112)
		Irrigable land (113)
		Non-irrigated farmland (114)
		Land for vegetable (115)
	Garden plot	Orchard (121)
		Land for mulberry field (122)
		Tea plantation (123)
		Other garden plot (125)
	Wood land	Woodland (131)
		Sparsely forested woodland (133)
		Young afforested woodland (134)
		Cleared land after logging (135)
		Land for tree nursery (136)
	Other land	Pond for irrigation (154)
		Pond for vegetation (155)
		Pond for vegetation (K) (155K)
		Grain-sunning ground (158)
Construction land	Residential land	Residential in urban areas (201)
		Residential in town areas (202)
		Residential in rural areas (203)
		Industrial and mining land (204)
		Land for special use (206)
	Transportation land	Land for railway (261)
		Land for highway (262)
	Water conservancy land	Reservoir (271)
		Water Facility (272)
Unused land	Unused land	Wasted land (311)
		Sand land (314)
		Barren earth (315)
		Exposed rock and shingle land (316)
		River (321)
		Lake (322)
		Reed marsh (323)
		Shoal (324)

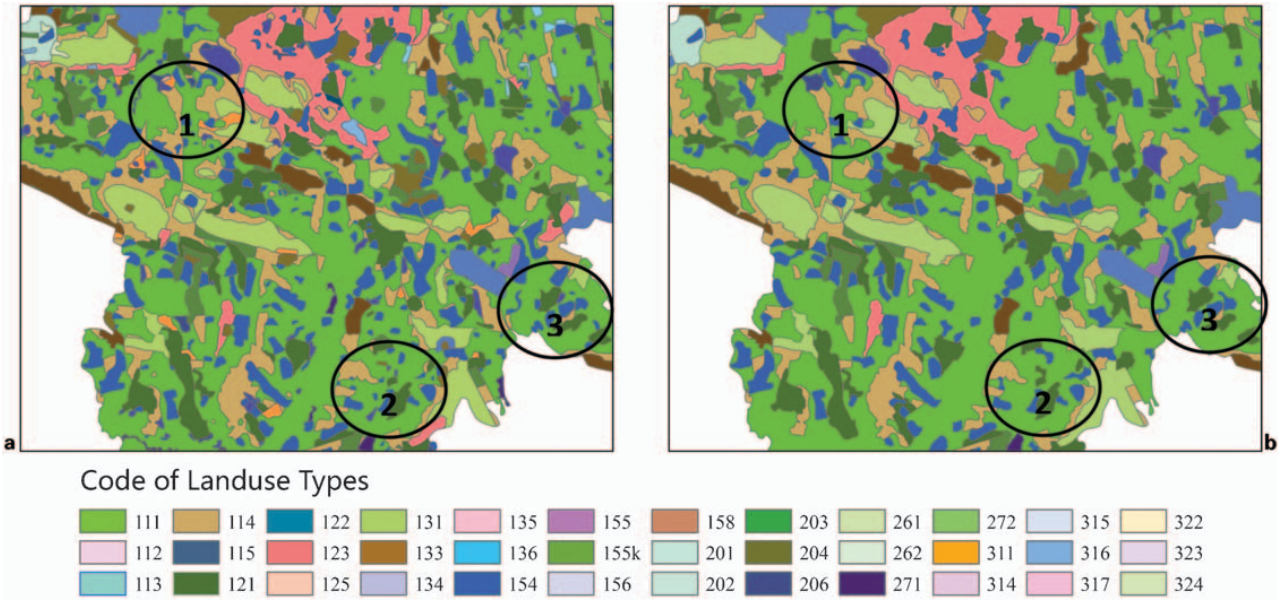

Figure 4. Land-use maps before generalization (a) and after generalization (b) of small-area patches at 1:50 000 scale. The display scale is not real 1:50 000, for clarity purposes: (a) before generalization, (b) after generalization

present the association degrees with the land-use types 111, 114 and 154, respectively. A high bar corresponds with a large association degree. Residence in rural area (203) has the highest association confidence and thus equals 2.0 in total.

Figure 3b aligns the land-use codes in a descending order according to their preliminary importance. The blue block at the bottom represents the preliminary importance, while the red block is the increase after adjustment and the total height of each pillar gives the final importance. As a consequence, the importance of each land-use type is increased, and the importance order of the land-use types is changed. Taking grain-sunning ground (158) as an example, we notice that the value of its preliminary importance is close to 0 and lower than those of tea plantation (123) and industrial and mining land (204). It has, however, a strong relationship with type 111, as the association confidence is close to 1 (Figure 3a). After adjustment, its importance is improved beyond those of types 123 and 204. The explanation is that grains harvested from the patches of type 111 generally need to be dried on grain-sunning grounds of type 158.

Next, the final importance values are ranked into three classes with the natural break classification, yielding in total four classes including the least important class defined above. These four classes are then assigned with the ordinal labels 1–4 representing the importance ranks from the most important class to the least important class (Table 3).

Setting the minimum area threshold

Referring to the Radical Law and the Technical Report on Land use Survey of Guangdong Province in China in 1998 (Gao *et al.*, 2004a), we set A_{min} and A_{max} in equation (8) equal to 1.3 and 15 mm^2 on the target map, respectively. The minimum area threshold (*MA*) for each land-use rank is then determined (Table 4).

Generalizing the small-area patches

Following Table 1, we processed the small-area patches one by one starting from the least important land-use types. Figure 4 shows the part of the test maps at the scale of 1:50 000 before generalization (Figure 4a) and after generalization (Figure 4b).

From Figure 4a we observe many small dark blue patches belonging to type 154 (pond for irrigation), as one of the

Table 4. Minimum area thresholds for the 4-ranks land-use types. The third column is the corresponding number of small-area patches of each rank

Importance rank	Minimum area threshold (mm^2)	Number of small-area patches
1	1.3	785
2	5.9	362
3	10.4	231
4	15	45

Table 3. Importance ranks of land-use types. The third column is the corresponding area ratio of land-use types of each rank

Importance value	Land-use types	Importance rank (*I*)	Area ratio (%)
1.13–1.41	111, 114, 154, 203, 322	1	63
0.92–1.14	113, 115, 121, 131, 133, 155, 158, 206, 261, 272	2	23
0.24–0.93	123, 136, 204, 311, 315	3	12
Support <0.4%	112, 122, 125, 134, 135, 155k, 155, 201, 202, 262, 271, 314, 316, 317, 323, 324	4	2

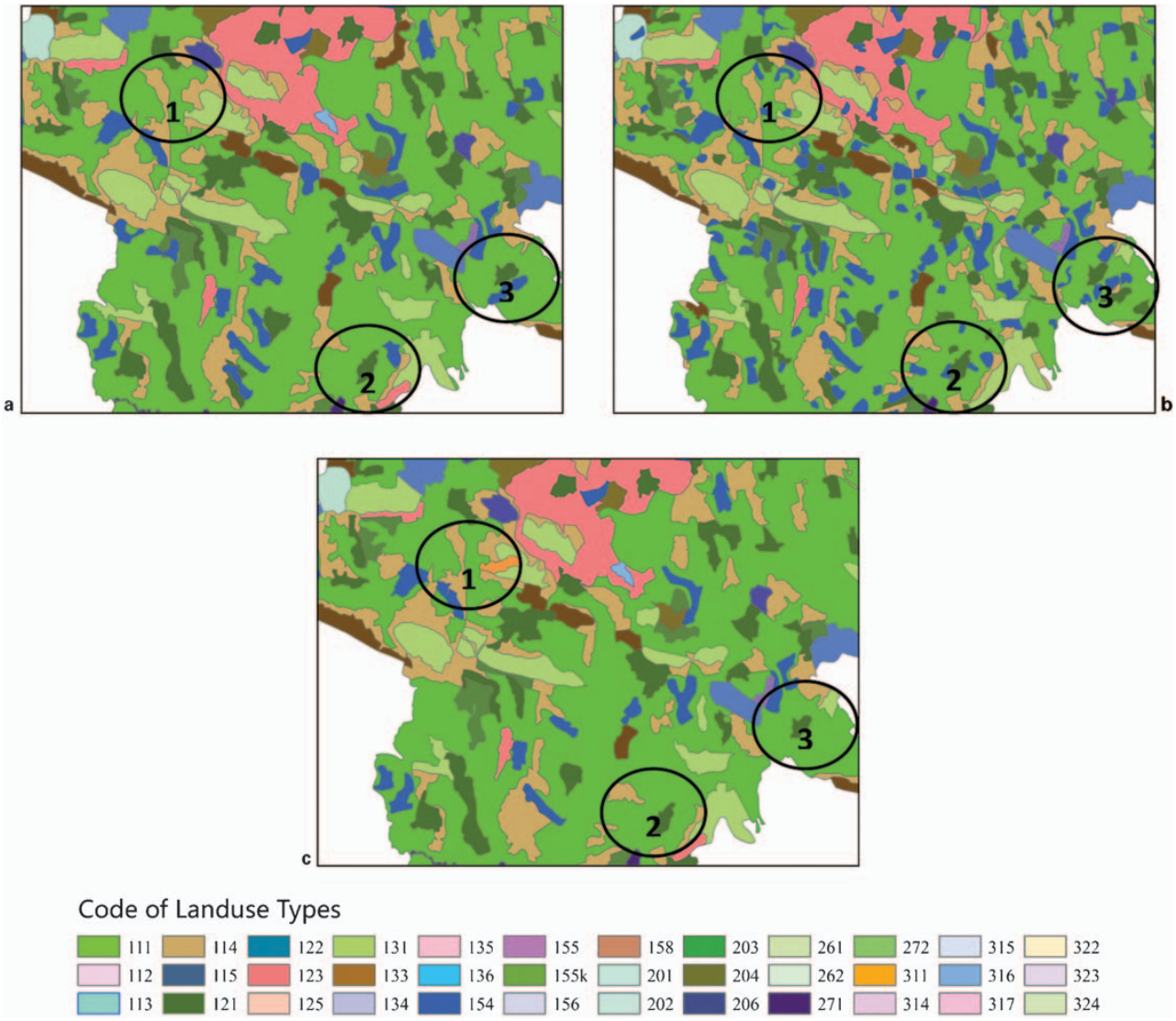

Figure 5. The generalized results of Test 1 (a) using the single minimum area threshold, Test 2 (b) using the single processing, and Test 3 (c) using the single minimum area threshold and single processing

most important land-use types. Some of these patches were merged, while others were aggregated in line with Table 1. The black circles indicate examples of the generalized patches on the two maps. Small-area patches of other land-use types were also generalized following the same procedure. On the final outcome (Figure 4b), only patches larger than the corresponding area thresholds remained. Visually, the structure of land-use patches is preserved and the whole map looks much clearer after removal of the small-area patches.

To analyze the influence of the choice for the minimum area thresholds and the quality of the data generalization, three additional tests following different procedures were implemented. For a convenient description, the test yielding the map on Figure 4b is called Test 0. The three additional tests are called Test 1, Test 2 and Test 3, respectively. Their details are as follows:

Test 1: a fixed minimum area threshold was set as 8.148 mm^2 on the map, being equal to the mean value of A_{min} and A_{max} in Test 0 for all land-use types, while the generalization processing on small-area patches was implemented in line with Table 1. The intention for Test 1 is to identify the influence of the generalization processing controlled with the importance ranks of land-use types (Figure 5a).

Test 2: the minimum area thresholds in Table 4 were used to detect small-area patches, but the single processing of *Spatial context 3* in Table 1 was used to generalize the small-area patches. Test 2, therefore, fixed the generalization processing but the minimum area threshold for each land-use type is determined according to its importance rank (Figure 5b).

Test 3: the minimum area threshold was fixed at 8.148 mm^2 for each land-use type and the single processing of *Spatial context* was used to generalize the small-area patches. Test 3, therefore, uses the fixed area threshold and processing without any consideration of the land-use type importance at all (Figure 5c).

The black circles shown in Figure 5 indicate the corresponding areas in Figure 4. The map in Figure 5c

appears to be over-generalized because too many small patches have disappeared and the overall structure of the spatial distribution of the patches is lost. More small patches have remained in Figure 5a, but some parts are also over-generalized, for example, the 154-type patches in the black circles 1 and 2. The map in Figure 5b appears to be similar to the map in Figure 4b. It indicates that the importance of land-use types has more influences on the minimum area threshold than on the generalization processing.

EVALUATION OF GENERALIZED RESULTS

Evaluation on geometric aspect

After the four test generalizations, the number of land-use types is 26, 25, 26 and 25, and the dominance index is 0.879, 0.978, 0.899 and 0.995, respectively, on the four output maps (Table 5). Compared with the 36 land-use types and the dominance index ($\boldsymbol{D}$=1.13) of the original map, about 10 land-use types are eliminated on the generalized maps and their area is released to other land-use types. The disappeared land-use types have very few numbers of patches (the number ratio is smaller than 0.9% for each type) and very limited area proportions (smaller than 0.1%). As a result, the area proportion of the remained land-use types trends to be more equal on the landscape.

Compared with the original map, the area ratio and number ratio of the land-use types in Rank 1 on the map yielded in Test 0 are increased. The increased area is partially due to the disappeared land-use types and partially due to the land-use types of rank 2, 3 and 4 according to the procedure in Table 1. The result of Test 2 is more similar to that of Test 0. Test 3 produced a totally different result. The number ratios from Rank 2 to 4 are even larger than those of the original map and the number ratio of Rank 1 is decreased, while the area ratio is increased with only 0.01. Therefore, the result of Test 3 loses too many details of the important land-use types at the target scale instead remains the many characteristics of minor land-use types.

Evaluation on semantic aspect

According to the procedure in Table 1, a small-area patch (O_s) releases its area in different ways on the three spatial contexts. *Spatial context 1* leaves the area of O_s in the same land-use type; for *Spatial context 2,* O_s contributes its area to its super-classes; for *Spatial context 3*, O_s releases its space to a totally different land-use type. In order to evaluate the semantic accuracy with equations (9)–(11), therefore, we set $\mu_{O_s}^{C_A}$ equal to 1, 0.5 and 0 for *Spatial context 1*, *2* and *3*, respectively.

According to Table 6, the map of Test 0 has the highest semantic accuracy, and the semantic accuracy of Test 2 is higher than that of Tests 1 and 3. In addition, we calculated the semantic accuracy at the map level (i.e. at the landscape) to be equal to 0.986, 0.973, 0.976 and 0.964 for the four tests, respectively. Test 0, likewise, has the highest accuracy and Test 2 prforms better than Test 1 and Test 3.

DISCUSSION

This study intends to improve the objectivity and adaptability of generalization constraints and process of land-use data generalization by introducing the importance of land-use types to assist in setting the minimum area threshold and selecting the generalization operators for small-area patches. The paper focuses on how to identify the importance of land-use types and how to utilize it during the process of land-use data generalization.

The dominance index gives a general overview of the land-use distribution at the landscape. It is helpful to know about initially the basic characteristics of land use before generalization. MADM considers various factors into identifying the importance of land-use types and the spatial associations between different land-use types are introduced to determine the final importance of each land-use type. Such importance is utilized in constructing the mathematic function to set the minimum area thresholds of land-use types and controlling the generalization processing of small-area patches. Compared with the traditional methods, the mathematic function provides more objective and adaptive values of the minimum area thresholds for different land-use types.

Comparing the four tests, we found that the result of Test 0, i.e. using the proposed methods, shows visually a more reasonable layout than the results of the other Tests. In addition, according to the quantitative evaluation on both geometric and semantic aspects, the map obtained from Test 0 retains the general characteristics of the most important land-use types as the geometric complexity of the whole map is decreased.

Comparing Tests 1 and 2, we found that the influence of the land-use type importance on the minimum area

Table 6. Semantic accuracy of each rank on the maps yielded in the four tests

Importance rank	Test 0	Test 1	Test 2	Test 3
Rank 1	0.987	0.973	0.978	0.967
Rank 2	0.983	0.971	0.976	0.963
Rank 3	0.987	0.971	0.972	0.956
Rank 4	0.984	0.933	0.975	0.929

Table 5. The area ratio and number ratio of land-use types of each rand on the original map and the four maps yielded in the four tests

	Original map		Map of Test 0		Map of Test 1		Map of Test 2		Map of Test 3	
Importance rank	Area ratio	Number ratio	Area ratio	Number ratio	Area ratio	Number ratio	Area ratio	Number ratio	Area ratio	Number ratio
Rank 1	0.626	0.685	0.657	0.848	0.647	0.658	0.655	0.835	0.640	0.614
Rank 2	0.234	0.186	0.225	0.101	0.223	0.199	0.224	0.109	0.227	0.222
Rank 3	0.118	0.107	0.102	0.043	0.110	0.114	0.103	0.048	0.112	0.134
Rank 4	0.022	0.022	0.017	0.008	0.021	0.029	0.017	0.008	0.021	0.031

thresholds is more intensive than on the generalization processing. Therefore, for an efficient implementation, the minimum area thresholds are variable according to the land-use type importance, while the generalization processing can adopt a single operator without considering the difference of the importance.

The selection of attributes and their weights in MADM, however, are variable with characteristics and specific applications related to land use. Moreover, different geometric conflicts may consider different spatial contexts and accordingly appropriate procedure is selected to solve the conflicts.

This study focused on polygon objects in land-use dataset excluding line and point objects. In some practical land-use datasets, some features, such as highways and rivers with significant economic values, are represented as line objects. The future work will extend the presented methods to consider all kinds of objects and their inter-relationship into land-use data generalization.

CONCLUSIONS

This study proposed a three-step method to identify effectively the importance of land-use types. It consists of compiling a diversity index, applying a multiple attribute decision model and carrying out a spatial association analysis. The importance of land-use types was utilized in a mathematic function to determine the minimum area thresholds for these land-use types. The importance and the minimum area thresholds were critical generalization constraints to control the generalization processing of small-area patches.

The study showed that generalization constraints constructed with the minimum area thresholds assisted in generating reasonable and objective outputs during land-use data generalization. There was a clear effect of the choice for the importance of land-use types. The generalized outputs were different for the different contexts of land-use database and depended upon the application of land-use data generalization. An incorrect specification of threshold would lead to a decrease of semantic accuracy.

BIOGRAPHICAL NOTES

Dr Wenxiu Gao works at State Key Lab of Information Engineering in Surveying, Mapping & Remote Sensing (LIESMARS) in Wuhan University. Her research interests include geo-visualization, map generalization and geographic information standard. She has an MS degree in cartography and PhD in Photogrammtry and Remote Sensing, both from Wuhan University.

ACKNOWLEDGEMENTS

This study was supported by National Natural Science Foundation of China (grant no. 41023001 and 41021061).

REFERENCES

Bader, M. and Weibel, R. (1997). 'Detecting and Resolving Size and Proximity Conflicts in the Generalization of Polygon Maps', in **18th International Cartographic Conference**, pp. 1525–1532, Stockholm, Sweden.

Beard, M. K. (1991). 'Constraints on rule formation', in **Map Generalization: Making Rules for Knowledge Representation**, ed. by Buttenfield, B. P. and Mcmaster, R. B., pp. 121–135, Longman, London.

Belton, V. and Stewart, T. J. (2002). '**Multiple Criteria Decision Analysis: An Integrated Approach**', Kluwer Academic Publ., Boston, MA.

Castilla, G. and Hay, G. J. (2007). 'Uncertainties in land use data', **Hydrology and Earth System Sciences**, 11, pp. 1857–1868.

Cheng, T. and Li, Z. (2006). 'Effect of generalization on area features: a comparative study of two strategies', **The Cartographic Journal**, 43, pp. 157–170.

Edwardes, A. and Mackaness, W. (2000). 'Modelling knowledge for automated generalization of categorical maps - a constraint based approach', in **GIS and GeoComputation (Innovations in GIS 7)**, ed. by Atkinson, P. and Martin, D., pp. 161–173, Taylor & Francis, London.

Gao, W., Gong, J., Yang, L., Jiang, X. and Wu, X. (2012a). 'Detecting geometric conflicts for generalisation of polygonal maps', **The Cartographic Journal**, 49, pp. 21–29.

Gao, W., Gong, J. & Zhilin, L. (2004a). 'Thematic knowledge for the generalization of land use data', **The Cartographic Journal**, 41, pp. 245–252.

Gao, W., Song, A. and Gong, J. (2004b). 'Constraint-based Generalization of Soil Map', in **the XXth CONGRESS of ISPRS**, pp. 205–209, Istanbul, Turkey, Jul 12–23.

Gao, W., Stein, A., Yang, L., Wang, Y. & Fang, H. (2012b). 'Improving Representation of Land-use Maps Derived from Object-oriented Image Classification', **Transactions in GIS**, in press.

Han, J. and Kamber, M. (2006b). '**Data Mining: Concepts and Techniques**', Morgan Kaufmann Publishers.

Haunert, J.-H. and Wolff, A. (2006). 'Generalization of Land Cover Maps by Mixed Integer Programming', in **The 14th International Symposium on Advances in Geographic Information Systems**, pp. 75–82, Arlington, VA, Nov 1–4.

Haunert, J.-H. and Wolff, A. (2010). 'Area aggregation in map generalisation by mixed-integer programming', **International Journal of Geographical Information Science**, 24, pp. 1871–1897.

Hietala-Koivu, R., Lankoski, J. and Tarmi, S. (2004). 'Loss of biodiversity and its social cost in an agricultural landscape', **Agriculture, Ecosystems & Environment**, 103, pp. 75–83.

Kilpelainen, T. (2000). 'Knowledge acquisition for generalization rules', **Cartography and Geographic Information Science**, 27, pp. 41–50.

Liu, Y., Molenaar, M. and Kraak, M.-J. (2002). 'Semantic Similarity Evaluation Model in Categorical Database Generalization', in **Symposium on Geospatial Theory, Processing and Applications**, Ottawa, Canada, Jul 9–12.

Mackaness, W. A., Perikleous, S. & Chaudhry, O. Z. (2008). 'Representing forested regions at small scales: automatic derivation from very large scale data', **The Cartographic Journal**, 45, pp. 6–17.

Martinez, S., Ramil, P. and Chuvieco, E. (2010). 'Monitoring loss of biodiversity in cultural landscapes. New methodology based on satellite data', **Landscape and Urban Planning**, 94, pp. 127–140.

McMaster, R. B. & Shea, S. (1992). '**Generalisation in Digital Cartography**', The Association of American Geographers, Washington D.C.

Muller, J. C. and Wang, Z. (1992). 'Area-patch generalisation: a competitive approach', **The Cartographic Journal**, 29, pp. 137–144.

O'Neill, R. V., Krummel, J. R., Gardner, R. H., Sugihara, G., Jackson, B., Deangelis, D. L., Milne, B. T., Turner, M. G., Zygmunt, B., Christensen, S. W., Dale, V. H. and Graham, R. L. (1988). 'Indices of landscape pattern', **Landscape Ecology**, 1, pp. 153–162.

Oosterom, P. V. (1995). 'The GAP-tree, an approach to 'on-the-fly' map generalization of an area partitioning', in **GIS and Generalization – Methodology and Practice**, ed. by Müller, J.-C., Lagrange, J.-P. and Weibel, R., Taylor & Francis, London.

Podolskaya, E. S., Anders, K.-H., Haunert, J.-H. and Sester, M. (2007). 'Quality assessment for polygon generalization', in **The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, Enschede, The Netherlands.

Riitters, K. H., O'neill, R. V., Hunsaker, C. T., Wickham, J. D., Yankee, D. H., Timmins, S. P., Jones, K. B. and Jackson, B. L. (1995). 'A factor analysis of landscape pattern and structure metrics', **Landscape Ecology**, 10, pp. 23–39.

Saura, S. and Martinez-Mlian, J. (2001). 'Sensitivity of landscape pattern metrics to map spatial extent', **Photogrammetric Engineering & Remote Sensing**, 67, pp. 1027–1036.

Walsh, S. E., Soranno, P. A. and Rutledge, D. T. (2003). 'Lakes, wetlands, and streams as predictors of land use/cover distribution', **Environmental Management**, 21, p. 16.

Zhang, X., Ai, T., Stoter, J., Kraak, M.-J. and Molenaar, M. (2011). 'Building pattern recognition in topographic data: examples on collinear and curvilinear alignments', **GeoInformatica**, pp. 1–33.