

## Classifying natural-language spatial relation terms with random forest algorithm

Shihong Du, Xiaonan Wang, Chen-Chieh Feng & Xiuyuan Zhang

To cite this article: Shihong Du, Xiaonan Wang, Chen-Chieh Feng & Xiuyuan Zhang (2016): Classifying natural-language spatial relation terms with random forest algorithm, International Journal of Geographical Information Science, DOI: [10.1080/13658816.2016.1212356](https://doi.org/10.1080/13658816.2016.1212356)

To link to this article: <http://dx.doi.org/10.1080/13658816.2016.1212356>



Published online: 25 Jul 2016.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



# Classifying natural-language spatial relation terms with random forest algorithm

Shihong Du<sup>a</sup>, Xiaonan Wang<sup>a</sup>, Chen-Chieh Feng<sup>b</sup> and Xiuyuan Zhang<sup>a</sup>

<sup>a</sup>Institute of Remote Sensing and GIS, Peking University, Beijing, China; <sup>b</sup>Department of Geography, National University of Singapore, Singapore, Singapore

## ABSTRACT

The exponential growth of natural language text data in social media has contributed a rich data source for geographic information. However, incorporating such data source for GIS analysis faces tremendous challenges as existing GIS data tend to be geometry based while natural language text data tend to rely on natural language spatial relation (NLSR) terms. To alleviate this problem, one critical step is to translate geometric configurations into NLSR terms, but existing methods to date (e.g. mean value or decision tree algorithm) are insufficient to obtain a precise translation. This study addresses this issue by adopting the random forest (RF) algorithm to automatically learn a robust mapping model from a large number of samples and to evaluate the importance of each variable for each NLSR term. Because the semantic similarity of the collected terms reduces the classification accuracy, different grouping schemes of NLSR terms are used, with their influences on classification results being evaluated. The experiment results demonstrate that the learned model can accurately transform geometric configurations into NLSR terms, and that recognizing different groups of terms require different sets of variables. More importantly, the results of variable importance evaluation indicate that the importance of topology types determined by the 9-intersection model is weaker than metric variables in defining NLSR terms, which contrasts to the assertion of 'topology matters, metric refines' in existing studies.

## ARTICLE HISTORY

Received 26 October 2015

Accepted 7 July 2016

## KEYWORDS

Natural-language spatial relations; topological terms; metric variables; random forest; social media data; geographical information retrieval

## 1. Introduction

The exponential growth of natural language text data in social media has contributed a rich data source for geographic information. Frequently found in these data are the natural language spatial relation (NLSR) terms or predicative words /phrases of various spatial relations that represent relational knowledge and spatial locations about spatial objects (Egenhofer and Mark 1995, Du *et al.* 2015, Du and Guo 2015). Therefore, research in natural language spatial description, first started in user-friendly query and interaction, has now expanded into geographical information retrieval from the web (Purves *et al.* 2007, Smart *et al.* 2007, Abdelmoty *et al.* 2009), pattern recognition from image (Vanegas *et al.* 2011, Takemura *et al.* 2012), geographic information collection (Crampton

*et al.* 2013), socioeconomic patterns analysis (Li *et al.* 2013), socio-spatial inequality exploration (Shelton *et al.* 2015), and disaster management (Adam *et al.* 2012). For these research efforts, the key is to analyze geometric data and text data in an integrative manner. However, the fundamental difference between the geometric representations in mainstream GIS data, which is based on absolute coordinate systems to describe various spatial dimensions of features, and the natural language spatial description in social network data, which is based on relative coordinate systems and uses qualitative relations and natural languages to describe the locations of geographic features, makes it challenging to integrate the two data sources for further analysis. To use such description in GIS, it is imperative to establish the correspondences between the two representations.

For NLSR terms, such translation can be understood as a transition between three levels: (1) the geometric measures at the bottom level, (2) the formalized relations at the middle level, and (3) the NLSR terms at the top level. The geometric measures are quantitative and derived from geometric coordinates, for example, distances and shapes. The formalized relations are qualitative concepts recognized in formal models. For topological relations, the 9-intersection model (Egenhofer and Herring 1991) and region connection calculation (Cohn *et al.* 1997) are two well-known formal models. For extracting directional and topological relations between regions, a formal model combining F-histogram and Allen's relationships was proposed (Matsakis and Nikitenko 2005). For modeling relative relations between regions, projective relations were used (Clementini and Billen 2006). Despite being closer to natural languages than geometric measures, many concepts in these models are not common natural language terms. A means to translate geometric measures to NLSR terms are therefore necessary for computers to handle the NLSR terms.

Existing research on this issue, however, offers partial solutions. Topological relations have been extensively studied in the context of NLSR terms. Egenhofer and Shariff (1998) used metric variables, including splitting, alongness, and closeness to refine the formalized relations in 9-intersection model, and analyzed the value ranges of these variables. Their result was later used to calibrate 59 English NLSR terms (Shariff *et al.* 1998). However, the two studies fall short of supporting real translation from geometric configurations to NLSR terms as they only reported the maximum, minimum, and median values, or the ranges of each variable. In addition, due to the lack of mathematical theory, the calibration of the NLSR terms is too simple. The line-line topological relations distinguished by the 9-intersection model were also refined using splitting ratios and closeness measures (Nedas *et al.* 2007). However, these studies were limited to line-line topological relations than to the correspondence between the geometric representation and NLSR terms. To provide more detailed characterization of line-line spatial relations, Xu (2007) first defined a set of metric variables and then classified the line-line NLSR terms using decision tree. The results showed correlations between the NLSR terms and both formalized topological relations and metric variables. Skubic *et al.* (2003) built a model for robots to generate natural language spatial descriptions using F-histogram. However, the model is robot egocentric and designed specifically to describe directional relationship between disjoint regions. It thus cannot be used to transform various geometric configurations in a geographic scene to topologically linguistic ones. Bartie *et al.* (2013) provide an egocentric projective model to produce

linguistic description of the configuration of visible cityscape objects. It does not aim at the extraction of spatial relations in natural language texts and has limited relevancy to this research theme.

In establishing the correspondences between geometric representation and NLSR terms, besides the methods to generate the mapping rules, the identification of important variables is crucial as different variables measure different aspects of NLSR terms. Namely, different NLSR terms may depend on different variables, and thus evaluating variable importance is crucial to improve the characterization and the classification accuracy of NLSR terms. For this purpose, Shariff *et al.* (1998) used the standard scores of metric variables. However, the score is not ideal for finding variable distinction as it does not consider the correlations between variables and may lead to misleading importance measures.

Existing studies are thus limited in three aspects. First, they distinguish limited numbers of NLSR terms. In social media and natural languages, however, over dozens of terms are constantly involved. The added complexity requires a new, robust method to transform geometric configurations to NLSR terms commonly used in social media and natural languages. Second, an effective method to establish the mapping from geometric configurations to NLSR terms is needed. Existing mapping methods are built on simple methods, such as cluster analysis and decision tree, with reliability that is less desirable. Moreover, the decision tree algorithm tends to over-fit samples, leading to low classification accuracies. Better methods should be explored to establish such mapping. Third, the importance of each variable to defining each NLSR term has been not fully evaluated. The three limitations highly demand that new machine learning algorithms should be adopted to translate geometric configurations into NLSR terms.

Based on the previous research on NLSR, this study aims to: (1) define a set of explanatory variables (including one nominal variable and 18 metric variables) to represent line-region topological relations and further to explain NLSR terms; (2) learn the mapping between the defined explanatory variables and the 69 English NLSR terms through the random forest (RF) algorithm (Breiman 2001); and (3) measure the importance of each variable to recognizing NLSR terms by RF (in R3.1.3) algorithm. To address these three aims, experiments are conducted to train classification model, evaluate classification accuracy, and measure variable importance. The results demonstrate that the presented approach can achieve a large accuracy and metric variables account for NLSR terms more than the topological type.

## 2. Explanatory variables

For interpreting NLSR terms, this section defines for line-region configurations a set of explanatory variables comprising one nominal variable and 18 metric variables. The nominal variable uses one of the 19 qualitative concepts determined by the 9-intersection model to describe the meaning of a line-region configuration, while the metric variables refine and distinguish different configurations with the same qualitative relation. Considering the NLSR terms, such as *cross*, *run along*, and *within*, this section uses metric variables to depict intersection, along, inclusive or disjoint configurations between lines and regions. All the metric variables are the ratios of the lengths or areas of regions and the lengths of lines so that these variables are invariant under scale

transformation. They will be used to bridge the gap between NLSR terms or text-based descriptions and the geometric configurations.

## 2.1. Topological relations

Topological relations are qualitative and formal descriptions of relational knowledge about geometric representations. The 9-intersection model is employed in this study to translate geometric configurations into relational descriptions. For a region  $A$ , the 9-intersection model expresses it as three subsets interior ( $A^\circ$ ), boundary ( $\partial A$ ), and exterior ( $A^-$ ). Similarly, a line  $B$  is represented by three subsets  $B^\circ$ ,  $\partial B$ , and  $B^-$ . In total, the nine intersections between the three subsets of a simple line and a simple region can determine 19 exhaustive line-region topological relations (Appendix Figure A1). However, this model is not designed to effectively facilitate the identification of various geometric configurations which are distinguished by different NLSR terms. Accordingly, some metric variables (e.g. splitting, alongness, and closeness metrics) are introduced to refine the line-region topological relations. The topological type is included in the set of explanatory variables as a nominal variable that ranges from 1 to 19 to represent R1–R19 in Appendix Figure A1, respectively.

## 2.2. Metric refinement

The 18 metric variables for refining the line-region relations are:

- Four splitting metrics, including  $IAS$ ,  $OAS$ ,  $ITS$ , and  $PS$ , to refine the configurations where a line and a region have nonempty intersection.
- Six alongness metrics, including  $La$ ,  $PA$ ,  $IPA$ ,  $ILA$ ,  $OPA$ , and  $OLA$ , to measure the tendency of a line along the boundary of a region.
- Eight closeness metrics, including  $OAC$ ,  $OLC$ ,  $IAC$ ,  $ILC$ ,  $OAN$ ,  $OLN$ ,  $IAN$ , and  $ILN$ , to refine the configurations where a line is disjoint from or entirely inside a region, and to measure the relative distance from a line's boundary or interior to a region's boundary or interior.

Eight of these variables, namely  $IAS$ ,  $OAS$ ,  $ITS$ ,  $PS$ ,  $La$ ,  $PA$ ,  $OAN$ , and  $IAN$ , are directly adapted while the others are revised from Shariff *et al.* (1998). Tables 1 and 2 provide the definitions and ranges of all variables. Compared with F-histogram or projective relations, these metric variables are more suitable for line-region topological refinement in a non-egocentric view.

## 3. Methodology

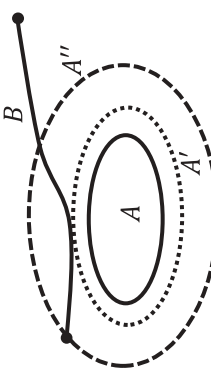
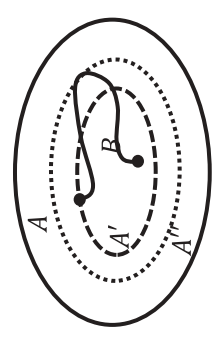
Classifying NLSR terms involves first training a mapping model linking the 19 explanatory variables with the 69 NLSR terms, and second translating geometric configurations into NLSR terms by incorporating the explanatory variables extracted from geometric configurations in the classification model. To build such a robust and accurate classification model and quantify the importance of each explanatory variables, a large number of samples and a better training algorithm are required. In

Table 1. Definitions of quantitative variables.

Category	Name	Illustration	Definition	Annotation
Splitting metrics	Inner area splitting (IAS)		$\min(\text{area}(D), \text{area}(A - D)) / \text{area}(A)$	• $\text{area}(D)$ : the left part of aggregated area of bounded areas in set $A^{\circ} \cup B^{-}$ .
	Outer area splitting (OAS)		$\text{area}(E) / \text{area}(A)$	• $\text{area}(E)$ : the aggregated area of bounded areas in set $A^{-} \cap B^{-}$ .
	Inner traversal splitting (ITS)		$\text{length}_i / \text{length}(B)$	• $\text{length}_i$ : the aggregated length of components in set $A^{\circ} \cap B^{\circ}$ .
	Perimeter splitting (PS)		$\text{length}_{\max} / \text{length}_{\text{boundary}}$	• $\text{length}_{\max}$ : the length of the longest component in set $\partial A \cap B^{-}$ .
Alongness metrics	Line alongness (La)		$\text{length}(cd) / \text{length}(B)$	• $\text{length}(cd)$ : the aggregated length of components in set $\partial A \cap B^{\circ}$ .
	Perimeter alongness (PA)		$\text{length}(cd) / \text{length}_{\text{boundary}}$	• $\text{length}_{\text{boundary}}$ : the length of $A$ 's boundary.
	Inner perimeter alongness (IPA)		$\text{length}(ef) / \text{length}(\text{boundary}(A'))$	• $A'$ : the minimum buffer polygon of $A$ that touches $B$ .
	Inner line alongness (ILA)		$\text{length}(ef) / \text{length}(B)$	• $B'$ : the line produced by buffering $B$ so that $B$ just sweep through $\partial A'$ .
	Outer perimeter alongness (OPA)		$\text{length}(ef) / \text{length}(\text{boundary}(A'))$	• $e$ and $f$ : the farthest two points of the intersection of $B'$ and $\partial A'$ .
	Outer line alongness (OLA)		$\text{length}(ef) / \text{length}(B)$	• $\text{length}(ef)$ : the length of the component $ef$ in $\partial A'$ .

(Continued)

Table 1. (Continued).

Category	Name	Illustration	Definition	Annotation
Closeness metrics	Outer area closeness (OAC)		$\Delta EA_i / \text{area}(A)$	• $A''$ : the minimum buffer polygon of $A$ that $\partial A$ intersects $\partial B$ .
	Outer line closeness (OLC)		$\Delta ED_i / \text{length}(B)$	• $\Delta EA_i$ : the area between $\partial A$ and $\partial A'$ ;
	Outer area nearness (OAN)		$\Delta EA_{A_i} / \text{area}(A)$	• $\Delta ED_i$ : the distance between $\partial A$ and $\partial A'$ ;
	Outer line nearness (OLN)		$\Delta ED_{A_i} / \text{length}(B)$	• $\Delta EA_i$ : the area between $\partial A$ and $\partial A''$ ;
	Inner area closeness (IAC)		$\Delta EA_i / \text{area}(A)$	• $\Delta ED_i$ : the distance between $\partial A$ and $\partial A''$ .
	Inner line closeness (ILC)		$\Delta ED_i / \text{length}(B)$	
	Inner area nearness (IAN)		$\Delta EA_{A_i} / \text{area}(A)$	
	Inner line nearness (ILN)		$\Delta ED_{A_i} / \text{length}(B)$	

**Table 2.** Ranges of metric variables.

Category	Variables	Range
1	TA	1, 2, 3, ..., 19
2	IAS	[0.0, 0.5]
3	ITS, La, PA, IPA, ILA, OPA, OLA	[0.0, 1.0]
4	PS, IAC, IAN	(0.0, 1.0)
5	OAS, OAC, OLC, ILC, OAN, OLN, ILN	(0.0, +∞)

addition, some measures should be adopted to evaluate the contributions of each variable to recognizing each term.

### 3.1. Sample collection

Each sample collected for the experiments includes a line-region sketch and the corresponding NLSR terms. For each sketch, at least one NLSR term is chosen to describe the line-region configuration. All the NLSR terms are English words or phrases describing the meanings of the topological relations between a line and a region, such as *cross* or *go through*. They apply to geometric configurations where a line (i.e. the subject) forms a topological relation with a region (i.e. the object). No direction terms (e.g. *north* and *south*) or metrics (e.g. *50 meters*) are involved in this study as their combination with topological terms are more complicated and beyond the scope of this study. The NLSR terms are collected from the published literature (Shariff *et al.* 1998, Leopold *et al.* 2015) and natural-language texts (e.g. the text on Wikipedia articles). Online dictionaries including Oxford English Dictionary (OED 2016) and Longman Dictionary (Longman 2016) were consulted to confirm the semantic annotations for the collected NLSR terms. They ensure the correct understanding of the NLSR terms and hence the development of a correct classification model.

The second set of samples is the records of explanatory variables extracted from the line-region sketches produced by 41 subjects, all of which are college students proficient in English but with different professional backgrounds. To eliminate semantic discrepancy in the understanding of the collected NLSR terms, all the subjects are required to align their understandings with the semantic annotation of those NLSR terms acquired in the previous step. When producing these sketches, each subject was asked to use a specialized program developed with ArcEngine 10.0 to firstly choose the targeted NLSR term(s) from the 69 terms and secondly draw a line representing a road and a region representing a park so that each line-region sketch matches the description of *a road + NLSR term(s) + a park*. Since some NLSR terms have similar semantics, such as *intersect* and *transect*, a sketch may correspond to more than one NLSR term. Therefore, the subjects are allowed to choose more than one term for each sketch, but they are also instructed to distinguish NLSR terms with different line-region sketches as far as possible. In addition, subjects do not need to consider three-dimensional configurations. For example, the configuration of ‘a bridge spans a river’ should be drawn as a two-dimensional map. The procedure of choosing the targeted NLSR term(s) and drawing a line-region sketch to match with the prescribed description would continue until every NLSR term has enough number of sketches (i.e. nearly thirty sketches at least) for training a robust model. The



resulting sketches were manually checked by the authors and native English speakers. More sketches were added specifically by the authors for covering all possible geometric configurations for each NLSR term, and then the subjects were asked to choose the appropriate terms for the added sketches. Once all the sketches are collected, the values of all 19 explanatory variables were extracted. If an explanatory variable has no valid definition for some line-region sketches, an out-of-range value is assigned. To ensure the usability of these explanatory variable records, the line-region sketches and the corresponding explanatory variable records were examined to remove erroneous samples, such as a mismatch of the line-region sketch and NLSR terms.

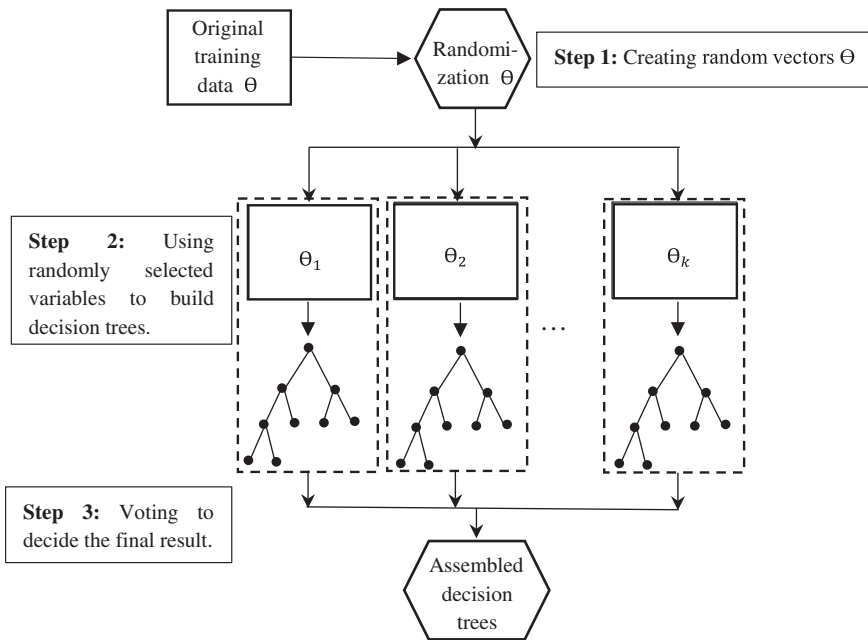
### **3.2. Classifying NLSR terms with random forest**

Given that the explanatory variables of line-region configurations are high-dimensional and are strongly correlated, and their values are highly diverse (i.e. they fall in distinct ranges with some samples with no valid values for certain variables), a robust classifier to build the correct mapping between the variables and the terms is critical. RF is a classifier composed of multiple decision trees (Breiman 2001). It excels in improving prediction accuracy, overcoming over-fitting issue, handling larger number of samples, and resolving the autocorrelation among variables due to its training and classification mechanisms (Breiman 2001). At the training stage, each decision tree is trained independently by using a randomly selected subset of samples and a subset of variables. At the classification stage, each decision tree predicts a label for each sample, and the prediction of a RF is the result of the votes from all decision trees.

For sample selection, RF employs a bagging algorithm to train each decision tree independently with a subset of samples (Figure 1). Bagging algorithm in RF helps increase the classification accuracy and evaluate the generalization error of the forest, i.e., the performance of the learned model adapting to new data. As for the variable selection, RF uses randomly chosen variables instead of all variables to reduce the correlation among decision trees while maintaining their classification capacities. The larger the differences among classes made by the model, the more accurate the decision tree. Moreover, RF is robust to outliers and noise and can provide useful internal evaluation of error, strength, correlation, and variable importance. Accordingly, RF algorithm is adopted to learn the mapping from the explanatory variables to the NLSR terms, and to translate the geometric configurations into NLSR terms.

### **3.3. Evaluating variable importance with a large number of samples**

In addition to classifying NLSR terms, RF can measure the importance of variables on recognizing these NLSR terms using a variable permutation method (Breiman 2002). This method randomly permutes all the values of a variable and then treats the difference in the prediction accuracy before and after the permutation as the measurement of the variable's importance. The larger the difference, the more important the variable.



**Figure 1.** Illustration of the procedure of RF (Pang-Ning *et al.* 2005). In the learning process, some subsets ( $\theta_1, \theta_2, \dots, \theta_k$ ) are selected randomly with replacement from the original samples ( $\theta$ ) (Step 1). Second, for each subset, a decision tree is created using randomly selected variables (Step 2). Generally, GINI coefficient or information gain can be used to form node splitting in this step. When the growing procedure stops, a RF is built, which in this study means the learning of mapping relationships is completed. In the classification process, RF predicts the NLSR term for each geometric sketch by combining the predictions of all decision trees (Step 3).

### 3.4. Experimental design

Two experiments were conducted to establish the mapping from the explanatory variables to NLSR terms, evaluate the accuracy of the model and provide variable importance measure.

#### 3.4.1. Experiment one: building classification model

For the collected 2493 samples, each corresponds to one or several NLSR terms and the values of the 19 explanatory variables. The mapping between the NLSR terms and the 19 explanatory variables thus amounts to build a classification model from the 2493 samples. To achieve this goal, a RF classifier is trained by using randomly chosen subsets of the 2493 samples, with each NLSR term as the label of each sample and the explanatory variables as the corresponding variables. The accuracy of the model is then evaluated by using the out-of-bag (OOB) samples, that are not chosen for training RF, can be used to determine the appropriate number of decision trees in a RF classifier.

#### 3.4.2. Experiment two: grouping NLSR terms and quality evaluation

Experiment two aims to determine the grouping schemes of NLSR terms and evaluate how the grouping schemes affect classification performance. Grouping NLSR terms

based on their level of semantic similarity is a necessary step to improve classification accuracy due to the potential of similar or identical values in certain explanatory variables of semantically similar NLSR terms. However, the schemes used for grouping NLSR terms can have considerable influences on the accuracy and performance of a RF classifier.

The grouping schemes were identified according to the following three criteria: (1) cognitive judgment evaluated by human, (2) similarity of terms provided by a RF classifier, and (3) separability measured by class distance. For cognitive judgment, subjects participating in the experiment first distinguish NLSR terms according to the line-region sketches and their understanding of the NLSR terms, and second merge similar NLSR terms based on personal judgements. The measure leverages human's cognition on natural language and the statistics of sample data for grouping NLSR terms, but it may introduce subjective error. For the similarity of NLSR terms provided by RF classifiers, grouping NLSR terms can be repeated according to confusion matrix produced during the training stage of each RF classifier. Confusion matrix can be used for the stated purpose because the value at row  $m$  and column  $n$  in the matrix represents the number of OOB samples belonging to term  $m$  while being predicted as term  $n$ , and thus an indicator of NLSR term similarity. The larger the value, the more similar the two terms. For the separability of terms, the average distance (Li *et al.* 2003) among samples belonging to different terms is used. The explanatory variables of a sample belonging to a NLSR term can be represented as a vector, thus a set of vectors is required for the samples belonging to the same term. The average distance between two terms refers to the Euclidean distance between their average vectors minus their average radii. The average vector of a term comprises the average values of explanatory variables of the vector set, and the average radius of a term refers to the average Euclidean distance between its average vector and each vector in the vector set. The distance among the average vectors measures the dispersion among different terms, while the average diameter measures the dispersion in one term.

The three criteria above are combined to produce a final grouping scheme. The randomForest algorithm is then employed to learn and classify the NLSR terms for 1000 iterations before the classification accuracy and variable importance measure are evaluated.

## 4. Results

### 4.1. NLSR terms and the associated explanatory variables

Experiment 1 and 2 collected a large number of data samples (2493) with each data sample including a line-region NLSR term and the corresponding values of the explanatory variables computed from line-region sketches. Below the characteristics of these data are summarized.

#### 4.1.1. NLSR terms of line-region configurations

Semantic annotations for the 69 NLSR terms are reported in Table A1 in the Appendix. The characteristics of these terms include the followings:

- Different terms may be related to the same line-region topology type, such as *in* and *within*, and *starts near* and *starts just outside*, and so on.
- Some terms can only be distinguished by the direction of a line instead of topology types, such as *enters* and *exits*, and *goes into* and *goes out of*.
- Some terms limit the location of part of instead of the whole of a line. For example, *starts just inside* can describe both the configurations where a line crosses the boundary of a region from inside to outside and a line is completely inside a region. The former is similar or identical to the configurations represented by *starts just outside* in the values of explanatory variables, while the latter is very different from the ones represented by *starts just outside* in the values of explanatory variables.
- Distinguishing the semantics of some terms is context-dependent. For example, distinguishing *ends at* and *ends in* may rely on the size of a region, with the former used in a small region while the latter used in a large region.

#### 4.1.2. Explanatory variables of line-region configurations

According to the definitions in Table 1, the values of the explanatory variables can be computed for all 2493 line-region sketches. Table 3 provides the number of sketches corresponding to each NLSR term. For all NLSR terms, the average, the maximum, and the minimum number of sketches are 37, 54 and 29, respectively. Therefore, the number of sketches for each term is large enough to establish reliable correspondence between explanatory variables and NLSR terms.

Table 4 reports the number of topology types associated with each NLSR term. A quick analysis shows that each NLSR term can be related to multiple topology types. Five of the NLSR terms—*ends just inside*, *runs along*, *runs along boundary*, *starts and ends*

**Table 3.** Number of samples corresponded by each NLSR term.

NLSR terms	Number	NLSR terms	Number	NLSR terms	Number
along edge	54	contained in edge	29	ends just outside	37
avoids	47	contained within	34	ends near	37
be adjacent to	38	crosses	34	ends on	32
bisects	44	cuts	35	ends outside	37
break into	40	cuts across	35	enters	34
by passes	46	cuts through	32	entirely outside	40
comes from	35	divides	36	exits	32
comes into	39	enclosed by	29	goes across	29
comes out of	35	encloses	37	goes away from	33
comes through	32	ends at	32	goes by	35
connected to	35	ends in	36	goes into	33
connects	36	ends just inside	36	goes out of	37
goes through	32	pass through	38	starts in	37
goes to	35	reach into	34	starts just inside	36
goes up to	35	run from	38	starts just outside	38
in	40	runs across	35	starts near	39
inside	39	runs along	38	starts outside	38
intersect	34	runs along boundary	39	stretch from	34
leads to	33	runs into	36	stretch over	28
leaves	33	run through	36	transects	32
near	39	separate	37	traverses	34
outside	44	spans	36	within	39
passes	32	splits	36	starts and ends in	37

**Table 4.** Number of topology types corresponded by each NLSR term.

NLSR terms	Number	NLSR terms	Number	NLSR terms	Number
along edge	9	contained in edge	8	ends just outside	7
avoids	1	contained within	4	ends near	3
be adjacent to	3	crosses	5	ends on	7
bisects	3	cuts	6	ends outside	5
break into	7	cuts across	3	enters	5
bypasses	1	cuts through	3	entirely outside	1
comes from	8	divides	9	exits	5
comes into	5	enclosed by	3	goes across	2
comes out of	6	encloses	3	goes away from	7
comes through	3	ends at	6	goes by	4
connected to	8	ends in	7	goes into	5
connects	9	ends just inside	10	goes out of	5
goes through	3	pass through	4	starts in	8
goes to	9	reach into	7	starts just inside	8
goes up to	8	run from	7	starts just outside	5
in	2	runs across	3	starts near	5
inside	2	runs along	10	starts outside	6
intersect	4	runs along boundary	10	stretch from	8
leads to	7	runs into	10	stretch over	8
leaves	7	run through	3	transects	3
near	3	separate	6	traverses	2
outside	2	spans	5	within	2
passes	6	splits	4	starts and ends in	10

*in*, and *runs into*—have the largest number (i.e. 10) of topology types. For the 24 terms: *along edge*, *break into*, *comes from*, *comes out of*, *connected to*, *connects*, *goes to*, *goes up to*, *leads to*, *leaves*, *passes*, *contained in edge*, *cuts*, *divides*, *ends at*, *ends in*, *reach into*, *run from*, *separate*, *goes away from*, *starts in*, *starts just inside*, *stretch from*, and *stretch over*, each of them is related to more than five topology types and imposes weaker restrictions on line-region geometric configurations compared to topology types. For example, the term *runs along boundary* is concerned with three geometric configurations: a line is inside a region’s interior and along its boundary, a line is inside a region’s exterior and along its boundary, and a line intersects and is along a region’s boundary. The collected sketches should cover these configurations because subjects use *runs along boundary* to describe all these configurations. For the three configurations, multiple topology types may occur due to different configurations of the six subsets of a line and a region. The statistical results in Table 4 indicate that topology types determined by the 9-intersection model play limited roles in distinguishing NLSR terms.

**4.2. The trained random forest: translating geometric configurations to NLSR terms**

Using the chosen terms and the values extracted from line-region sketches, a RF classifier was trained and used to translate the line-region configurations to NLSR terms. Figure 2 illustrates the translation process of a trained RF where the maximum depth of each tree is set to 4 and the symbols G, I, N, R, S at the leaf nodes refer to the terms *goes to*, *in*, *near*, *runs along boundary*, and *starts and ends in*, respectively. Each non-leaf node contains an explanatory variable, and each edge represents a condition for making a decision. For a line-region sketch, translation can be conducted using the RF classifier in the following steps. We use the geometric sketch in Figure 3 as an

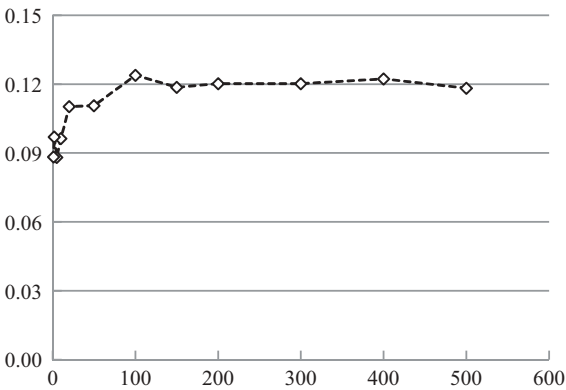


Figure 2. Translating process of the RF.

example. First, the values of explanatory variables are calculated according to definitions in Section 2. For brevity, the upper table in Figure 2 shows only the values of the explanatory variables useful for the decision process. Second, the values of metric variables are examined against each decision tree to predict one NLSR term for each

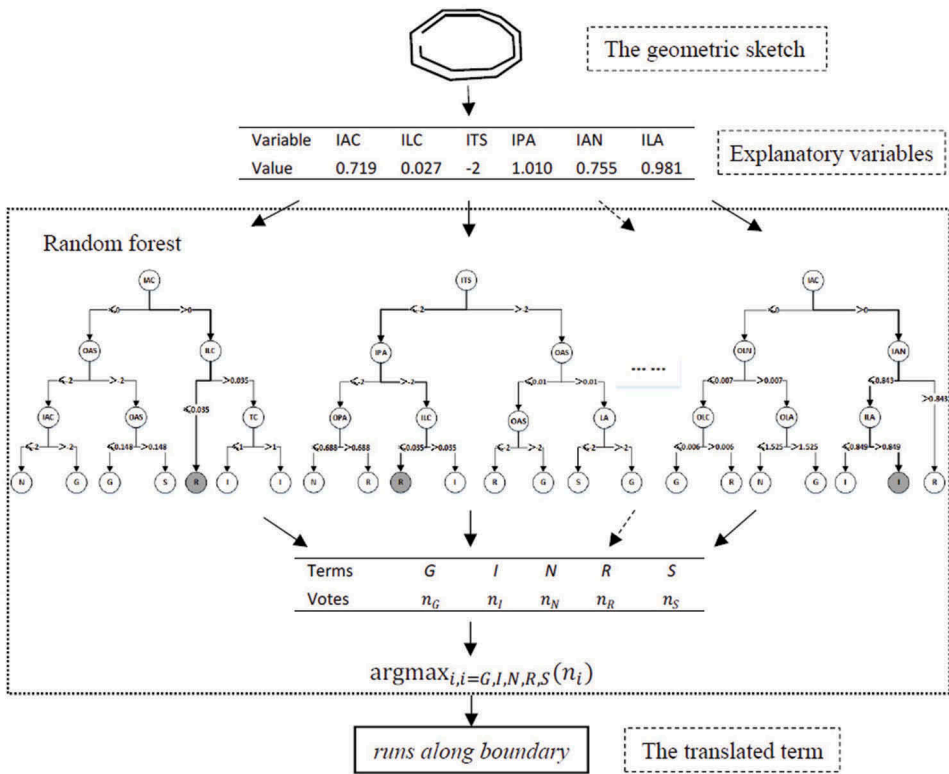


Figure 3. Classification accuracies of RF classifiers with different numbers of decision trees in experiment one.

tree. For example, for the first tree, since the satisfied rule is  $IAC > 0$  and  $ILC \leq 0.035$ , the geometric sketch is translate to the term *runs along boundary*. For the second tree, since the satisfied rule is  $ITS \leq -2$ ,  $IPA > -2$ , and  $ILC \leq 0.035$ , the sketch is also translate to the term *runs along boundary*. However, for the last tree, since the satisfied rule is  $IAC > 0$ ,  $IAN \leq 0.843$ , and  $ILA > 0.849$ , the sketch is translate to the term *in*. The discrepancies in the terms obtained by different trees result from the diversity of trees built on randomly chosen samples and variables. Third, the obtained NSLR terms of all the trees are combined by a majority voting as shown in the lower table in Figure 2. Finally, the term with the highest votes from all the trees is chosen as the final result of the RF (the formula in Figure 2). In this case, the sketch is translate to the term *runs along boundary* by the RF classifier. As shown in Figure 2, the accuracy and robustness of the RF is strengthened by synthesizing the decisions of all decision trees in the majority voting process.

4.3. Random forest classification

4.3.1. Influences of decision tree sizes on classification accuracy

Figure 3 shows the changes of classification accuracy according to the numbers of decision trees based on the results of experiment one. The classification accuracy of randomForest algorithm stabilizes at the most accurate level, at 12.39%, when the number of decision trees reaches approximately 100. Accordingly, the number of decision trees was set at 100 for training RF classifier and evaluating accuracy in the following experiments. Note that 12.39% (corresponding to around 0.12 on the y-axis at 100 on the x-axis in Figure 3) is very low partly because too many terms among the 67 NSLR terms are similar, leading to the low separability of these terms and thus the low classification accuracy.

4.3.2. Influences of grouping schemes on classification accuracy

To examine how classification accuracy can be improved by the grouping schemes, in experiment two the 69 NSLR terms were merged into seven groups (Table 5) using the three criteria stated in Section 3.4.2.

The resulting groups (Table 5) have the following characteristics.

Table 5. Groups of NSLR terms.

No.	NSLR terms
1	<i>starts and ends in</i>
2	<i>runs along boundary</i> , runs along, along edge, contained in edge, encloses, enclosed by
3	<i>goes to</i> , leads to, goes up to
4	<i>in</i> , inside, within, contained within
5	<i>goes through</i> , bisects, splits, run through, pass through, comes through, cuts across, cuts through, spans, goes across, stretch over, runs across, cuts, intersect, passes, transects, traverses, crosses, break into, divides, separate
6	<i>goes into</i> , reach into, comes into, enters, runs into, ends outside, comes from, goes out of, run from, stretch from, comes out of, exits, leaves, ends in, ends on, starts in, starts outside, connected to, connects, ends at, starts just inside, ends just inside, starts near, starts just outside, ends just outside, ends near
7	<i>near</i> , be adjacent to, bypasses, avoids, goes away from, goes by, outside, entirely outside

- (1) The first group restricts the boundary of a line to be inside the region of a polygon. According to the sketches, the NLSR terms in this group elicit the drawings of a line going out of a region interior and later coming back to its interior. Variable *OAS* may contribute heavily to recognizing this group because *OAS* corresponds to valid values for this group, while is out of its range (e.g. -2) for other groups.
- (2) The second group is associated with a line following closely the outline of a polygon region, thus the variables *La*, *PA*, *ILA*, and *OLA* are important to define these terms. Nonetheless, the role of the topology is vague for these NLSR terms due to the lack of specification on whether the line is in the interior or exterior of the region or on its boundary.
- (3) The third group considers configurations where a line leads to a region, but different people may have different understandings on these terms. For example, for the interior intersection of the line and the region, some subjects consider it as empty while others consider it as non-empty. Except for variables *ILC*, *IAC*, *ILN*, *IAN*, *IPA*, and *ILA*, the others are all in their normal range, but they may be similar with the ones of the seventh group if the interior-boundary intersection is empty for all the samples or with the ones of the sixth group if the intersection is non-empty.
- (4) The semantics of the NLSR terms in the fourth group is perhaps the clearest: the intersection of the interior of a line and the exterior of a region is empty. In this case, variables *EAS*, *OAC*, *OLC*, *OAN*, and *OLN* are out of their ranges and assigned an abnormal value (-2), allowing them to be easily distinguished from the terms with normal values.
- (5) The fifth group represents the configurations where the intersection between the interiors of a line and a polygon region is non-empty. Accordingly, variables *IAS*, *PS*, *OAS*, *La*, *PA*, and *ITS* are valid, while other variables are out of their ranges. Since the exterior-boundary intersection of a line and a region, and the crossing of the interiors of a line and a region, cannot be determined (e.g. for the configuration representing *divide*, a line divides a region into three areas, while for the one representing *bisects*, a line divides a region into two approximately equal areas.), multiple topological types are involved in this group.
- (6) The sixth group exposes limitations on one of the two end-points of a line and express roughly the configuration of a line and a region. For instance, the topology type related to *starts inside* may be R9, R10, R12, R13, R17, R18, or R19 in [Figure A1](#) in the [Appendix](#). Therefore, it is difficult to recognize this group by the topological type alone.
- (7) The characteristics of the seventh group is similar to the fourth group although this group applies to the configurations where a line is outside a region. In this case, variables *IAS*, *IAC*, *ILC*, *IAN*, and *ILN* are out of their ranges, thus they can be distinguished from other terms with normal values. Difference may exist in some relative variables related to the distance or the non-empty boundary intersection between a line and a region.

In summary, the NLSR terms in the same group are much similar in configurations and explanatory variables, while the ones in different groups may share some configurations. Therefore, the grouping scheme is fundamental to training and classification.



To build a better model for classifying NLSR terms, a grouping method is introduced to combine cognitive judgment, feedback of classification result, and separability measure of samples in Section 3.4.2. But different emphasis on the three aspects produce different grouping schemes. This study conducts the classification using two schemes. The first scheme emphasizes the role of cognition. All seven groups in Table 5 are used and the first NLSR term in each group is considered as the group label. The second scheme emphasizes the role of feedback of classification results and five groups are obtained for classification. The third, fifth, and sixth groups in the first scheme are merged into one group in the second scheme, namely, *goes to*, while the other groups are same in the two schemes.

For the first scheme, the classification accuracy is 79.90%. For the second scheme, the accuracy is 87.29%, a significant improvement from the first scheme.

4.3.3. Comparison of different grouping schemes

The two experiments are compared to analyze the influence of grouping schemes on classification accuracies. Before grouping, the accuracy ranges from 10% to 20%, while after the grouping, the accuracy increases to 70–90%, indicating that grouping NLSR terms is reasonable.

4.4. Results of variable importance evaluation

4.4.1. Importance measure of explanatory variables

Since both the grouping schemes achieved relatively high classification accuracy, they are chosen for evaluating the variable importance. Figures 4–10 and 11–15 illustrate respectively the variable importance measurements for the first and the second schemes produced by RF. The horizontal axes refer to the variables while the vertical axes represent the normalized scores of variable importance. Based on these figures, the variables important for recognizing the groups of NLSR terms in the first scheme are as follows: (1) *IAS* and *PS* for the group *goes through* (Figure 4); (2) *ILC* and *ILN* for the group *in* (Figure 5); (3) *OAC*, *OLC* and *OPA* for the group *near* (Figure 6); (4) *OLA* and *OAN* for the group *goes to* (Figure 7); (5) *IAS* and *PS* for the group *goes into* (Figure 8); (6) *OPA*

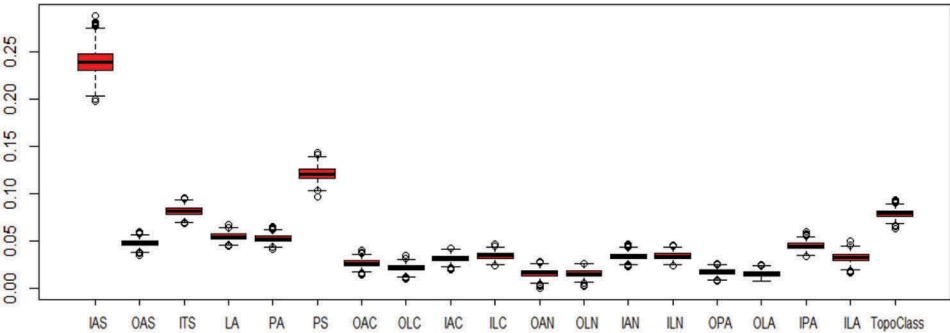


Figure 4. Variable importance measurements of the group of *goes through* in the first grouping scheme.

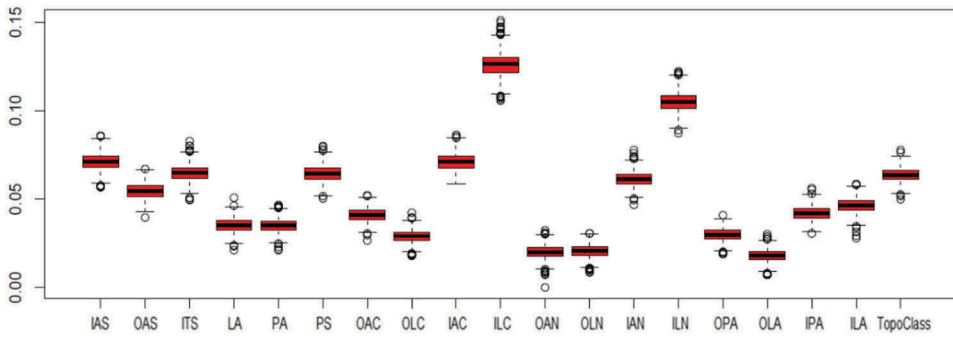


Figure 5. Variable importance measurements of the group of *in* in the first grouping scheme.

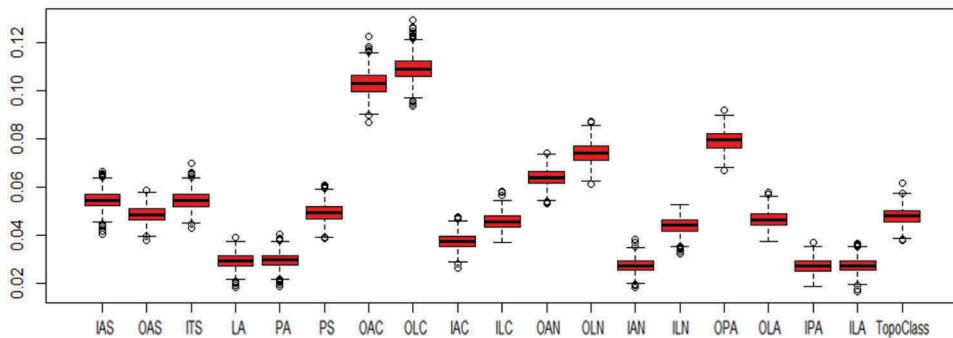


Figure 6. Variable importance measurements of the group of *near* in the first grouping scheme.

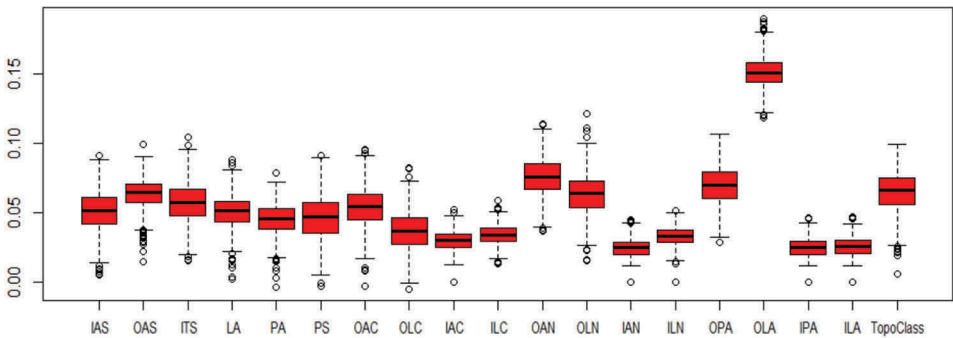


Figure 7. Variable importance measurements of the group of *goes to* in the first grouping scheme.

and La for the group *runs along boundary* (Figure 9); and (7) OAS and IAS for the group *starts and ends in* (Figure 10).

Similarly, the variables important for recognizing the groups of terms in the second scheme are: (1) PS, OLA, and OAS for the group *goes to* (Figure 11), (2) ILC, ILN, and OAS for the group *in* (Figure 12), (3) OLC, OAC, and OPA for the group *near* (Figure 13), (4) OPA, La, and PA for the group *runs along boundary* (Figure 14), and (5) OAS, IAS, and PS for the group *starts and ends in* (Figure 15). The results show that the variables important

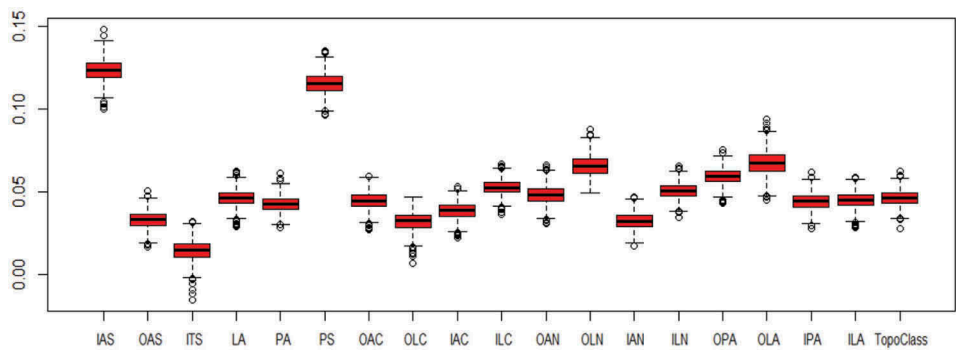


Figure 8. Variable importance measurements of the group of *goes into* in the first grouping scheme.

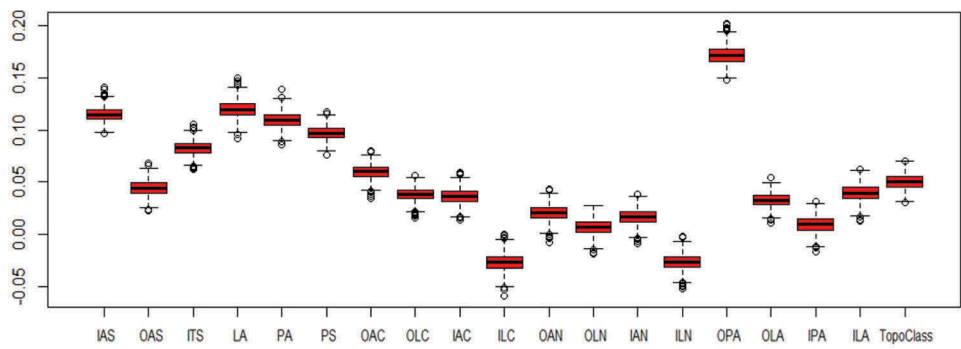


Figure 9. Variable importance of the group of *runs along boundary* in the first grouping scheme.

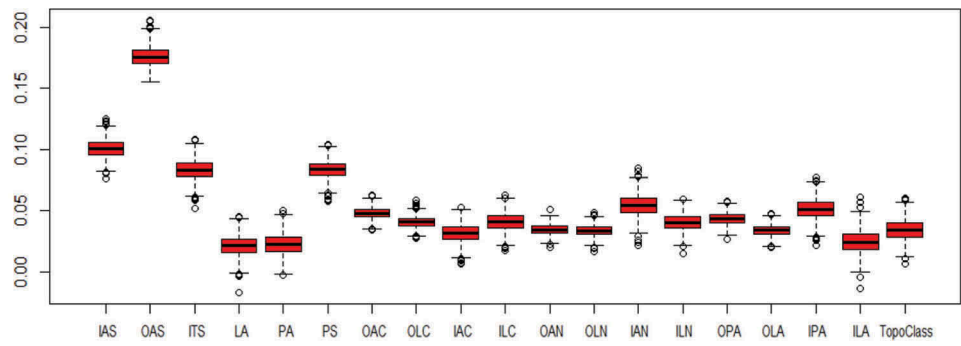
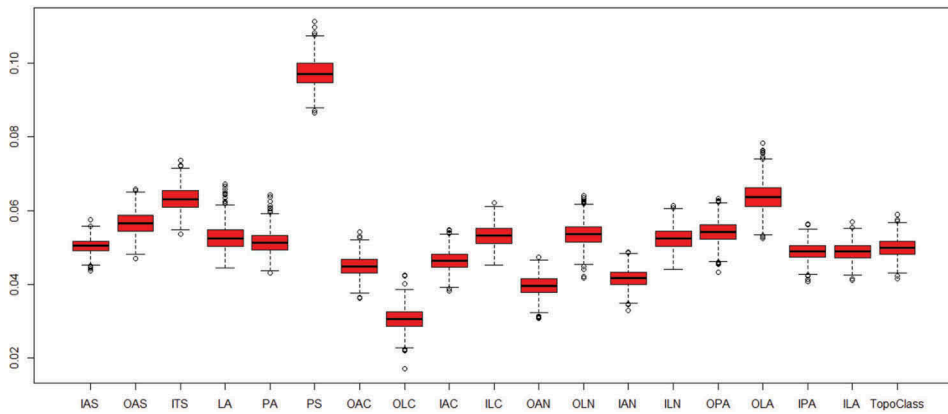


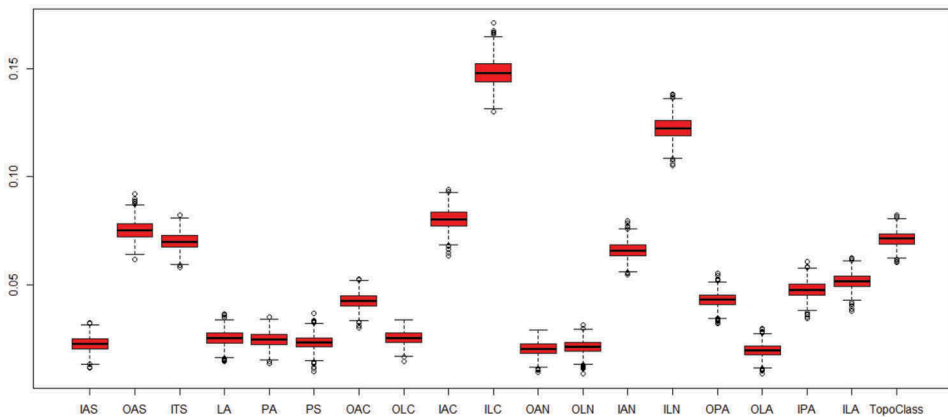
Figure 10. Variable importance of the group of *starts and ends in* in the first grouping scheme.

for groups *in*, *near*, *runs along boundary*, and *starts and ends in* are the same for the two grouping schemes.

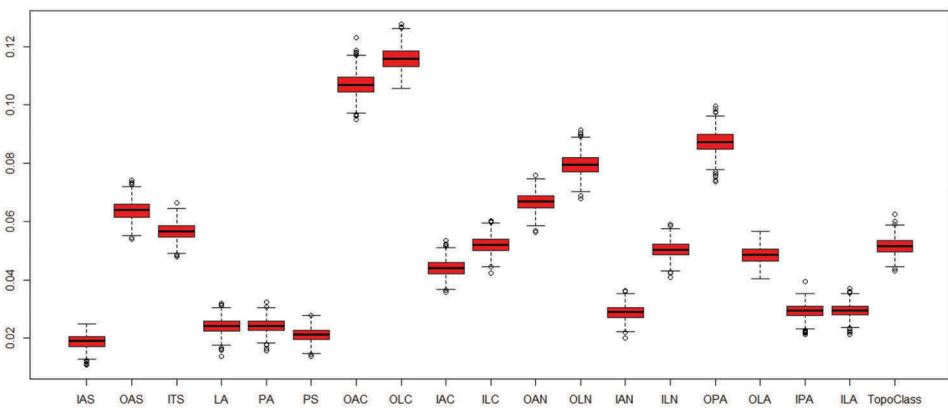
Meanwhile, it can be seen that the variations of the importance of variables for different groups can be significant. Taking the first grouping scheme as an example, variable *OAS* is important for the group *starts and ends in*, but not for the other six groups; and *PS* is important for the group *goes through* and *goes into*, but not for the



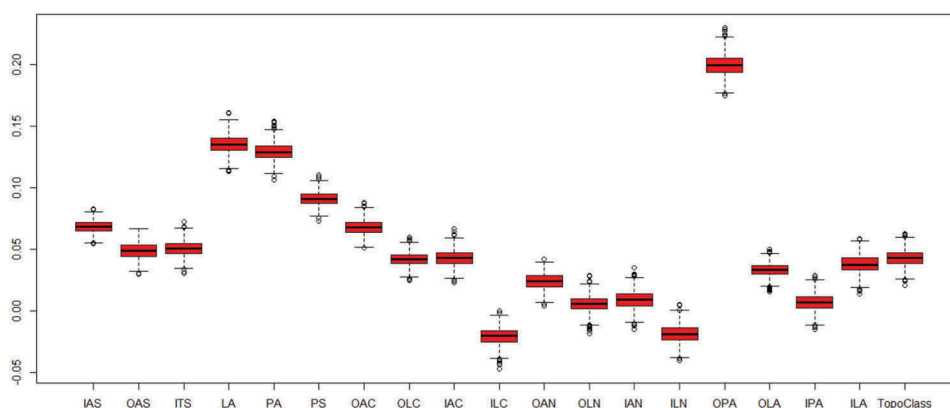
**Figure 11.** Variable importance measurements of the group of *goes to* in the second grouping scheme.



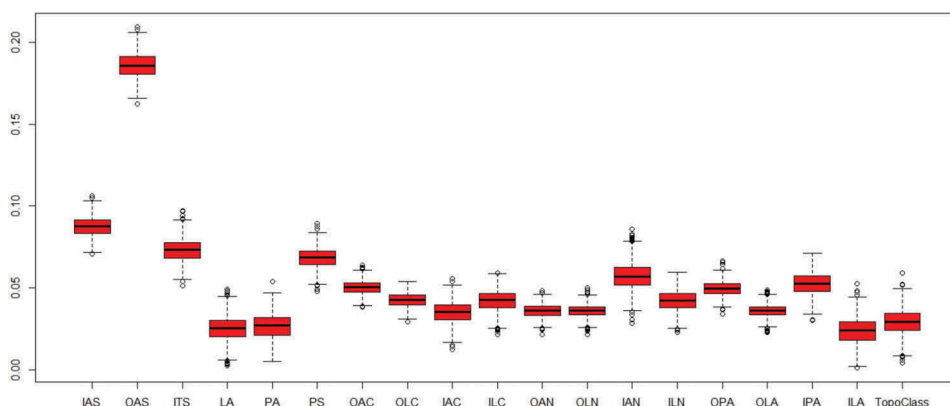
**Figure 12.** Variable importance measurements of the group of *in* in the second grouping scheme.



**Figure 13.** Variable importance measurements of the group of *near* in the second grouping scheme.



**Figure 14.** Variable importance of the group of *run along boundary* in the second grouping scheme.



**Figure 15.** Variable importance of the group of *starts and ends* in the second grouping scheme.

other five groups. Such variations may be attributed to the characteristics of the variables. OAS applies to the configurations where new region(s) in the region's exterior is (are) formed by a line's interior and a region's boundary, while PS applies to the configurations where the intersection of a line's interior or boundary and a region's boundary is non-empty. The configurations where the values of OAS and PS are in the normal ranges correspond to different topological types. To some extent, this implies the dependence of metric variables on the topological types.

In addition, Figures 16 and 17 illustrate the variable importance measurements for all the groups of NLSR terms instead of the ones for each group. The two figures show that variable importance measurements change drastically with different grouping schemes. For the first scheme (Figure 16), the groups *goes to*, *goes into*, and *goes through* are considered separately, which evidently increases the importance of IAS and decreases the importance of OPA. For the second scheme (Figure 17), the importance of IAS decreases considerably while the importance of PS increases considerably. This indicates that variable importance is subject to the criteria of measuring variable importance and the schemes of grouping NLSR terms.

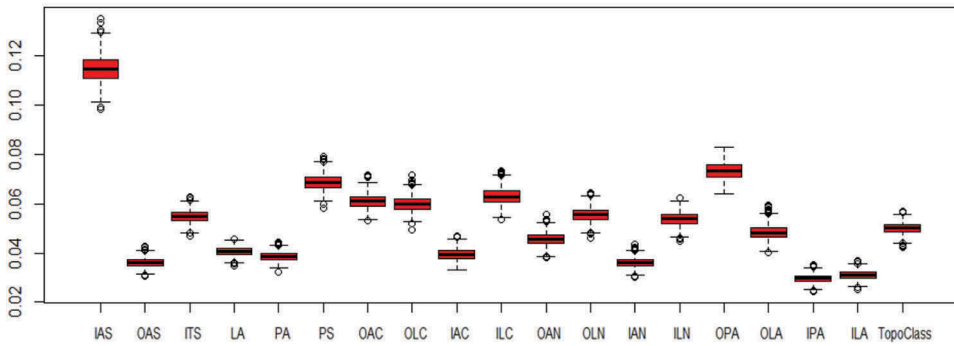


Figure 16. Variable importance measurements for all the groups in the first scheme.

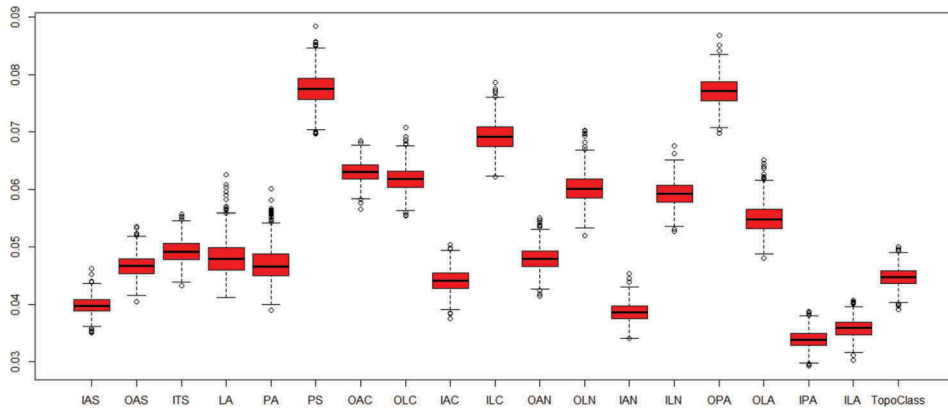


Figure 17. Variable importance measurements for all the groups in the second scheme.

#### 4.4.2. Comparison of importance of topological type and metric variables

According to the variable importance measurements with the original data of grouped NLSR terms in both schemes, the importance of topology type is less than 5%, which is smaller than the importance of metric variables: *PS*, *OPA*, *ILC*, *OAC*, *OLC*, *OLN*, and *ILN*. This indicates that topological types determined by the 9-intersection model have limited roles in classifying NLSR terms, while the metric variables are more important.

### 5. Comparison with existing studies

This study differs from existing studies in several ways. First, a more robust and precise mapping model from geometric representations to NLSR terms is developed using the RF algorithm. Among existing studies, the method used by Shariff *et al.* (1998) only relates maximum, minimum, and median values or the ranges of each variable to the terms and considers each variable separately. Moreover, their method ignores the distribution of each variable, the correlation between variables, and the dependences among the terms. Accordingly, this method cannot effectively and efficiently implement the translation from geometries to NLSR terms. The study by Xu (2007) focuses

on line-line relations instead of line-region relations, and uses decision tree algorithm to relate metric variables to line-line terms. The decision tree algorithm tends to overfit samples and cannot overcome the strong correlation between variables. The RF algorithm, on the contrary, excels in handling high-dimensional and strongly-correlated variables and are more robust than other methods (Breiman 2000, 2002). Moreover, it considers the dependencies among the 69 NLSR terms. Although the much larger number of NLSR terms (i.e. 69) and the extension of topology relations to the line-region configuration, the level of accuracy achieved in this study is comparable to that in Xu (2007). Accordingly, the model established in this study is more accurate than existing studies. Moreover, the work (Breiman 2001) in the field machine learning already has proved that RF classifier is more precise than a single decision tree.

Second, a significantly larger number of samples (over 2000) are used to learn the mapping model and they are more-or-less evenly distributed across the NLSR terms. Existing studies (e.g. Egenhofer and Shariff 1998, Shariff *et al.* 1998) only use dozens of samples and cannot guarantee an even distribution of samples on different terms. Therefore, this study can produce a statistically more reliable mapping model than existing studies.

Third, this study focuses on topology-related NLSR terms, while existing work focuses on directional terms (Skubic *et al.* 2003, Matsakis and Nikitenko 2005, Clementini and Billen 2006, Bartie *et al.* 2013). Specifically, the two studies, i.e., Skubic *et al.* (2003) and Matsakis and Nikitenko (2005), first use F-Histograms to describe quantitative directional relations, and then generate egocentric linguistic descriptions about directions for robots navigation, while the study by Bartie *et al.* (2013) adopts projective relations to describe direction relations for location-based services. Note that the directional terms in these studies are modeled from a egocentric view instead of geographic view. Generally, the egocentric terms take the observers as centers to define the direction relations of other objects relative to observers, thus this kind of terms is suitable at small scales. However, this study mainly addresses the topology-related NLSR terms and is based on a large or geographic scale. The work by Clementini and Billen (2006) presents a model for describing ternary projective relations between regions, which is based on the collinearity invariant of three points. However, our study is interested in binary topological terms which are totally different from the ternary projective relations (Clementini and Billen 2006). Finally, this study focuses on line-region relations, while existing studies focus mainly on regions.

Fourth, a few topology-related fuzzy terms are defined for image understanding, such as *along* (Takemura *et al.* 2012), *surround* (Vanegas *et al.* 2011), *adjacency* (Bloch 2005). These terms are still suitable for regions instead of line-regions, thus they are quite different from the terms in this study, and only a limited number of terms are addressed. More importantly, these terms are formalized separately, i.e., different models are used for different terms, thus the dependencies between terms are ignored, leading to the possibility that they cannot be distinguished in some complex cases. Contrarily, this study examined the 69 terms and adopts a unified model to distinguish all the terms. Accordingly, it can work well in complex cases.

Finally, there is only one existing study that addresses the issue on variable importance evaluation about NLSR terms (Shariff *et al.* 1998). However, the standard scores

used in their studies do not consider the distribution of the values of each variable, the correlation between variables, and the dependences among the terms. This study makes full use of the RF algorithm to evaluate the variable importance by considering the three factors above. The results indicate that the importance of topology types determined by the 9-intersection model is weaker than metric variables, which contrasts to the commonly accepted view of 'topology matters, metric refines' in existing studies. However, the possible reason is that the metric variables contain topological information already, which may reduce the importance of topological type. Interestingly, RF classifier can overcome the correlation between variables, thus it can help to find correct conclusion.

## 6. Conclusions

Meeting the needs of the general public to use geographic information to solve daily affairs has always been the direction of GIS development and research (Egenhofer and Mark 1995). Establishing the relationship between the quantitative, computational and structured languages in GIS and the qualitative and fuzzy natural languages is among the urgent issues to be handled. Based on previous research, this study takes line-region explanatory variable system as example to build up the relationship between quantitative explanatory variable system and qualitative NLSR terms, and obtains good performance.

With the advent of Big Data Era and the development of data mining technique, the source of geographic information is wider and the application of geographic data is more various. And extracting spatial information from natural-language text is among the research focuses (Zhang *et al.* 2009, Loglisci *et al.* 2012). This study provides a feasible method for topological relation translation in natural-language texts by establishing the mapping between the 69 English NLSR terms and the explanatory variable system of line-region configurations.

In the theoretical research of GIS, topological relation draws the most attention. The inference of 'topology matters, metric refines' is widely recognized. Nevertheless, in this study, the importance of topological type determined by the 9-intersection model and other metric variables for NLSR term classification is measured, and the results show that the importance of topological type is smaller than that of metric variables. However, this conclusion may result from the dependence of topological information on metric variables, it is necessary to further address the relative importance measure of topological type and metric variables.

This study establishes a crisp model to classify NLSR terms. Future work will build a fuzzy model to recognize NLSR terms because they are inherently fuzzy in human languages and many collected terms are synonyms, near-synonyms, hypernyms and hyponym. The established fuzzy model should distinguish those terms similar in semantics. In addition, some NLSR terms are related to their surroundings or spatial contexts, which are ignored in this study. Therefore, spatial contexts will be taken into consideration for building a more robust model to handle NLSR terms. Finally, the topology- and direction-based terms will be combined to produce more complicated terms.



## Acknowledgements

The work of the first author is supported by the National Natural Science Foundation of China (No. 41171297). The work of the second author is supported by the National University of Singapore Academic Research Fund (R-109-000-112-112). Comments from the editor and three anonymous reviewers are greatly appreciated.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

This work was supported by the the National Natural Science Foundation of China; [41171297]; the National University of Singapore Academic Research Fund; [R-109-000-112-112].

## References

- Abdelmoty, A.I., et al., 2009. Supporting frameworks for the geospatial semantic web. In: N. Mamoulis, et al., eds. *Advances in spatial and temporal databases, lecture notes in computer science* 5644. Berlin: Springer-Verlag, 355–372.
- Adam, N.R., Shafiq, B., and Staffin, R., 2012. Spatial computing and social media in the context of disaster management. *IEEE Intelligent Systems*, 27 (6), 90–96. doi:[10.1109/MIS.2012.113](https://doi.org/10.1109/MIS.2012.113)
- Bartie, P., Clementini, E., and Reitsma, F., 2013. A qualitative model for describing the arrangement of visible cityscape objects from an egocentric viewpoint. *Computers, Environment and Urban Systems*, 38, 21–34. doi:[10.1016/j.compenvurbsys.2012.11.003](https://doi.org/10.1016/j.compenvurbsys.2012.11.003)
- Bloch, I., 2005. Fuzzy spatial relationships for image processing and interpretation: a review. *Image and Vision Computing*, 23(2), 89–110. doi:[10.1016/j.imavis.2004.06.013](https://doi.org/10.1016/j.imavis.2004.06.013)
- Breiman, L., 2000. Randomizing outputs to increase prediction accuracy. *Machine Learning*, 40 (3), 229–242. doi:[10.1023/A:1007682208299](https://doi.org/10.1023/A:1007682208299)
- Breiman, L., 2001. Random forests. *Machine Learning*, 45 (1), 5–32. doi:[10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324)
- Breiman, L., 2002. *Manual on setting up, using, and understanding random forests v3*. 1. Berkeley, CA: Statistics Department, University of California.
- Clementini, E. and Billen, R., 2006. Modeling and computing ternary projective relations between regions. *IEEE Transactions on Knowledge and Data Engineering*, 18 (6), 799–814. doi:[10.1109/TKDE.2006.102](https://doi.org/10.1109/TKDE.2006.102)
- Cohn, A.G., et al., 1997. Qualitative spatial representation and reasoning with the region connection calculus. *Geoinformatica*, 1 (3), 275–316. doi:[10.1023/A:1009712514511](https://doi.org/10.1023/A:1009712514511)
- Crampton, J.W., et al., 2013. Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, 40 (2), 130–139. doi:[10.1080/15230406.2013.777137](https://doi.org/10.1080/15230406.2013.777137)
- Du, S., Feng, C.-C., and Guo, L., 2015. Integrative representation and inference of qualitative locations about points, lines, and polygons. *International Journal of Geographical Information Science*, 29 (6), 980–1006. doi:[10.1080/13658816.2015.1004333](https://doi.org/10.1080/13658816.2015.1004333)
- Du, S. and Guo, L., 2015. Similarity measurements on multi-scale qualitative locations. *Transactions in GIS* (in press). doi:[10.1111/tgis.12179](https://doi.org/10.1111/tgis.12179)
- Egenhofer, M.J. and Herring, J.R., 1991. Categorizing binary topological relations between regions, lines and points in geographic databases. Unpublished technical report. University of Maine.
- Egenhofer, M.J. and Mark, D.M., 1995. *Naive geography*. In: A.U. Frank and W. Kuhn, eds. *Spatial information theory: a theoretical basis for GIS, lecture notes in computer sciences* 988. Berlin: Springer-Verlag, 1–15.

- Egenhofer, M.J. and Shariff, A.R.B., 1998. Metric details for natural-language spatial relations. *ACM Transactions on Information Systems*, 16 (4), 295–321. doi:[10.1145/291128.291129](https://doi.org/10.1145/291128.291129)
- Leopold, J.L., Sabharwal, C.L., and Ward, K.J., 2015. Spatial relations between 3D objects: the association between natural language, topology, and metrics. *Journal of Visual Languages & Computing*, 27, 29–37. doi:[10.1016/j.jvlc.2014.11.008](https://doi.org/10.1016/j.jvlc.2014.11.008)
- Li, H., Zhang, J., and You, Z., 2003. A separative criterion based on class distance (in Chinese). *Computer Engineering & Application*, 39 (26), 97–99.
- Li, L., Goodchild, M.F., and Xu, B., 2013. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, 40 (2), 61–77. doi:[10.1080/15230406.2013.777139](https://doi.org/10.1080/15230406.2013.777139)
- Loglisci, C., et al., 2012. Toward geographic information harvesting: extraction of spatial relational facts from Web documents. In: J. Vreeken, et al., ed. *IEEE 12th international conference on data mining workshops*, 10 December, Brussels. Los Alamitos, CA: IEEE Computer Society, 789–796.
- Longman, 2016. *The Longman Dictionary of Contemporary English* [online]. Available from: <http://www.ldoceonline.com> [Accessed 20 July 2016].
- Matsakis, P. and Nikitenko, D., 2005. Combined extraction of directional and topological relationship information from 2D concave objects. In: F.E. Petry, V.B. Robinson, and M.A. Cobb, eds. *Fuzzy modeling with spatial information for geographic problems*. Berlin: Springer-Verlag, 15–40.
- Nedas, K.A., Egenhofer, M.J., and Wilmsen, D., 2007. Metric details of topological line-line relations. *International Journal of Geographical Information Science*, 21 (1), 21–48. doi:[10.1080/13658810600852164](https://doi.org/10.1080/13658810600852164)
- OED, 2016. *Oxford English Dictionary* [online]. Available from: <http://www.oed.com> [Accessed 20 July 2016].
- Pang-Ning, T., Steinbach, M., and Kumar, V., 2005. *Introduction to data mining*. London: Addison Wesley.
- Purves, R.S., et al., 2007. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, 21 (7), 717–745. doi:[10.1080/13658810601169840](https://doi.org/10.1080/13658810601169840)
- Shariff, A.R.B., Egenhofer, M.J., and Mark, D.M., 1998. Natural-language spatial relations between linear and areal objects: the topology and metric of English-language terms. *International Journal of Geographical Information Science*, 12 (3), 215–246.
- Shelton, T., Poorthuis, A., and Zook, M., 2015. Social media and the city: rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning*, 142, 198–211. doi:[10.1016/j.landurbplan.2015.02.020](https://doi.org/10.1016/j.landurbplan.2015.02.020)
- Skubic, M., et al., 2003. Generating multi-level linguistic spatial descriptions from range sensor readings using the histogram of forces. *Autonomous Robots*, 14 (1), 51–69. doi:[10.1023/A:1020927503616](https://doi.org/10.1023/A:1020927503616)
- Smart, P.D., et al., 2007. A framework for combining rules and geo-ontologies. In: M. Marchiori, J.Z. Pan, and C. de Sainte Marie, eds. *1st International conference on web reasoning and rule systems, lecture notes in computer science 4524*. Berlin: Springer-Verlag, 133–147.
- Takemura, C.M., Cesar, R.M., and Bloch, I., 2012. Modeling and measuring the spatial relation “along”: regions, contours and fuzzy sets. *Pattern Recognition*, 45 (2), 757–766. doi:[10.1016/j.patcog.2011.06.016](https://doi.org/10.1016/j.patcog.2011.06.016)
- Vanegas, M.C., Bloch, I., and Inglada, J., 2011. A fuzzy definition of the spatial relation “surround” – application to complex shapes. In: S. Galichet and J. Montero, eds. *The 7th conference of the European society for fuzzy logic and technology*, 18–22 July, Aix-Les-Bains. Amsterdam: Atlantis Press, 844–851.
- Xu, J., 2007. Formalizing natural-language spatial relations between linear objects with topological and metric properties. *International Journal of Geographical Information Science*, 21 (4), 377–395. doi:[10.1080/13658810600894323](https://doi.org/10.1080/13658810600894323)
- Zhang, C., et al., 2009. Rule-based extraction of spatial relations in natural language text. In: *International conference on computational intelligence and software engineering*, 11–13 December 2009. Wuhan: IEEE, 1–4. doi:[10.1109/CISE.2009.5363900](https://doi.org/10.1109/CISE.2009.5363900)

Appendix

Table A1. Semantic annotations of the 69 NLSR terms.

NLSR terms	Meanings
along edge	In a line that follows the edge of something; from one end to or towards the other end of the edge of something; beside the long edge of something.
avoids	To keep away from something.
be adjacent to	Next to or near something.
bisects	To divide something into two equal parts.
break into	To be separated into two or more parts.
bypasses	To go around or avoid a place.
comes from	As anything from a source, be derived from.
comes into	To enter or be brought into contact.
comes out of	To emerge from, to extend or lead out of a place.
comes through	To move to or toward something from one end or side of something to the other.
connected to	To join something.
connects	To join together two or more things, to be joined together.
contained in edge	To be had inside the edge of something or as part of the edge.
contained within	To be included within a certain space.
crosses	To lay across another, to place over another, of things to lie or pass across, to intersect.
cuts	Of a line to pass through or across, to cross a line or surface, intersect.
cuts across	To cross, to pass straight through or across, to go across something in order to make your route shorter.
cuts through	To make a path or passage through something by cutting.
divides	To separate or make something separate into parts.
enclosed by	To be bounded on all sides(a portion of space), to be surrounded.
encloses	To surround something, to bound on all sides(a portion of space)
ends on	Placed so as to present the end directly towards any object.
ends at	To carry through to the end at someplace.
ends in	To carry through to the end in a certain space.
ends just inside	To end in the inner part of something by a small amount.
ends just outside	To end on the outer side of something by a small amount.
ends near	To end to, within, or at a short distance, to or in close proximity
ends outside	To end at a position or area adjacent to and beyond the outer side or surface of something.
enters	To go or come in, to go within the bounds of something.
entirely outside	Completely on the outside or outer side of, external to.
exits	To leave (a building, road, etc.); to get out of.
goes across	To go from one side to the other of something with clear limits, such as an area of land, a road, or a river.
goes away from	To leave a place from.
goes by	To go past, pass.
goes into	To enter a position within a space or thing.
goes out of	To be no longer present in something.
goes through	To go from one end or side of something to the other.
goes to	Go towards something.
goes up to	To go from one place to another.
in	Within an area or a space.
inside	Within something, on or to the inner part of something.
intersect	To divide an area by crossing it, of lines to meet or cross each other.
leads to	To direct or guide by going on in advance to someplace.
leaves	To depart from, to go away form.
near	At a nearer distance, with a smaller interval in space.
outside	On the outer side of, external to.
pass through	To continue on one's course through a place.
passes	To go or travel to, into, a place or destination, to go(from one place) to or into another.
reach into	To arrive at or get to inner of something.
run from	To derive from.
run through	To be present in every part of something, to pass quickly through something.
runs across	To meet something.
runs along	Be in line with; form a line along.
runs along boundary	To run following the boundary.
runs into	To merge with, to collide with.

(Continued)

Table A1. (Continued).

NLSR terms	Meanings
separate	To put apart, make a division between.
spans	To stretch right across something, from one side to the other.
splits	To divide, or to make something divide, into two or more parts.
starts and ends in	The origin and end of something is inside something.
starts in	The origin of something is inside something.
starts just inside	To start in the inner part of something by a small amount.
starts just outside	To start on the outer side of something by a small amount.
starts near	To start to, within, or at a short distance, to or in close proximity.
starts outside	To start at a position or area adjacent to and beyond the inner side or surface of something.
stretch from	To start at someplace to spread over an area of land.
stretch over	To stretch from one side to another side, across an open space.
transects	To cut transversely.
traverses	To go or travel across or over.
within	Inside the range or limits of something.

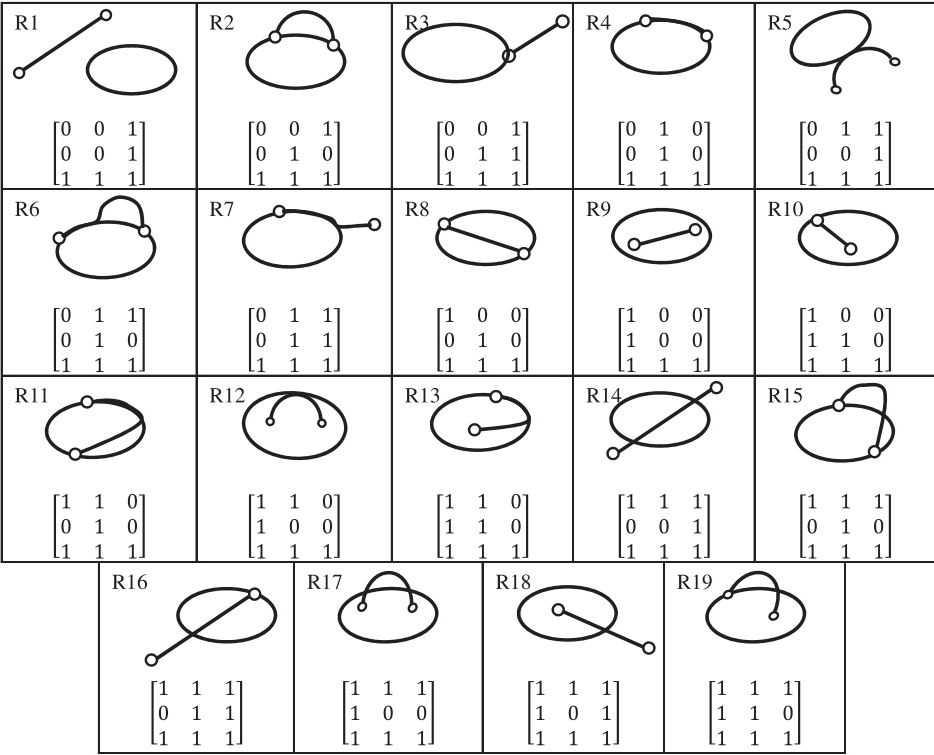


Figure A1. The 19 line-region topological relations.