GEOMORPHOMETRY 2020
PERUGIA, ITALY
PARTNERS SPONSORS
JUNE 22 - 26 2020

A case-based classification strategy of automatically selecting terrain covariates for modeling geographic variable–environment relationship

Cheng-Zhi Qin[§], Peng Liang
State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS
Beijing, China
[§] qincz@lreis.ac.cn

A-Xing Zhu
Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application and School of Geography
Nanjing Normal University
Nanjing, China

Department of Geography
University of Wisconsin-Madison
Madison, Wisconsin, USA

Abstract—**It is valuable for modelers (especially those non-experts) in real applications to automatically select a proper set of terrain covariates for modeling the relationship between a geographic variable (phenomenon) and its environment, which is the basis for many predictive mapping of geographic variables (such as soil properties, landslide susceptibility, and species habitat suitability). For this aim, in this study we propose a case-based classification strategy which is designed based on two considerations. The first is that the cases created from existing predictive mapping applications with terrain covariates determined by experts would contain the implicit and non-systematic knowledge on selecting covariates according to specific application contexts. The second is that a binary classifier for each of terrain covariate candidates can be trained by the cases collected in advance and then be applied to a new application for automatically determining if the corresponding covariate should be selected or not for the new application. The proposed strategy can relieve users' burden on using traditional statistical methods of selecting terrain covariates for mapping in a study area, i.e., collecting a large number of samples in the study area in advance, and preparing dataset of all terrain covariate candidates of the study area. Two methods based on the proposed strategy were implemented, i.e., the random forests (RF) method, and the logistic regression (LR) method. With the application domain of digital soil mapping (DSM), we built a case base containing 191 DSM cases which totally use 38 terrain covariates, and then conducted a leave-one-out experiment for evaluation. Experimental results show that RF with the proposed strategy performed better.**

I. INTRODUCTION

Spatial distribution of geographical variables (phenomena; such as soil properties, landslide susceptibility, and species habitat suitability) is estimated increasingly by predictive mapping through modeling the relationship between geographical variables and environmental covariates [1]. Among those environmental covariates, terrain covariates are those mostly used (even exclusively used) [2,3], due to not only the substantial relationship between geographical variables and terrain [1] but also the high availability of digital elevation model (DEM) data for deriving diverse terrain covariates [4].

Selection of a proper set of terrain covariates is crucial for building a reliable model for depicting a geographic variate–environment relationship. Ignorance of important terrain covariate(s) will obviously impact the reliability of the built model of the relationship. Besides, inclusion of unnecessary terrain covariates may introduce errors to the model result.

For modelers of predictive mapping (especially those non-experts), it is still a challenge to select a proper set of terrain covariates for real applications. The selection of proper terrain

Cheng-Zhi Qin, Peng Liang,A-Xing Zhu (2020) A case-based classification strategy of automatically selecting terrain covariates for building geographic variable-environment relationship:

in Massimiliano Alvioli, Ivan Marchesini, Laura Melelli & Peter Guth, eds., Proceedings of the Geomorphometry 2020 Conference, doi:10.30437/GEOMORPHOMETRY2020_9.

covariates highly depends on the domain knowledge related to the application context (such as the target of predictive mapping, geographic characteristics of study area, and data resolution), while nowadays so many topographic attributes are candidates of terrain covariates [2,4]. Such application context knowledge, although crucial in modeling, is often implicit, non-systematic, and hard to be presented in a clear form (such as rules) for modelers [4,5].

Currently there still has no effective method of automatically selecting a proper set of terrain covariates for predictive mapping, so to lower the burden of modelers. Some statistical methods have been designed to select terrain covariates for predictive mapping [6-9]. However, they need modelers to collect a large number of samples in the application area in advance, and to prepare dataset of all candidates of terrain covariates of the area. Then it could be tested if each individual of terrain covariate candidates is statistically related to the geographical variable of predictive mapping. Such requirements of using these statistical methods are heavy burden on modelers and often unpractical in real applications. Thus the statistical methods are with limited applicability and also hard to be automated.

In this study we propose a case-based classification strategy of automatically selecting terrain covariates for modeling geographic phenomenon–environment relationship. Two methods are designed based on the proposed strategy and evaluated based on an experiment of selecting terrain covariates for digital soil mapping (DSM).

II. Methods

The case-based classification strategy is proposed based on two considerations. The first is that the cases created from existing applications with terrain covariates determined by experts would contain the application context knowledge on selecting terrain covariates. Artificial intelligent domain provides "case" as a suitable way to formalizing the prior and non-systematic knowledge [10]. Case-based method has been primarily explored in digital terrain analysis [5] and showed promising performance in using application context knowledge to support automatic modeling. The second consideration is that a binary classifier for each individual of terrain covariate candidates can be trained by the collected cases in advance and then be applied to a new application. Then it could be automatically determined if the corresponding covariate should be selected for the new application [11]. Unlike those existing statistical methods of selecting terrain covariates, the proposed strategy need neither collecting a large number of samples in the study area of the new application, nor preparing the dataset of all candidates of terrain covariates in the study area. Both the training process and applying process of the proposed strategy can be automated, once the case base was built. Therefore, the proposed strategy should be reasonable and practical for automatically selecting a proper set of terrain covariates for predictive mapping.

A. *Case formalization*

Similar to the normal design of cases, a case which records an existing applications with terrain covariates determined by experts is designed as two parts, i.e., the problem part, and the solution part. The problem part of a case describes the application context information of the case. In this study, the target of predictive mapping, geographic characteristics of application area, and data resolution are recorded in the case problem part. Specifically, the geographic characteristics of application area are described by two factors (i.e., the area size, and the terrain complexity of the application area) and further formalized as four quantitative attributes (i.e., the size of application area, the total relief, the standard deviation of elevation, and the mean slope of the application area) [5,11].

In this study, the solution part of a case records those terrain covariates adopted by the corresponding application case.

According to above-designed case formalization, a case base could be built through collecting and formalizing existing applications with terrain covariates determined by experts. These existing applications could be collected from scientific publications and open technical reports of real applications of predictive mapping.

B. *Binary classification methods*

Above-built case base can be used to train a binary classifier for each terrain covariate appearing in the solution part of cases in the case base. The input features of the classifiers are attribute values of the case problem part. The output of each classifier is a Boolean value, that is, whether the corresponding terrain covariate should be selected for modeling the geographical variate–environment relationship under the input features (i.e., the application context of the case). Then, such classifiers trained for each terrain covariates can be used to automatically select terrain covariates for a new application case, according to the problem part of the new case.

In current study we consider two popular binary classification methods, i.e., the random forests (RF) method, and the logistic regression (LR) method. RF [12] is a typical ensemble machine learning method widely used for classification. Its advantages include noise resistance, working well on imbalanced data, and free of variable distribution [12]. It has been proposed for the case-based strategy of automatically selecting of terrain covariates [11].

LR is a generalized linear model for classification, which can produce the probability of classification and require no assumption on the data distribution. LR could be potentially available for the proposed strategy in this study.

III. Experiment

Above-presented two methods with the proposed strategy were evaluated through an experiment of selecting terrain covariates for digital soil mapping (DSM). DSM is often conducted by modeling the soil–environment relationship for a study area, in which terrain covariates are mostly used and even exclusively used [3]. While results from RF with the proposed strategy was recently published in Ref. [11], this study focuses on comparison between RF and LR with the proposed strategy.

A. Case base of DSM

In this experiment we adopted a DSM case base built in Ref. [11]. The DSM case base contains 191 cases spread around the world (Figure 1), which were collected from the scientific papers published in DSM-related journals in recent years. A total of 38 terrain covariates appear in the case base. Details of the case base are referred to Ref. [11].

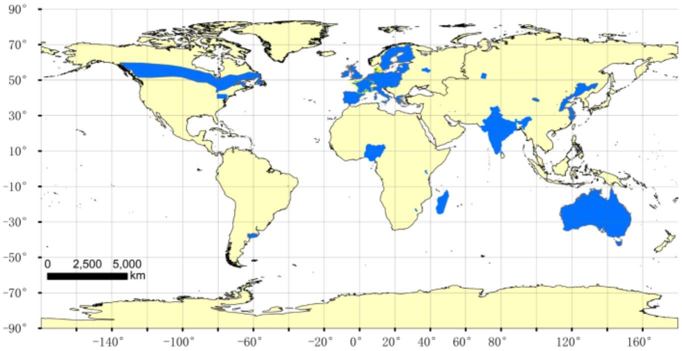

Figure 1. Map of DSM cases.

B. Experimental design

A leave-one-out experiment was designed to evaluate the performance of the two methods with the proposed strategy.

A so-called Novice method was also test as a reference method for comparison with the proposed methods. The Novice method simulates a normal way of selecting covariates by novices, that is, adopting those terrain covariates which were most frequently appeared in the case base, according to the count of covariates adopted in the original solution (i.e., the solution part) of an evaluation case [11].

Three quantitative evaluation indices widely used for classification accuracy evaluation were calculated for the results from the tested methods when applying to a new application case (i.e., evaluation case) in the leave-one-out experiment.

$$recall = \frac{TP}{TP + FN} \qquad (1)$$

$$precision = \frac{TP}{TP + FP} \qquad (2)$$

$$F1\text{-}score = \frac{2 * (precision * recall)}{precision + recall} \qquad (3)$$

where *TP*, *FN*, and *FP* mean True Positives, False Negatives, and False Positives, respectively. The *recall* index is the ratio of covariates selected correctly by the method under test to those covariates adopted in the original solution of the evaluation case. The *precision* index is the ratio of covariates correctly selected by the method under test to those covariates selected out by the method. The *F1-score* ranges from 0 (the worst performance) to 1 (the best performance).

Mean and standard deviation (Std.) of these three evaluation indices were calculated for comparing the performance of the methods under test.

C. Experimental results

Table 1 shows that RF with the proposed strategy performed best. RF correctly selected most of covariates for evaluation cases, and meanwhile selected less covariates which were not in the original solution of evaluation cases. Although LR with the proposed strategy performed better than the Novice method according to the *precision* index, LR performed worst according to the *recall* index and *F1-score* index. By comparison, LR often selected less covariates for evaluation cases, which resulted in lower values of *recall* index. Note that the imbalance among frequency of individual covariates appeared in the case base plays a challenge on the classification methods, RF showed its advantage of working well on imbalance data, while LR performed poorly on current limited case base with imbalance data.

TABLE I. Mean and standard deviation (Std.) of the evaluation indices from the methods under test.

Method	Evaluation Index	Mean	Std.
RF with the proposed strategy	*recall*	**0.644**	0.380
	precision	**0.704**	0.391
	F1-score	**0.624**	0.362
LR with the proposed strategy	*recall*	0.414	0.350
	precision	0.546	0.407
	F1-score	0.332	0.275
Novice	*recall, precision, F1-score*	0.474	0.321

IV. Conclusions

In this study we propose a case-based classification strategy of automatically selecting terrain covariates for modeling geographic phenomenon–environment relationship. Two methods (i.e., RF, and LR) based on the proposed strategy were implemented and compared. A leave-one-out experiment based on a DSM case base shows that RF with the proposed strategy performed best.

REFERENCES

[1] Zhu, A-X., G. Lu, J. Liu, C.-Z. Qin, C. Zhou, 2018. "Spatial prediction based on Third Law of Geography". Ann. GIS 24, 225–240.

[2] Hengl, T., and H.I. Reuter (Eds), 2008. "Geomorphometry: Concepts, Software, Application". Developments in Soil Science - Volume 33, Elsevier, 765 p.

[3] McBratney, A., M. Mendonça Santos, B. Minasny, 2003. "On digital soil mapping". Geoderma 117, 3–52.

[4] Wilson, J.P. (Eds), 2018. Environmental Applications of Digital Terrain Modeling. Hoboken, NJ: Wiley Blackwell.

[5] Qin, C.-Z., X.-W. Wu, J.-C. Jiang, A-X. Zhu, 2016. "Case-based knowledge formalization and reasoning method for digital terrain analysis – application to extracting drainage networks". Hydrol. Earth Syst. Sci. 20, 3379–3392.

[6] Lagacherie, P., A.-R. Sneep, C. Gomez, S. Bacha, G. Coulouma, M.H. Hamrouni, I. Mekki, 2013. "Combining Vis–NIR hyperspectral imagery and legacy measured soil profiles to map subsurface soil properties in a Mediterranean area (Cap-Bon, Tunisia)". Geoderma 209–210, 168–176.

[7] Adhikari, K., A.E. Hartemink, B. Minasny, R.B. Kheir, M.B, Greve, M.H. Greve, 2014. "Digital mapping of soil organic carbon contents and stocks in Denmark". PloS One 9, e105519.

[8] de Carvalho Junior, W., P. Lagacherie, C. da Silva Chagas, B. Calderano Filho, S.B. Bhering, 2014. "A regional-scale assessment of digital mapping of soil attributes in a tropical hillslope environment". Geoderma 232–234, 479–486.

[9] Vaysse, K., and P. Lagacherie, 2015. "Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France)". Geoderma Reg. 4, 20–30.

[10] Kaster, D.S., C.B. Medeiros, H.V. Rocha, 2005. "Supporting modeling and problem solving from precedent experiences: the role of workflows and case-based reasoning". Environ. Model. Softw. 20, 689–704.

[11] Liang, P., C.-Z. Qin, A-X. Zhu, Z.-W. Hou, N.-Q. Fan, Y.-J. Wang, 2020. "A case-based method of selecting covariates for digital soil mapping". Journal of Integrative Agriculture, doi: 10.1016/S2095-3119(19)62857-1.

[12] Breiman, L., 2001. "Random Forests". Mach. Learn. 45, 5–32.