# Efficient soil loss assessment for large basins using smart coded polygons

**Citation for published version:**
Borthwick, A, Ni, J, Wu, A, Li, T & Yue, Y 2014, 'Efficient soil loss assessment for large basins using smart coded polygons' Journal of environmental informatics, vol. 23, no. 2, pp. 47-57. DOI: 10.3808/jei.201400264

**Digital Object Identifier (DOI):**
10.3808/jei.201400264

**Link:**
Link to publication record in Edinburgh Research Explorer

**Document Version:**
Peer reviewed version

**Published In:**
Journal of environmental informatics

# Efficient Soil Loss Assessment For Large Basins Using Smart Coded Polygons

**Jinren Ni**

Key Laboratory of Water and Sediment Sciences, Ministry of Education, Beijing, PR China; Department of Environmental Engineering, Peking University, PR China

**Ao Wu**

Key Laboratory of Water and Sediment Sciences, Ministry of Education, Beijing, PR China; Department of Environmental Engineering, Peking University, PR China

**Tianhong Li**

Key Laboratory of Water and Sediment Sciences, Ministry of Education, Beijing, PR China; Department of Environmental Engineering, Peking University, PR China

**Yao Yue**

Key Laboratory of Water and Sediment Sciences, Ministry of Education, Beijing, PR China; Department of Environmental Engineering, Peking University, PR China

**Alistair GL Borthwick**

School of Engineering, The University of Edinburgh, The King's Buildings, Edinburgh EH9 3JL, U.K.

**Corresponding Author**

Jinren Ni, Department of Environmental Engineering, Peking University, NO 5 Yiheyuan Road, Beijing 100871, PR China

Email: nijinren@iee.pku.edu.cn

Tianhong Li, Department of Environmental Engineering, Peking University, NO 5 Yiheyuan Road, Beijing 100871, PR China

Email: lth@pku.edu.cn

**Abstract**

Soil erosion is a severe ecological problem. Most conventional methodologies for soil-erosion assessment are appropriate for small or medium river basins. This paper presents an approach to soil-erosion intensity assessment in large basins, utilizing coded polygons identified by spatially overlapping gradation levels of primary environmental factors. Efficient assessment of soil-erosion intensity is achieved by matching the coded polygons to selected polygons pre-assigned to reference groups. A case study is presented for the soil-erosion assessment of the Yellow River Basin. It is found that the calculated and observed soil-erosion intensities are in close agreement for 86% of the total area. Sensitivity analysis indicates that acceptable results are obtained using a 5% sample of the original 9921 coded polygons, thus reducing substantially the computational load. Direct comparisons between the polygon codes in the reference and test groups show that uncertainty is reduced with respect to previous methods. This is confirmed by the reduction in information entropy from 7.49 to 1.33. The proposed approach should be of particular use in the cost-effective assessment of soil erosion in large basins.

**Keywords**

Coded polygons, Soil erosion assessment, Yellow River Basin, Information classification, Semi-quantitative model

## 1. Introduction

Soil erosion causes 84% of land degradation worldwide (Eswaran et al., 2001) and leads to other severe environmental problems such as river sedimentation and non-point pollution (Pimentel et al., 1995; UNEP, 2007; Telles et al., 2011). The global area of land degraded by water erosion covers nearly 1,100 Mha and is predominantly located in Asia and Africa (Oldeman, 1994). In China, the gross quantity of eroded soil exceeds 5 billion tons per year, accounting for about 8% of the World's total (Jing et al., 2005). The Second National Survey of Soil Erosion indicated that 37% of China's land area was affected by water and soil loss, with an even larger area undergoing soil erosion and deposition processes (Jing et al., 2005).

In the 20th Century, the primary factors influencing soil erosion were fully investigated, including precipitation, vegetation, soil type, and land management (Zingg, 1940; Smith and Whitt, 1948; Meyer, 1984). Several empirical models were proposed for assessing the status of soil erosion, based on knowledge of the environmental factors and physical processes involved. The Universal Soil Loss Equation (USLE) was proposed by the U.S. Department of Agriculture (Wischmeier and Smith, 1965; Meyer, 1984), and later revised as RUSLE (Renard et al., 1997). Although USLE/RUSLE has been used worldwide (Wang and Jiao, 1996; Biesemans et al., 2000; Li et al., 2010; Dabney et al., 2011; Xu et al., 2011), it is not always exactly applicable and has occasionally been misused (Wischmeier, 1976; Boardman, 2006). USLE works best for regions in the USA (Stocking, 1995; Vrieling et al., 2002), with amendments necessary for other areas. Moreover, the original USLE model was derived from plot experiments and so is only directly applicable at plot-scale (Terranova et al, 2009; Kinnell 2010). For large-scale applications, the study areas have to be separated into cells or sub-basins until the resulting units are sufficiently small for USLE to be correctly implemented (Millward and Mersey, 1999; Chen et al., 2011; Iroum et al, 2011; Shinde et al., 2011). Ideally, the parameters required for each unit should be derived using 3S technology (Global Positioning System, Remote Sensing, and Geographic Information System). Remote sensing can provide high-resolution images and GIS enables rapid spatial analysis, incorporating the DEM dataset, slope calculations, division of river basins, and so on. However, such data requirements are presently beyond the capabilities of many developing countries in Asia and Africa where soil erosion is particularly severe (Stocking, 1995; Ananda and Herath, 2003; Vrieling, 2006). Physically-based models have been developed, including CREAMS (Chemicals, Runoff and Erosion from Agricultural Management Systems; Knisel, 1980), AGNPS (Agricultural Nonpoint Pollution Source; Young et al., 1989), WEPP (Water Erosion Prediction Project; Nearing et al., 1989), ANSWERS (Areal Nonpoint Source Watershed Environment Response Simulation; Beasley et al., 1980), and HSPF (Hydro-logic Simulation Program Fortran; Johanson et al., 1984). Physically-based models are calibrated through empirical coefficients or exponents for practical applications (Borah and Bera, 2003; Aksoy and Kavvas, 2005), and thus are highly dependent on data accessibility (Boardman, 2006; De Vente et al., 2006), especially when applied to the assessment of large areas (Mutekanga et al., 2010). Semi-quantitative models such as PSIAC (PSIAC, 1968) and FSM (Verstraeten et al., 2003) have less strict data requirements (De Vente and Poesen, 2005; Haregeweyn et al., 2005), but their applications to large basins are still limited owing to the divergence in empirical parameters for different small basins. With the aid of 3S technology, physically-based models (Vrieling, 2006; Tian, 2010) could be used for larger areas, but new challenges arise in how to deal with the massive quantity of data. For DMMP, uncertainty resulting from the discrimination analysis needs to be further minimized.

Ni et al. (2008) proposed a Discrimination Method based on Minimum Polygons (DMMP) for assessment of soil erosion based on the overlay analysis of spatial multi-factors. An erosion index (*EI*) is used for each polygon by multiplying the normalized environmental factors by weights determined using the Analytic Hierarchy Process (Saaty, 1980). Representative polygons are selected and then clustered into reference groups according to erosion grade, whereas the others are assigned to test groups. For each reference group, a discrimination rule is derived between the soil-erosion grades of minimum polygons and their *EI*s in order to assess the soil-erosion severity level within each polygon in the test groups.

This paper proposes a smart coding system (SCS) to encode graded information on each environmental factor. Increasingly large areas are represented by multiple coded polygons derived from the overlay of coded factors,. This permits efficient assessment of the severity of soil erosion in large basins such as the Yellow River Basin.

## 2. Methodology

### 2.1. Classification and Coding Schema for Geographic Information (CCSGI)

Geographic information is often comprehensive and derived from different sources, including maps, numerical data and texts describing geographical entries. To facilitate data handling, Classification and Coding of Information (CCI) transforms geographic information into a set of coding elements via certain prescribed rules. Coding is based on information classification according to independent attributes (Figure 1). Standard methods for CCI include hierarchic classification and faceted classification (SAQSIQ, 2002). For CCSGI, it is supposed that hierarchic classification is suitable for qualitative information, whereas faceted classification is suitable for detailed quantitative information. CCSGI unites qualitative and quantitative information by applying these two classification methods together.

[Place Figure 1 here]

Hierarchic classification is widely used in many fields given that hierarchic structures are commonplace (Boulton and Wallace, 1973; Zheng, 2000; Dale and Wallace, 2005;

Dale et al., 2010). Figure 2 shows the dendrogram structure of a hierarchy with defined levels. In hierarchic classification, the population is divided into $N$ classes, and then each class is further subdivided into independent refined sub-classes at the next level, based on the hierarchic relationships between sub-classes and their node-class. This process repeats until all terminal classes i.e. class-$k$ at level-$j$ (Figure 2) contain enumerable or numeric information that are inappropriate for hierarchic classification but suitable for faceted classification. For a given level of hierarchic classification, a coding template is derived that consists of the terminal classes at this level. The coding template concisely conveys synthetic information concerning the geographic unit, and is represented by the following set:

$$W = \{X \mid X_1, X_2, \ldots X_i, \ldots, X_T\} \tag{1}$$

where $X_i$ is an item in the coding template and $T$ is the dimension of the set or the number of attributes considered.

[Place Figure 2 here]

Each item of this coding template is relatively independent and describes a single attribute of the geographic unit. At different levels of the hierarchic classification, the coding template changes. Therefore, this classification method adapts to different scales at different levels (Dale and Wallace, 2005).

For each item quantified by enumerative or numeric information in the coding template, the faceted classification method is further used to categorize the information into a specific state or facet according to predefined partitioning rules. Each facet or state may represent several enumerable values or a range of detailed values between two thresholds. Hence, information on the population can be reduced to multi-states. Item $X_i$ in set Ω is given as follows:

$$X_i = \{X_i \mid x_i^1, x_i^2, ..., x_i^j, ..., x_i^k\} \quad (2)$$

where $x_i^j$ is the state $j$ of $X_i$; $k$ is the total number of states belonging to $X_i$.

This classification scheme is inherently able to describe the subject domain using simplified quantitative information (Prieto-Diaz, 1991; Herring, 2007). Moreover, a specific numeric code is assigned for each state/facet and considered as a substitute for the source information. As classification information, the code is much more tolerant to data deficiency and inaccuracy than the quantitative numeric information. In other words, faceted classification helps the data requirement to be fulfilled.

In short, a mass of given geographic information is partitioned into $T$ classes by hierarchic classification rules. Subsequently, the codes are obtained by faceted classification rules as follows:

$$C = \{C \mid (c_1, c_2, ..., c_i, ..., c_T), c_1 \in X_1, c_2 \in X_2, ..., c_i \in X_i, ..., c_T \in X_T\} \quad (3)$$

where $c_i$ is the code of element $X_i$ in set Ω.

The code value $c_i$ is either assigned an ordered integer ranging from 1 to $k$, or else values based on its application so as to facilitate easy expansion of the coding system (and hence its usefulness). Adaptability of the code template at different levels in the hierarchical classification facilitates tolerance to data deficiency and inaccuracy; in other words, the CCSGI is self-adaptive at different spatial scales for data of moderate scarcity in a large basin.

### 2.2. Selection, Classification, and Coding of Soil Erosion Environmental Factors

The CCSGI is implemented in the selection, classification and coding of soil erosion environmental factors in order to complete the representation of environmental factors. Although information describing the environmental factors might be scale-dependent, the factors are generally classified under four main headings of climate, topography, soil, and vegetation (Ni et al., 2008). Figure 3 depicts the hierarchical classification scheme of environmental factors systematically selected for soil erosion. Here Level 1 is at the highest level, whereas Level 4 the lowest level in the hierarchy. The attributes at Level 1 are more qualitative than those at lower levels. Macroscopic variables appear at Level 2 corresponding to basin-scale. For example, the climate variable at Level 1 is further specified as annual precipitation at Level 2 for soil loss caused by rainfall. At Level 3, the topographical variables are further specified as length and gradient, and slope pattern. Similarly, the vegetation could be interpreted more specifically than vegetation cover at the lower levels. Attributes representing precipitation, gully density and soil type may remain but be resampled at higher spatial resolution. It should be noted that the rain regime is more important in small than in large basins (Nearing et al 2005, Fang et al 2012).

[Place Figure 3 here]

Table 1 lists the faceted classification codes for each environmental factor at Level 2, based on the standard released by the Ministry of Water Resource (MWR), China (2008), which has been widely cited in the literature (see e.g. Shi et al., 2004; Yang et al, 2005; Fu et al., 2006; Zhou et al., 2008; Liu et al., 2012). Table 1 lists the multi-states and corresponding ranges of values or facets corresponding to each state. For example, annual rainfall less than 300mm is coded as 2; soil erodibility of loess parent material is coded as 5. This makes the categorization scheme more reliable than conventional empirical methods such as simple clustering or equal division (MWR, 2008). Alternative methods like clustering discrimination could be used in cases where standardized classifications of factors such as vegetation type, slope length and slope pattern are lacking (MWR, 2008). For example, cover indices of different vegetation types (SEPA, 2006) could be simply calculated and graded for further coding.

[Place Table 1 here]

## 2.3. Comparison of Coding Sequences

CCSGI produces representations of environmental factors affecting soil erosion, and then SCS compares the derived codes (Figure 4). The code with information on graded environmental factors in a mini-polygon indicates the severity level of soil erosion in the same geographic unit.

[Place Figure 4 here]

For comparison, reference groups are established in terms of coding sequences of environmental factors, and rapid soil-erosion assessment is undertaken as follows.

(i) Coding of Mini-polygon

The mini-polygon is the basic spatial geographical unit for evaluation of soil erosion (Wang, 1993), and is directly derived from the overlay of environmental factors using GIS (Cowen, 1988; Burrough, 1992). By coupling CCSGI with tools in ArcGIS, the geographic information stored in a minimum polygon is further transformed into a coding sequence that is easy to handle. Via CCSGI, geographic maps of the grades of each environmental factor are generated in vector format. Using ArcGIS overlay analysis, a coding-sequence map is produced that contains all graded environmental factors, from which the mini-polygons are generated and coded. Detailed advice on ArcGIS tools is available at ArcGIS Resource Center ( http://resources.arcgis.com ).

(ii) Establishment of the Reference Group

A sample of coded mini-polygons is used to establish the reference groups. The remaining coded mini-polygons constitute the test groups. Random sampling is used for large numbers of coded polygons to ensure the reference groups are representative.

(iii) Matching of Polygons in the Test Group

Matching of coding sequences of test and reference polygons is the key step to predict the severity level of soil erosion in the mini-polygons. To measure the similarity of a pair of coding sequences, a coding sequence with $n$ bits is considered as an $n$-dimensional vector $\mathbf{c} = (c_1, c_2, \cdots, c_j, \cdots, c_n)^T$. Then, the cosine of the vector angle between two coding sequences is calculated from

$$a = \frac{\mathbf{c}_1 \mathbf{c}_2^T}{|\mathbf{c}_1||\mathbf{c}_2|} \tag{4}$$

in which $\mathbf{c}_1$, $\mathbf{c}_2$ are multi-dimensional vectors representing the two coding sequences to be compared. Taking the weights of the different factors into account, equation (4) becomes

$$\alpha' = \frac{\sum_{i=1}^{n} w_i c_{1,i} c_{2,i}}{|\mathbf{c}_1||\mathbf{c}_2|} \tag{5}$$

in which $w_i$ is the weight of factor $X_i$ with respect to soil loss; and $c_{1,i}$, $c_{2,i}$ are elements of vectors $\mathbf{c}_1$ and $\mathbf{c}_2$ respectively.

A series of similarity values $\alpha$ ($\alpha'$) is acquired through comparison of the coding sequences in the test and reference groups. Consequently, similar soil erosion grades are found in the mini-polygons with maximum similarity values.

## 3. Assessment of soil erosion status in the Yellow River Basin

### 3.1. Study Areas and Data Presentation

The Yellow River Basin covers a total area of 795,000 km$^2$. It flows through the Loess Plateau which is experiencing severe soil erosion. As shown in Figure 5, the annual gross rate of hydraulically-induced soil erosion in 1990s exceeded 5000 t/km$^2$ (MWR, 2002).

[Place Figure 5 here]

Referring to CCSGI, information on environmental factors is classified into the attributes at Level 2 in Figure 3. Datasets (i) ~ (v) are described as follows:

(i) Soil-erosion information extracted from 1:1,000,000 digital map of soil-loss intensity based on the 2nd National Soil Erosion Survey conducted in the 1990s by the Ministry of Water Resources, China and used as a data source for World Soil Information (Dijkshoorn et al., 2008). Figure 5 shows the soil erosion zonation map, with 6 grades ranging from slight erosion (Grade 1) to severe erosion (Grade 6).

(ii) Daily rainfall records at 66 hydrological stations in the Yellow River Basin available from 1990 to 1999 via China Meteorological Data Sharing Service System (http://cdc.cma.gov.cn/index.jsp).

(iii) Topography data extracted from a 90m resolution DEM, provided by International Scientific & Technical Data Mirror Site, Computer Network Information Center, Chinese Academy of Sciences (http://datamirror.csdb.cn). The DEM dataset was derived from SRTM (Shuttle Radar Topography Mission) digital elevation data V4.1.

(iv) Soil data from 1:1,000,000 digital map of soil type, provided by the Institute of Soil Science in Nanjing, Chinese Academy of Sciences (http://www.soil.csdb.cn/).

(v) Vegetation data from normalized difference vegetation index (NDVI) raster maps of 8 km resolution for the period from 1990 to 1999, obtained from the Environmental and

Ecological Science Data Center for West China, National Natural Science Foundation of China (http://westdc.westgis.ac.cn, source for this dataset is the VITO (Flemish Inst. Technological Research, Belgium), http://www.vgt.vito). The data form part of the GIMMS (Global Inventory Modelling and Mapping Studies)-NDVI dataset with temporal scale 15-days and spatial scale 8km. The annual NDVI is the averaged value within each year, from which the multiple annual NDVI is further derived.

Within the period of interest from 1990 to 1999, Dataset (i) is used for validation of assessment results of SCS, whereas Datasets (ii) ~ (v) are used as input information of SCS. The data are considered sufficiently accurate if they provide enough information is provided for the coding of each environmental factor based on Table 1.

## 3.2. Assessment Process

### 3.2.1. Data Processing

(i) Rainfall factor: Mean annual rainfall are derived from the daily rainfall at each meteorological station, and then a scatter map is created using ArcGIS with corresponding information on the latitudes and longitudes of the stations. Kriging interpolation is used to obtain a raster map of mean annual rainfall throughout the basin.

(ii) Topographical factors: Datum values of erosion surface elevation, gully density and relative height of terrain are determined using ArcGIS from the DEM (Tang and Yang, 2006).

(iii) Soil factor: Erodibility grades are assigned to different soil types according to the classification rules listed in Table 1.

(iv) Vegetation factor: Vegetation cover ($C$) is obtained from the NDVI map by (Zhao, 2003)

$$C = \frac{NDVI - NDVI_{\min}}{NDVI_{\max} - NDVI_{\min}} \tag{6}$$

where $NDVI_{\min}$ and $NDVI_{\max}$ are the minimum and maximum values of $NDVI$, respectively.

### 3.2.2. Coding and Identification of Mini-polygons

The CCSGI is used to encode the environmental factors by faceted classification. Table 1 indicates how the rainfall, topography and vegetation cover factors are graded according to standard classification rules. Coding maps are derived from the raw data on the environmental factors. All spatial gradation data at different scales are then transformed into vector format. Furthermore, all coded vector maps are overlaid and the mini-polygons generated. Each mini-polygon is identified by a specific coding sequence. The spatial accuracy of yielded polygons is determined by the minimum scale within the maps.

#### 3.2.3. Polygon Matching

The coded minimum polygons are randomly divided into reference and test groups. For each mini-polygon within the reference group, the grade of soil erosion intensity is determined as follows. Six grades of soil-erosion intensities are classified in reference polygons according to the 1990s' survey results. Polygon matching based on coding sequences is then undertaken to determine the soil-loss intensity of the test group. Equation (4) is used to examine the similarity of the coding sequence without considering the weights of the environmental factors. Figure 6 illustrates the pre-processing, coding, and classification procedure as applied to the assessment of soil erosion in the Yellow River Basin.

[Place Figure 6 here]

### 3.3. Evaluation Results

The Yellow River Basin is divided into 9916 coded polygons, of which ~90% of the total area is covered by polygons each of area less than 100 km$^2$, and ~75% by polygons each of area less than 50 km$^2$. Each polygon is represented by a corresponding coding sequence generated from graded environmental factors. Figure

7 shows the soil erosion intensity with a sample ratio ($SR$) of 5%, i.e. ratio of the number of coded polygons in reference groups to the total number of coded polygons.

[Place Figure 7 here]

To quantify the degree of consistency between the calculated and observed results, a variable defined as area overlap ratio ($R$) is introduced as follows:

$$R_i = \frac{\sum Ac_i}{\sum A_i} \tag{7}$$

where $R_i$ is the overlap ratio of the $i$-th grade soil erosion, $A_i$ is the surveyed area of mini-polygons with $i$-th grade soil erosion over the whole basin area, and $Ac_i$ is the area of mini-polygons with the same calculated and surveyed grades of soil erosion.

Figure 8(a) presents the area overlap ratios for the six soil erosion grades. The mean value of $R$ is about 86.1% (with a standard error of 1.2% for 8 sets of calculations) over the entire Yellow River Basin, whilst the minimum value of $R$ is 75% for the sixth grade. The overall accuracy is enhanced by the SCS approach, as is evident by comparison against the average $R$ of 76% by DMMP (Ni et al., 2008) for the same basin with the same input data. For the consistency ratio of each soil erosion grade in terms of the number of coded polygons, the accuracy ratio is 89.1% on average. Figure 8(b) depicts the detailed overlap ratios for each grade, showing that the minimum overlap ratio in terms of the number of coded polygons is 76.9% for the 6$^{th}$ grade of soil erosion intensity.

[Place Figure 8 here]

## 4. Discussion

Based on a Smart Coding System, the relationship has been properly established between environmental factors and soil erosion intensity. For the Yellow River Basin, a sample ratio of 5% achieves an average area overlap ratio of 86.1% with standard error of 1.2% over the whole study area. Moreover, the sensitivity analysis demonstrates that the sample ratio/number can be reduced further, with hardly any effect on prediction accuracy. Meanwhile, the modeling uncertainty also reduces compared to DMMP. SCS is not only applicable to larger basins but also more efficient through data compression via CCSGI.

### 4.1. Sensitivity Analysis of Sample Ratio/Number

A sensitivity analysis is undertaken to examine the influence of sample ratio/number on the predicted results. Figure 9 shows the change of mean area overlap ratio ($R$) as sample ratio ($SR$) is increased from 0.2% to 15%. At least 8 simulations are carried out for each $SR$ to avoid uncertainty from random sampling. It can be seen that $R$ increases monotonically whereas the standard error decreases with increasing $SR$. For $SR > 5\%$, $R$ and its standard error reach 95% and 0.5% respectively.

[Place Figure 9 here]

The relationship between the mean value of $R$ and the sample number ($SN$) of coded polygons in the reference group is investigated to test the minimum number of coded polygons required for satisfactory prediction of soil loss intensity. There is a positive correlation between $R$ and $SN$ (Figure 9). An overlap ratio of $R \sim 80\%$ is achieved for $SN \sim 200$, whereas further increase of $SN$ does not lead to any significant gain in $R$. To reduce workload, $SN = 200$ is sufficient as a reference value.

**4.2. Uncertainty of Assessment**

Similarity between coded polygons is related to uncertainty in application of the SCS, and is quantified using the vector cosine between each pair of coding sequences derived from CCSGI. The closer to unity the cosine value, the more reliable is the matching result. Figure 10 presents a histogram illustrating the percentages of coded polygons with different similarity bands; the values of similarities range from 0.96 to 1 with the majority close to 1. This distribution of similarities implies the assessment is highly reliable. SCS seems to have more advantages over discrimination analysis for assessing test groups (Ni et al., 2008) through discrimination using geographical information and reduction in uncertainty. A distance index, denoted $DI = \frac{|EI - EI_0|}{EI_0}$ where $EI$ is the erosion index of a test polygon and $EI_0$ is the central value of within its matched group, is now used to measure the relative distance from $EI$ to $EI_0$ and hence to indicate the uncertainty of the matching results. As $DI$ approaches 0, the matching result is more accurate (and less uncertain). Figure 11 plots the cumulative percentage of the number of *DI* values determined using discrimination analysis. Here, *DI* is generally not close to 0, with more than 50% of values greater than 0.5, and 20% greater than 1.

[Place Figure 10 here]

[Place Figure 11 here]

Information entropy is introduced to quantify the uncertainty of the assessed results derived from the DMMP and the SCS. Information entropy $\varphi$ indicates the uncertainty of information *Xi* based on its probability distribution *p(Xi)* as follows (Shannon, 1948; Li and Du, 2005):

$$\varphi = -\sum [p(X_i)\log_2 p(X_i)] \tag{8}$$

Larger information entropy means greater uncertainty. The calculated information entropies of coded-polygon *DI*s and similarities are $\varphi$ = 7.49 and $\varphi$ = 1.33 for DMMP and SCS respectively, confirming the higher reliability of SCS based on coding sequences.

### 4.3. Efficiency for Large Basins

SCS reduces data redundancy and hence promotes efficiency of data processing. For example, the number of polygons in the whole Yellow River Basin is reduced by nearly 90% (from 81,054 in DMMP to 9916 in SCS). For a given number $N$ of basin polygons and a sample ratio $SR$, the number of matches has previously been calculated from $N_m = SR(1- SR)N^2$. When $N$ is reduced by 90%, $N_m$ accounts for only 1.5% of the original number of matches required before CCSGI is implemented. Improved efficiency is to be expected as the number of polygons increases. By setting a sample ratio, the reduction in the total number of polygons also leads to a decrease in the number of polygons in reference group. For the Yellow River basin, only 200 coded polygons in the reference group are needed as matching polygons in the test group. SCS is therefore potentially useful for a cost-effective assessment of soil erosion in large basins.

## 5. Conclusions

Efficient assessment of soil loss is essential for sustainable river basin management. This paper proposes an approach based on a smart geo-coding system coupled with a rapid soil loss assessment framework. The system encodes the graded environmental factors in a generated polygon and thereby determines the soil erosion intensity in the polygon. Following the basic assumptions underpinning SCS, the soil erosion intensity values in polygons of the test group should be similar to corresponding values in polygons of the reference group, provided similar coding sequences are implemented. When SCS is applied to assessment of soil erosion intensity throughout the entire Yellow River Basin, satisfactory agreement is reached between the expected and observed results for about 86% of the total area. Sensitivity analysis indicates that the number of samples in the reference groups can be greatly reduced without loss of accuracy. Herein, reliable results are obtained using less than 200 reference samples from the 9916 coded polygons, which implies that only 2% representative polygons are required to ensure accurate assessment. SCS inherits most of the advantages of DMMP, including loose data requirement. By a simple coding-sequence matching of the polygons in reference and test groups, SCS significantly reduces computational load and uncertainty. SCS offers an alternative method for cost-effective assessment of soil loss or conservation in large river basins.

**Acknowledgments**

This work was supported by the National Natural Science Foundation of China with Grant No.51379010. Support from Collaborative Innovation Center for Regional Environmental Quality is also acknowledged

**References**

Aksoy, H. and Kavvas, M.L. (2005). A review of hillslope and watershed scale erosion and sediment transport models. Catena. 64(2-3), 247-271. doi:10.1016/j.catena.2005.08.008.

Ananda, J. and Herath, G. (2003). Soil erosion in developing countries: a socio-economic appraisal. Journal of Environmental Management. 68(4), 343-353. doi:10.1016/S0301-4797(03)00082-3.

Beasley, D.B., Huggins, L.F. and Monke, E.J. (1980). ANSWERS: a model for watershed planning. Transactions of the ASAE. 23(4), 938-944.

Biesemans, J., Meirvenne, M.V. and Gabriels, D. (2000). Extending the RUSLE with the Monte Carlo error propagation technique to predict long term average off-site sediment accumulation. Journal of Soil and Water Conservation. 55(1), 35-42.

Boardman, J. (2006). Soil erosion science: reflections on the limitations of current approaches. Catena. 68(2-3), 73-86. doi:10.1016/j.catena.2006.03.007.

Borah, D.K. and Bera, M. (2003). Watershed-scale hydrologic and nonpoint-source pollution models: review of mathematical bases. Transactions of the ASAE. 46(6), 1553-1566.

Boulton, D.M. and Wallace, C.S. (1973). An information measure for hierarchic classification. The Computer Journal. 16(3), 254-261.

Burrough, P.A. (1992). Development of intelligent geographical information systems. International Journal of Geographical Information Systems. 6(1), 1-11. doi:10.1080/02693799208901891.

Chen, T., Niu, R.Q., Li, P.X., Zhang, L.P. and Du, B. (2011). Regional soil erosion risk mapping using RUSLE, GIS, and remote sensing: a case study in Miyun Watershed, North China. Environmental Earth Sciences. 63(3), 533-541. doi:10.1007/s12665-010-0715-z.

Cowen, D.J. (1988). GIS versus CAD versus DBMS: what are the differences? Photogrammetric Engineering and Remote Sensing. 54, 1551-1555.

Dabney, S.M., Yoder, D.C., Vieira, D. and Bingner, R.L. (2011). Enhancing RUSLE to include runoff-driven phenomena. Hydrological Processes. 25(9), 1373-1390. doi:10.1002/hyp.7897.

Dale, M.B. and Wallace, C.S. (2005). Hierarchical clusters of vegetation types. Community Ecology. 1(6), 57-74. doi:10.1556/ComEc.6.2005.1.7.

Dale, P.E.R., Dale, M.B., Dowe, D.L., Knight, J.M., Lemckert, C.J., Choy, D.C.L., Sheaves, M.J. and Sporne, I. (2010). A conceptual model for integrating physical geography research and coastal wetland management, with an Australian example. Progress in Physical Geography. 34(5), 605-624. doi:10.1177/0309133310369617.

De Vente, J. and Poesen, J. (2005). Predicting soil erosion and sediment yield at the basin scale: scale issues and semi-quantitative models. Earth-Science Reviews. 71(1-2), 95-125. doi:10.1016/j.earscirev.2005.02.002.

De Vente, J., Poesen, J., Bazzoffi, P., Van Rompaey, A. and Verstraeten, G. (2006). Predicting catchment sediment yield in Mediterranean environments: the importance of sediment sources and connectivity in Italian drainage basins. Earth Surface Processes and Landforms. 31(8), 1017-1034. doi:10.1002/esp.1305.

Dijkshoorn, J.A., Van Engelen, V.W.P. and Huting, J.R.M. (2008). Soil and landform properties for LADA partner countries (Argentina, China, Cuba, Senegal and the Gambia, South Africa and Tunisia). ISRIC and GLADA report, ISRIC-World Soil

Information and FAO, Wageningen.

Eswaran, H., Lal, R. and Reich, P.F. (2001). Land degradation: an overview. Proc. 2nd. International Conference on Land Degradation and Desertification, Thailand, Khon Kaen, 20-35.

Fang, N.F., Shi, Z.H., Li, L., Guo, Z.L., Liu, Q.J. and Ai, L. (2012). The effects of rainfall regimes and land use changes on runoff and soil loss in a small mountainous watershed. Catena. 99, 1-8. doi: 10.1016/j.catena.2012.07.004.

Fu, B.J., Zhang, Q.J., Chen, L.D., Zhao, W.W., Gulinck, H., Liu, G.B., Yang, Q.K. and Zhu, Y.G. (2006). Temporal change in land use and its relationship to slope degree and soil type in a small catchment on the Loess Plateau of China. Catena. 65(1), 41-48. doi:10.1016/j.catena.2005.07.005.

Haregeweyn, N., Poesenb, J., Nyssena, J., Verstraeten, G., De Vente, J., Govers, G., Deckers, S. and Moeyersons, J. (2005). Specific sediment yield in Tigray-Northern Ethiopia: assessment and semi-quantitative modeling. Geomorphology. 69(1-4), 315-331. doi:10.1016/j.geomorph.2005.02.001.

Herring, S.C. (2007). A faceted classification scheme for computer-mediated discourse. Language@ Internet. 4(1), 1-37.

Iroumé, A., Carey, P., Bronstert, A., Huber, A. and Palacios, H. (2011). GIS application of USLE and MUSLE to estimate erosion and suspended sediment load in experimental catchments, Valdivia, Chile. Revista Técnica de la Facultad de Ingenieria Universidad del Zulia. 34(2), 119-128.

Jing, K., Wang, W.Z. and Deng, F.L. (2005). Soil Erosion and Environment in China. Science Press, Beijing.

Johanson, R.C., Imhoff, J.C., Davis, H.H., Kittle, J.L. and Donigian, A.S. (1984). Hydrologic Simulation Program-Fortran (HSPF): user's manual for Release 8. EPA, Environmental Research Laboratory, Athens, Georgia.

Kinnell, P.I.A. (2010). Event soil loss, runoff and the Universal Soil Loss Equation family of models: a review. Journal of Hydrology. 385(1-4), 384-397. doi:10.1016/j.jhydrol.2010.01.024.

Knisel, W.G. (1980). CREAMS: a field scale model for Chemicals, Runoff and Erosion from Agricultural Management Systems. USDA, Conservation Research Report No. 26, Washington, D.C..

Li, H., Chen, X.L., Lim, K.J., Cai, X.B. and Sagong, M. (2010). Assessment of soil erosion and sediment yield in Liao watershed, Jiangxi Province, China, Using USLE, GIS, and RS. Journal of Earth Science. 21(6), 941-953. doi:10.1007/s12583-010-0147-4.

Li, Y.D. and Du, H. (2005). Artificial intelligence with uncertainty. National Defense Industry Press.

Liu, Y., Fu, B.J., Lu, Y.H., Wang, Z. and Gao, G.Y. (2012). Hydrological responses and soil erosion potential of abandoned cropland in the Loess Plateau, China. Geomorphology. 138(1), 404-414. doi:10.1016/j.geomorph.2011.10.009.

Meyer, L.D. (1984). Evolution of the universal soil loss equation. Journal of Soil and

Water Conservation. 39(2), 99-104.

Millward, A.A. and Mersey, J.E. (1999). Adapting the RUSLE to model soil erosion potential in a mountainous tropical watershed. Catena. 38(2), 109-129. doi:10.1016/S0341-8162(99)00067-3.

Mutekanga, F.P., Visser, S.M. and Stroosnijder, L. (2010). A tool for rapid assessment of erosion risk to support decision-making and policy development at the Ngenge watershed in Uganda. Geoderma. 160(2), 165-174. doi:10.1016/j.geoderma.2010.09.011.

MWR (2002). The bulletin of soil and water loss of china. Report, Ministry of Water Resources of China.

MWR (2008). Standards for Classification and Gradation of Soil Erosion. Ministry of Water Resources of China.

Nearing, M.A., Foster, G.R., Lane, L.J. and Finkner, S.C. (1989). A process-based soil erosion model for USDA-water erosion prediction project technology. Transactions of the ASAE. 32(5), 1587-1593.

Nearing, M.A., Jetten, V., Baffaut, C., Cerdan, O., Couturier, A., Hernandez, M., Le Bissonnais, Y., Nichols, M.H., Nunes, J.P., Renschler, C.S., Souchere, V. and van Oost K. (2005). Modeling response of soil erosion and runoff to changes in precipitation and cover. Catena. 61, 131-154. doi: 10.1016/j.catena.2005.03.007.

Ni, J.R., Li, X.X. and Borthwick, A.G.L. (2008). Soil erosion assessment based on minimum polygons in the Yellow River Basin, China. Geomorphology. 93, 233-252.

doi:10.1016/j.geomorph.2007.02.015.

Oldeman, L.R. (1994). The global extent of soil degradation. In: Soil resilience and sustainable land use. Wallingford, UK: CAB International, 99-118.

Pimentel, D., Harvey, C., Resosudarmo, P., Sinclair, K., Kurz, D., McNair, M., Crist, S., Shpritz, L., Fitton, L., Saffouri, R. and Blair, R. (1995). Environmental and economic costs of soil erosion and conservation benefits. Science. 267(5201), 1117-1123. doi:10.1126/science.267.5201.1117.

Prieto-Diaz, R. (1991). Implementing faceted classification for software reuse. Communications of the ACM. 34(5), 88-97.

PSIAC (1968). Factors affecting sediment yield and selection and evaluation of measures for the reduction of erosion and sediment yield. Report of the water management subcommittee, Pacific Southwest Inter-Agency Committee (PSIAC).

Renard, K.G., Foster, G.R., Weesies, G.A., McCool, D.K. and Yoder, D.C. (1997). Predicting soil erosion by water: a guide to conservation planning with the Revised Universal Soil Loss Equation (RUSLE). National Technical Information Service, United States Department of Agriculture (USDA), Washington, DC.

Saaty, T.L. (1980). The Analytic Hierarchy Process. McGraw-Hill Company, New York.

SAQSIQ (2002). Basic Principles and Methods for Information Classifying and Coding. State Administration of Quality Supervision Inspection in Quarantine, China.

SEPA (2006). Technical Criteria for Eco-environmental Status Evaluation. State Environmental Protection Administration of China.

Shannon, C.E. (1948). The mathematical theory of communication. Bell System Technical Journal. 27, 379-423 and 623-656.

Shi, Z.H., Cai, C.F., Ding, S.W., Wang, T.W. and Chow, T.L. (2004). Soil conservation planning at the small watershed level using RUSLE with GIS: a case study in the Three Gorge Area of China. Catena. 55(1), 33-48. doi:10.1016/S0341-8162(03)00088-2.

Shinde, V., Sharma, A., Tiwari, K.N. and Singh, M. (2011). Quantitative determination of soil erosion and prioritization of micro-watersheds using remote sensing and GIS. Journal of the Indian Society of Remote Sensing. 39(2), 181-192. doi:10.1007/s12524-011-0064-8.

Smith, D.D. and Whitt, D.M. (1948). Evaluating soil losses from field areas. Agricultural Engineering. 29, 394-396.

Stocking, M. (1995). Soil erosion in developing countries: where geomorphology fears to tread. Catena. 25(1-4), 253-267. doi:10.1016/0341-8162(95)00013-I.

Tang, G.A. and Yang, Q. (2006). ArcGIS geography information system space analysis tutorial. Science Press, Beijing.

Telles, T.S., Guimarães, M.F. and Dechen, S.C.F. (2011). The costs of soil erosion. Revista Brasileira de Ciênciado Solo. 35(2), 287-298. doi:10.1590/S0100-06832011000200001.

Terranova, O., Antronico, L., Coscarelli, R. and Iaquinta, P. (2009). Soil erosion risk scenarios in the Mediterranean environment using RUSLE and GIS: an application model for Calabria (southern Italy). Geomorphology. 112(3-4), 228-245. doi:10.1016/j.geomorph.2009.06.009.

Tian, H.Y. (2010). Summary of the application of "3S" techniques in monitoring soil erosion. Proceedings of Symposium from Cross-Strait Environment & Resources and 2nd Representative Conference of Chinese Environmental Resources & Ecological Conservation Society, China, Linyi City, 91-95.

UNEP (2007). Global environmental outlook: environment for development (GEO-4). Valletta: United Nations Environment Programme.

Verstraeten, G., Poesen, J., Vente, D.J., Konincks, X. (2003). Sediment yield variability in Spain: a quantitative and semi-qualitative analysis using reservoir sedimentation rates. Geomorphology. 69(1-4), 315-331. doi:10.1016/S0169-555X(02)00220-9.

Vrieling, A., Sterk, G. and Beaulieu, N. (2002). Erosion risk mapping: a methodological case study in the Colombian Eastern Plains. Journal of Soil and Water Conservation. 57(3), 158-163.

Vrieling, A. (2006). Satellite remote sensing for water erosion assessment: a review. Catena. 65(1), 2-18. doi:10.1016/j.catena.2005.10.005.

Wang, F. (1993). A parallel intersection algorithm for vector polygon overlay. Computer Graphics and Applications, IEEE. 13(2), 74-81. doi:10.1109/38.204970.

Wang, W.Z. and Jiao, J.Y. (1996). Quantitative evaluation of factors influencing soil erosion in China. Bulletin of Soil and Water Conservation. 16(5), 1-20.

Wischmeier, W.H. (1976). Use and misuse of the Universal Soil Loss Equation. Journal of Soil and Water Conservation. 31(1), 5-9.

Wischmeier, W.H. and Smith, D.D. (1965). Predicting rainfall erosion losses from cropland east of the Rocky Mountains. United States Department of Agriculture (USDA), Agricultural Handbook No. 282, Washington, D.C..

Xu, Y.Q., Luo, D. and Peng, J. (2011). Land use change and soil erosion in the Maotiao River watershed of Guizhou Province. Journal of Geographical Sciences. 21(6), 1138-1152. doi:10.1007/s11442-011-0906-x.

Yang, X., Zhang, K., Jia, B. and Ci, L. (2005). Desertification assessment in China: an overview. Journal of Arid Environments. 63(2), 517-531. doi:10.1016/j.jaridenv.2005.03.032.

Young, R.A., Onstad, C.A., Bosch, D.D. and Anderson, W.P. (1989). AGNPS: a nonpoint-source pollution model for evaluating agricultural watersheds. Journal of Soil Water Conservation. 44(2), 168-173.

Zheng, Z. (2000). Constructing X-of-N attributes for decision tree learning. Machine Learning. 40(1), 35-75. doi:10.1023/A:1007626017208.

Zhao, Y.S. (2003). Theory and method of remote sensing application analysis. Beijing: Science Press.

Zhou, P., Luukkanen, O., Tokola, T. and Nieminen, J. (2008). Effect of vegetation cover on soil erosion in a mountainous watershed. Catena. 75(3), 319-325.

doi:10.1016/j.catena.2008.07.010.

Zingg, A.W. (1940). Degree and length of land slope as it affects soil loss in runoff. Agricultural Engineering. 21, 59-64.

**Table 1. Classification and gradation for environmental factors**

| Grade/ Code | Annual rainoff (mm) | Gully density ($km/km^2$) | Erosion base (m) | Relative Height (m) | Soil erodibility | Cover (%) |
|---|---|---|---|---|---|---|
| 1 | | <1 | 0 | <50 | | >90 |
| 2 | <300 | 1~2 | 1000 | 50~200 | black soils, chernozems, alpine/sub-alpine felty soils | 70~90 |
| 3 | 300~600 | 2~3 | 4000 | 200~500 | cinnamon soils, brown earths, yellow-brown earths | 50~70 |
| 4 | 600~1000 | 3~5 | | 500~1000 | yellow earths, red earths, latosols | 30~50 |
| 5 | 1000~1500 | 5~7 | | 1000~1500 | loess parent materials | 10~30 |
| 6 | >1500 | >7 | | >1500 | sandy soils, desert soils, loose weathering materials | <10 |

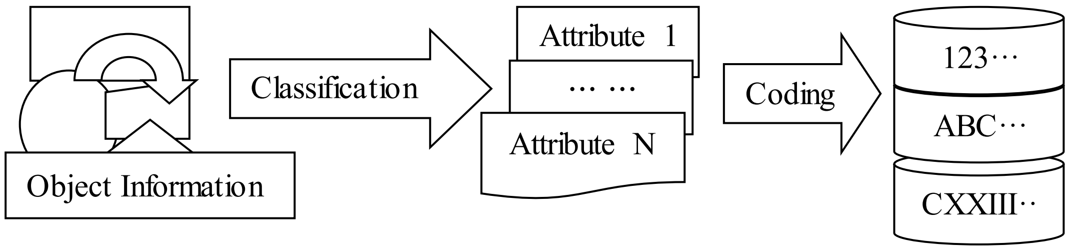



**Figure 1. Classification and Coding of Information (CCI)**

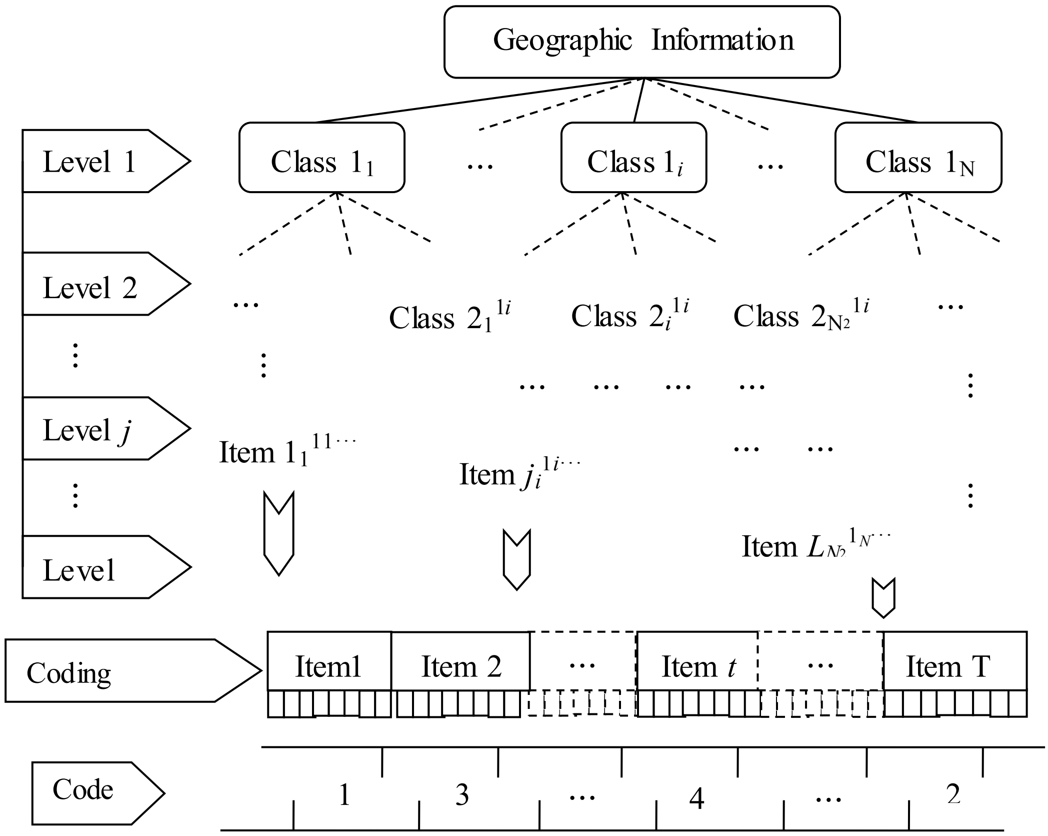



**Figure 2. Classification and Coding Schema for Geographic Information (CCSGI)**

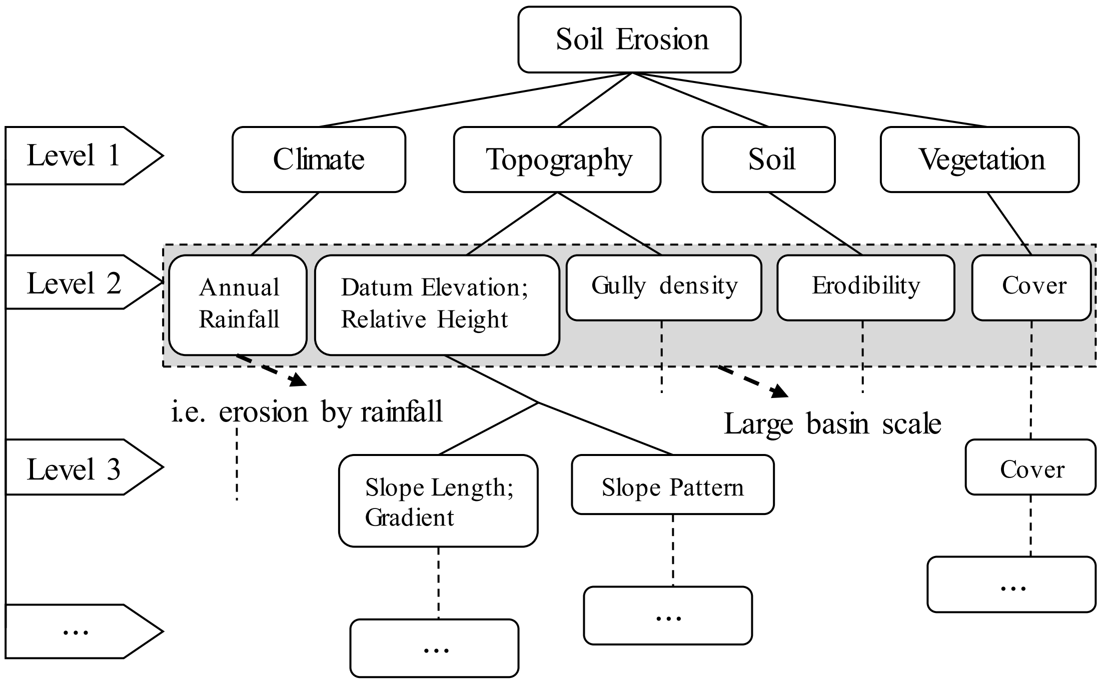
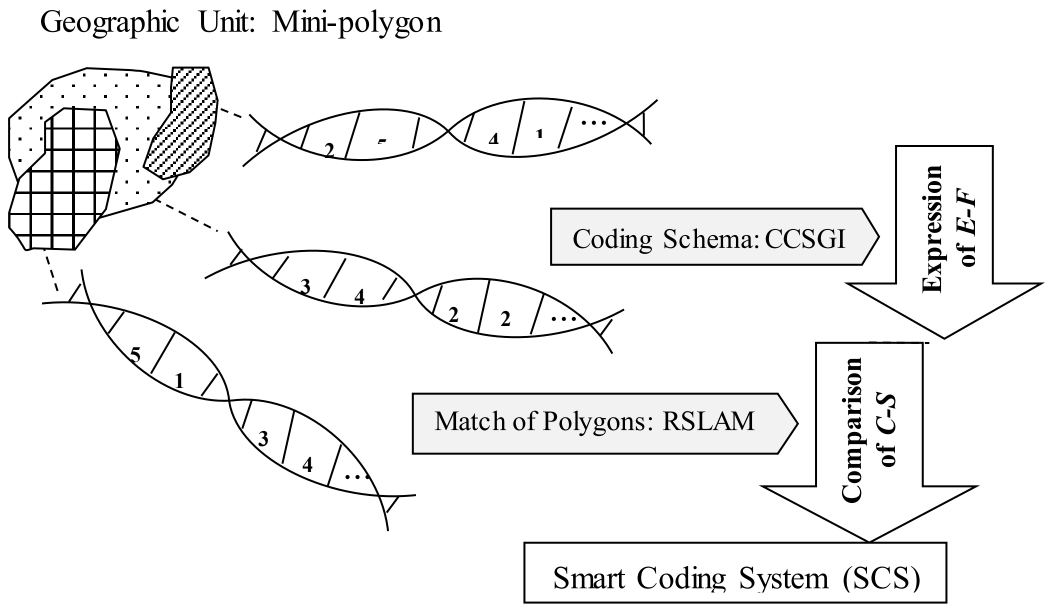



**Figure 3. Hierarchical classification of factors influencing soil erosion**




*E-F*: Environmental Factors; *C-S*: Coding Sequences.

**Figure 4. From Classification and Coding Schema for Geographic Information (CCSGI) to Smart Coding System (SCS)**

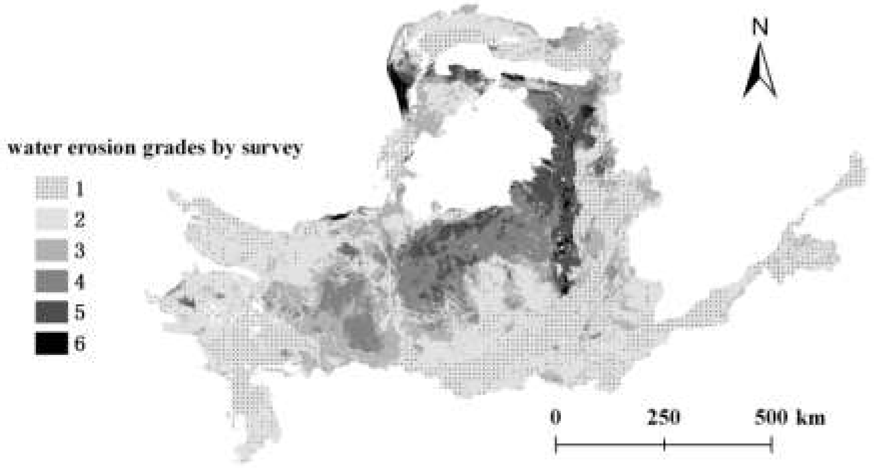



**Figure 5. Water-induced soil erosion in the Yellow River Basin (MWR, 1990s)**

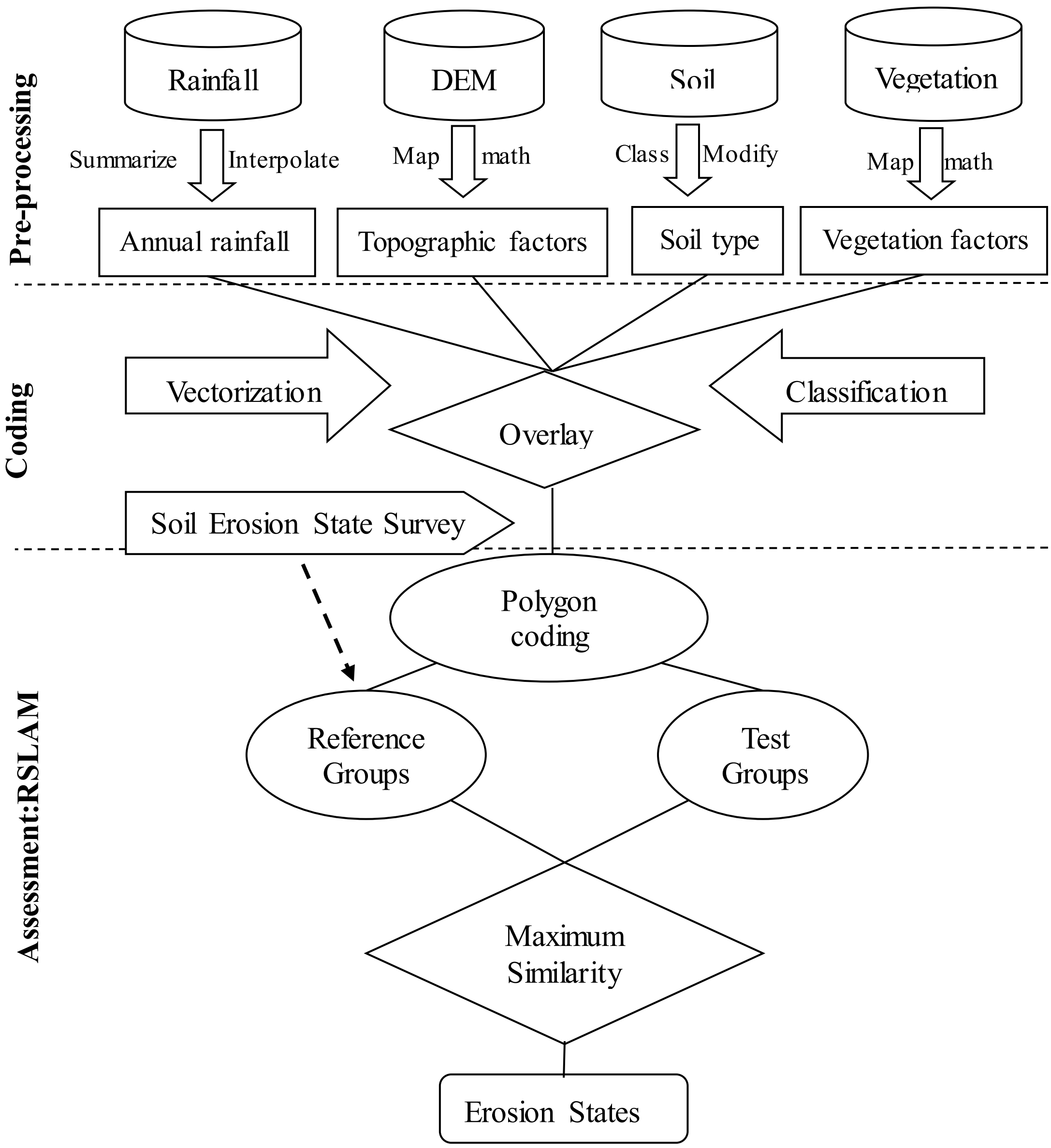



**Figure 6. Smart Coding System (SCS) for soil erosion assessment in the Yellow River Basin**

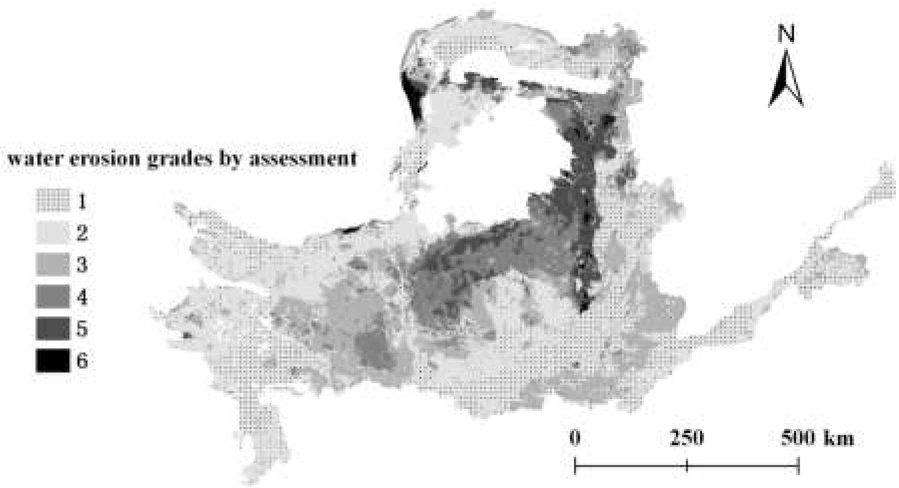



**Figure 7. Spatial distribution of soil erosion intensity in Yellow River Basin from Smart Coding System (SCS)**

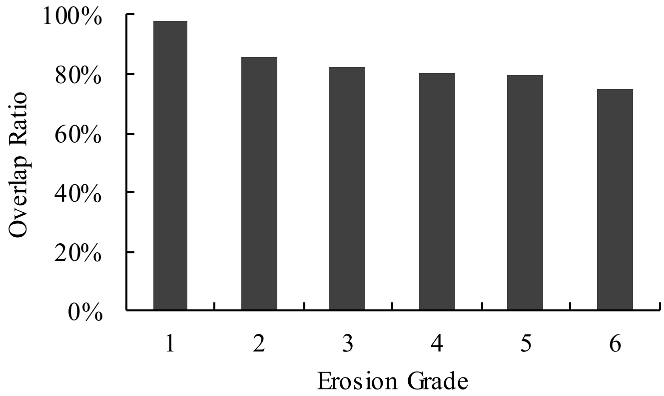



**Figure 8(a) Overlap ratio of observed and calculated areas of the same soil-erosion grade**

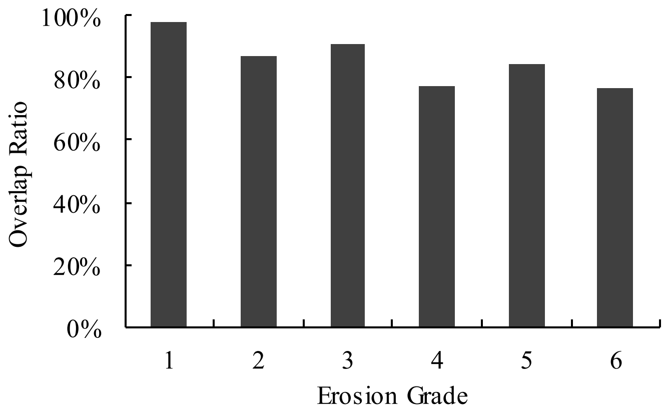



**Figure 8(b) Overlap ratio of observed and calculated numbers of coded polygons of the same soil-erosion grade**

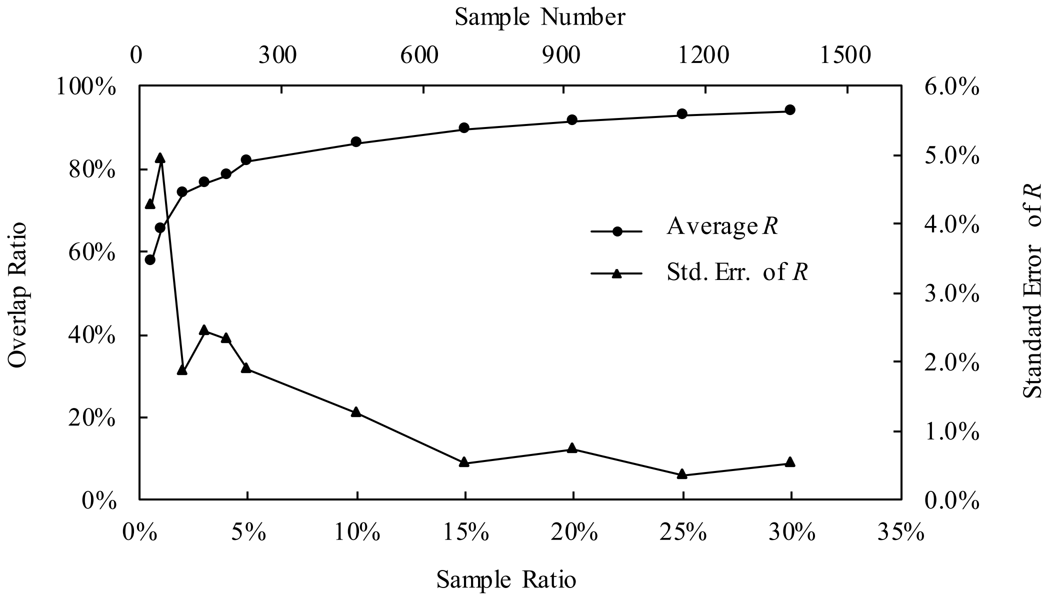



**Figure 9. Sensitivity of *R* with varying sample ratio**

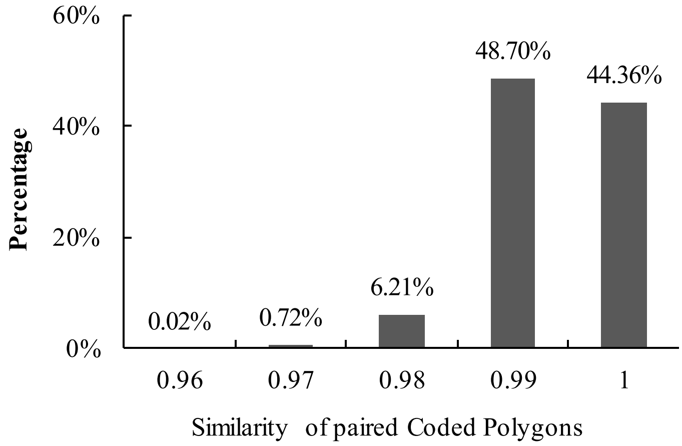



**Figure 10. Percentage distribution of similarities between paired coded polygons**

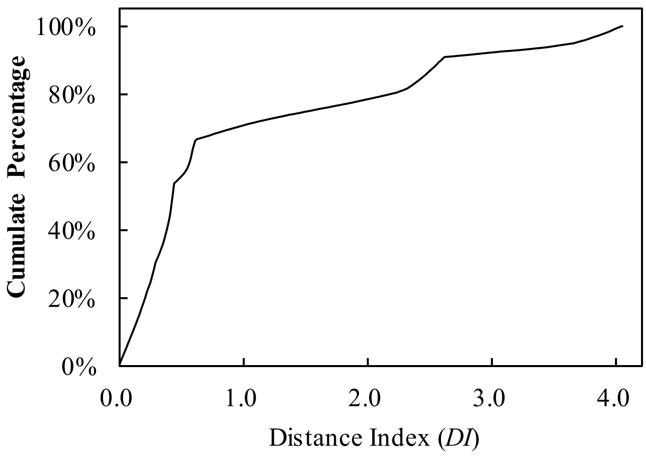



**Figure 11. Percentage distribution of distance index in discrimination analysis**