Methodology for Selection of Framework Data in China

Walter T. de Vries[1] and Wu Lan[2]

[1]International Institute for Geoinformation Science and Earth Observation (ITC), P.O.Box 6, 7500 AA Enschede, Netherlands.
[2]State Bureau of Surveying & Mapping, Beijing, China.

Abstract

China is in the process of developing a National Spatial Data Infrastructure (NSDI) based on so-called framework data. As China has a large number of sectors and users of geospatial data, an NSDI Coordinate Committee was set up in China in 2001 in order to administrate the design and implementation of this NSDI. One of the prominent problems is however still the content of the framework data.

This article describes a methodology for selection of themes and features of geo-spatial data as the framework data in China. This methodology is partly based on experiences in other parts of the world, such as the USA and in the UK, but aims to address the specific needs and users requirements in China. The methodology is founded on a two-staged users needs survey, from which a statistical cluster analysis is conducted. The aim of this analysis is to understand the importance of the features from users' point of view. An agglomerative hierarchical nesting method was chosen to meet the analysis requirements. The results reveal the difference of importance of the features. Two alternatives are derived from the clustering results. These alternatives are finalized for the contents of the framework data in China after refinement by analysis of spatial relationships between features.

I. INTRODUCTION

Geo-spatial data are data defined spatially (in location) by four dimensions (geometry and time) related to the Earth (Groot, 1998). Geo-spatial data are regarded as some of the most critical data underpinning economic and sustainable development, because it is estimated that almost over 80% of all information has a geo-spatial component. However, with regards to this information two main problems exist, which are contradictory in nature. One is the "information explosion", due to the rapid improvement of information technology and acquisition capability of geo-spatial data users and producers. The other is an "information shortage" due to the rapid increasing demands by users who encounter difficulties in accessing and obtaining suitable geo-spatial information for their various applications from the tremendous amount of data resources available. China is no exception to these global developments, and is therefore also encountering many the problems in the field of data sharing. Even though a series of national foundation databases have been developed, still a lot of thematic databases are under construction, and many sectors have started to produce and apply their own geo-spatial datasets without any coordination. As a result, all kinds of data sharing problems have emerged in China.

Administrations in China have been putting a lot of effort to solve these problems. In 2000, the State Council approved to establish the National Spatial Data Infrastructure (NSDI) Coordinate Commission, which is responsible for administrate and coordinate the establishment of the NSDI in China. One of their objectives has been to set up the specifications for framework data at the national level in China. However, how to develop such specifications is still a question, which should be solved in advance. For example questions like: what kinds of data should be included in the framework data? How to prioritise datasets to be included? How accurate should those data be? How much money can be spent? Who should be responsible for the development of whole dataset? Will need to be addressed first.

This article will focus on the development of a methodology of selection of framework data content in China. This methodology is partly based on experiences in NSDI user needs analyses and developments in other parts of the world, but aims primarily at addressing the specific needs and users requirements in China. The methodology is founded on a two-staged users needs survey, from which a statistical cluster analysis is conducted. The aim of this analysis is to understand the importance of the features from users' point of view. An agglomerative hierarchical nesting method was chosen to aggregate the users' requirements such that it becomes clear which are core data and which not. The analysis shows that it is possible to determine priorities of certain features. Two alternatives are derived from the clustering results. These alternatives are finalized for the contents of the framework data in China after refinement by analysis of spatial relationships between features.

II. NSDI AND FRAMEWORK DATA IN CHINA

Before entering in the analysis of user requirements for the framework data, it is important to define the framework data and the context of NSDI. At various levels - global, regional,

1082-4006/03/0901~2-48 $5.00

national, local - NSDIs are being developed and/or improved. Early examples are described by Rhind (1992), McLaughlin 1991) amongst others, while currently there is quite a lot of literature such as by Masser (1998), Groot and McLaughlin (2000) and an increasing number of conference proceedings. The concept of NSDI used in this article is taken from Groot (1998), see Figure 1. Based on this definition, the framework data are considered the backbone of any NSDI, as they as relevant for a large number of applications.

The exact definition and content of framework data has been researched from various perspectives. Generally it is found that the content of these framework data are not equal for every national context, as each national context has a specific national user community with specific interests and requirements. The goals of framework are listed by FGDC (2002), for example. At the global level, the ISO TC 211 Geomatics standardization activity is working on two related areas of endeavors that will assist in the global specification of content models and feature models for framework and non-framework data. At national level, a variable number of data layers may be considered to be common-use and of national or trans-national importance as "framework" data GSDI Cook book (2001), including a variety of layers. Onsrud (2001) conducted a survey "Survey of national spatial data infrastructure activities around the globe ". Some of results related to the contents of the framework data in some countries (Onsrud, 2001). The most commonly found framework data included data representing the Geodetic network, Land surface elevation / topographic data, Digital imagery, Government Boundaries / administrative boundaries, Cadastral / land ownership, Transportation / Roads, Hydrography/ river and lakes and Land use/ Land cover / vegetation.

These results may however not be completely duplicated by China, partly because of its different governmental structures, culture, environment and development objectives, but also largely because of different user requirements. In 1997, the National Geo-spatial Data Standardisation Co-ordination Committee was set up to co-ordinate and guide activities of standardisation related to geo-spatial data around the nation. Four government departments including the State Planning Commission, the Ministry of Science and Technology, the State Bureau of Surveying and Mapping (SBSM) proposed to the State Council to set up ' the National Geographic Information System Co-ordination Committee ' in 1998. Two main research projects on the NSII have been carried out by the State Planning Commission and the Ministry of Science and Technology: ' the Study of the Key Technology of National Resource Environment and District Economy Information System with National Spatial Information Infrastructure ' and "the demonstration of Information Sharing of Sustainable Development of China."

In 2001, the NSDI Coordinate Committee of China was established to guide and coordinate the geo-information production and distribution. In addition, more research has been conducted during past few years, involving most aspects of the NSDI such as data policy, computer network, data, data exchange standards and metadata etc. The SBSM has been implementing the National Foundational Geographic Information System (NFGIS), which is one of the largest national-wide

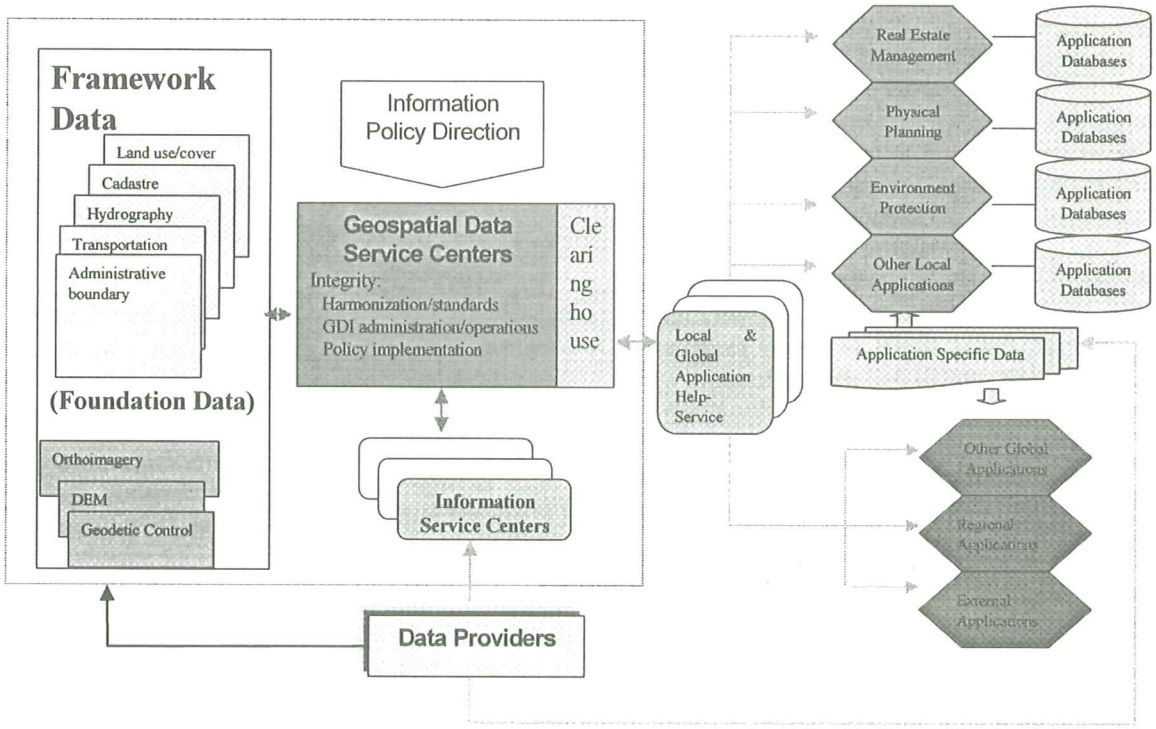

Figure 1. Concept of NSDI (from Groot and Mclanghlin, 2000)

geo-spatial databases. The Topographic Databases at the scale of 1:1,000,000 and 1:250,000 respectively, the Digital Elevation Model, the Geographic Name Databases, the Gravity Databases, and the Geodetic Databases have been established. The Topographic Databases at the scale 1:50,000 and Orthoimage Databases at different level are under construction.

There are a considerable number of other digital thematic data at government agencies, public organizations, academic sectors, and even private sectors in China. Wulan (2002) lists most of these geo-spatial data based on tables from research projects of the State Planning Committee of China. When reviewing these data, some of the problems include:

- Data are kept in many different locations in different systems. These systems are independent from each other, resulting in difficulties to share when needed. The systems are often poorly documented, with few agreed data definitions, no standard formats, and various management systems.
- No common reference system among the distributed databases.
- Lack of targeted polices, laws and regulations to coordinate and guide the geo-spatial data production and applications, resulting in duplication of efforts.
- Geo-spatial information services are still under-used. Users do not know where the geo-spatial data is stored, do not know which geo-spatial data is relevant to their work, and have no efficient means of accessing the data file automatically.
- The lack of geo-spatial data exchange and sharing mechanism causes, on one hand, relative low benefit of geo-spatial data use, and on the other hand, difficulty for some producers to get necessary information from other producers to integrate with or to update their own data bases thus allowing preparing useful data to other users.
- Due to the lack of mechanisms and means to obtain the feedback and demands information from users, the production of geo-spatial data is unfocused, and therefore limits its suitability for applications as well as leads to the waste of geo-spatial information resources.
- Tremendous increase of amount of geo-spatial data sets, due to the advanced technology and increased ability of users, but heterogeneous (different hosts, operating systems, different data sources and data structures).

Based on these constraints there is a high need for selection of framework data, which are based on common agreed standards and which incorporate the user requirements in such way that redundancy of data acquisition and provision is avoided.

III. SURVEY OF USER NEEDS

Given the enormity of available datasets and number of organisations involved, a user needs analysis was carried out through a two-staged process. The two-staged approach consisted of firstly discussions with experts, to identify the most prominent features of discussion. Only when this was finalized, it seemed appropriate to design a more detailed questionnaire and conduct more in-depth interviews. The surveys were conducted in three main areas: Guangzhou, Xi'an and Beijing. At the institutional level, a variety of departments were addressed from various organizations. Moreover, since China's public administration consists of a three-level administrative system, three levels of administration were considered: central, provincial and county level. The survey thus included representations of government, research and education, and private industry.

These initial discussions included 22 experts from various sectors: hydrology, agriculture, forest, transportation, economic statistics, land, marine, environment, meteorology, cultural relic, civil, scientific surveying, geology, mine, seismology, and surveying and mapping, etc.

Based on these initial discussions the original questionnaire was revised in such a way that:

- The starting point for all discussions and feature evaluations should be the topographic map of 1:50,000, since most features are included in 1:50,000 topographic maps.
- The range of the survey should be limited in the fields where the sectors often use the spatial data.
- Respondents could only mark Yes/No for every spatial feature that they would consider of critical importance
- Similar features were grouped into different types. For example, several terrain features such as plain, plateau, basin and desert were grouped together as they were considered to have equal importance geographically.

The questions were asked in two ways:

- Do you think the feature should be regarded as the framework feature?
- If you agree this, it means you agree the properties we considered for the feature. If you consider other properties, please add in the table.

In total there are 73 individuals addressed from different organizations. Among these there were people from different regions, working at various levels within organisations, with different financial budgets or different fields/sectors. Figure 2 shows the statistics of respondents per sector.

From these respondents 35 were at state level, 29 at provincial level and 9 at county level. This distribution seems in line with the use of spatial data, since the spatial data at scale 1:50,000 are generally used in the state level. At provincial level spatial data at scales 1:50,000 and 1:10,000 is commonly used. At county level spatial data are mainly at 1:10,000 scale and even bigger scales.

A draft detailed questionnaire was designed based on exist-

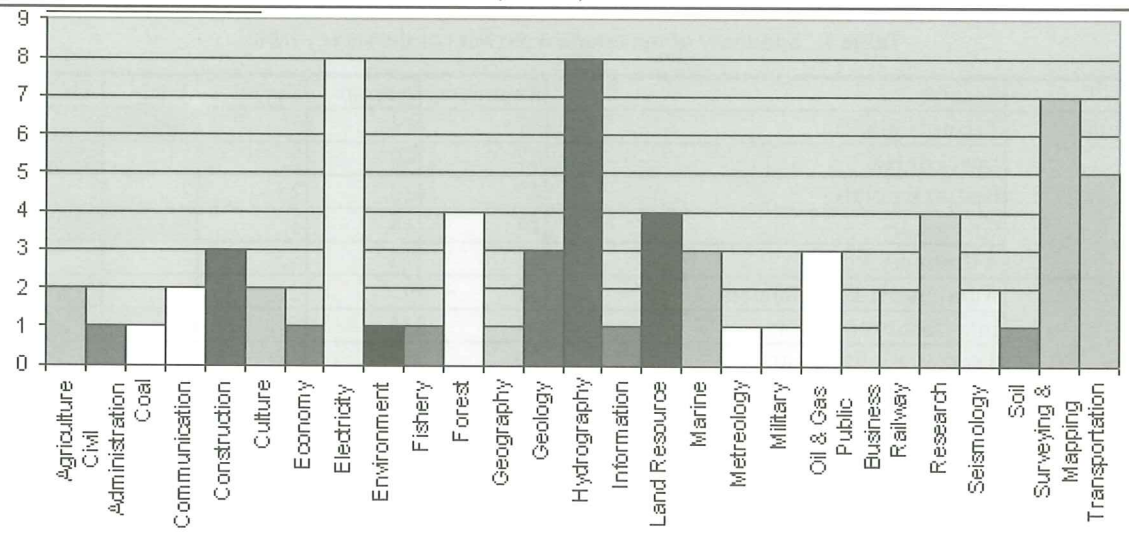

Figure 2. Surveyed organizations (classified according to fields)

ing features of the 1:50,000 topographic maps in China. The interviews included first general questions about the organization of the respondent, followed by more detailed discussions on which features were considered important and which not. The feature selection was carried out by first selecting a feature as important for framework data, followed by a certain weight for the relative importance of this feature compared to other features. In this way the details of the features and its priority for a particular user could be extracted.

To simplify the detailed questions, some features were aggregated into one group. For example, the transportation features were summarized to several groups: main road/railway, second road/railway, bridge, airport, bus station, railway station, tunnel, bank, and other accessories. The hydrology features are generalized by several types also: river/canal, lake and reservoir, dried river, swamp, sand beach, sluice, dam, waterfall, dyke, shore, fountain, well, hydrology station, ferry, dock, shipping assistant signs.

The first columns of Table 1 display a sample of the kind of statistical data that was collected. A list of possible framework data features was presented in such a way that respondents only needed to mark whether a certain feature needed to be included in a framework data set of not (this is reflected by either "accept" or "reject"). Some items in the questionnaire were not answered, mainly due to particular interests of users in their own features only. Although the respondents were requested to mark the properties of their interests, this was not always responded to.

Other than Onsrud (2001) there are no clear methodologies described in geoinformation related literature how to define the number and type of layers needed in each of the framework data sets. Where Onsrud (2001) is based on the actual situation in various countries, there are no explanations in this research how much of the existing layers of geodata are covered by the any of the users databases, and how much could be reduced when certain priorities are distributed over providers. For that reason, a statistical analysis was chosen as a method to select priorities in content of existing layers.

IV. CLUSTER ANALYSIS OF USERS PRIORITIES

Figures 3 and 4 show the distribution of the survey results after normalization. Figure 3 is the distribution grouped by the feature class. There are 10 feature classes each of which contains several "features" - some items are feature classes or important feature attributes (the figure also shows a feature "general class" referring to any remaining features, but this was excluded for further processing). For example, triangle point and level point are real features that belong to class control points. There are also features plain, desert and so on that belong to class terrain, although this is actually a feature sub-class. For reasons of simplicity and convenience, these are, however, all regarded as specific features. The value of vertical axis denotes the percentage of the "accept/agree" of each feature - which in essence denotes the probability of "accept/agree".

Figure 4 represents the same distribution but now in histogram form. The horizontal axis is the ratio of "accept" compared to "reject" per feature, and the vertical axis is the number of features per equal ratio.

Although indicative, the histogram reveals the variety of responses, and shows that a significant larger number of features is accepted than rejected. Yet, where to establish the cut-off point of which features to include and which not, is not immediately obvious, but could be done using the cluster analysis. The statistical theory of cluster analysis is searching for groups (clusters) of data, in such a way that objects belonging to the same cluster resemble each other, whereas objects in different clusters would be significantly different. The survey collected the number of "accept" and "reject" of

Table 1. Summary of the returned answers of the survey table.

Class	Feature	Reponses	Accept	Reject	Alt 1	Alt 2	Alt 3
Control Points	Triangle points	47	44	3			
	Level points	47	40	7			
	Bedrock points	47	22	25			
	Gps points	47	44	3			
	Astronomy Points	47	12	35			
Settlement	Administrative settlement	11	9	2			
	Integrated settlement	45	45	0			
	Separated settlement	39	11	28			
	All settlements	39	31	8			
	Destroyed settlement	33	0	33			
	Settlement type	33	2	31			
	Dense tent, cave	17	17	0			
	Separated Mongolian tent, cave	15	13	2			
Transportation	Main road/railway	55	54	1			
	Second road/railway	46	45	1			
	Minor road in remote area	52	46	6			
	Bridge	52	46	6			
	Airport	46	44	2			
	Bus station	45	19	26			
	Railway station	45	42	2			
	Tunnel	55	39	16			
	Road Bank	55	38	17			
	All roads in topo maps	55	16	39			
Hydrology	River, canal	49	48	1			
	Main lake (1-5), reservoir (1-5) National standard	49	48	1			
	Main river (class 1-7), canal (1-5); national standard	49	48	1			
	River, canal on the topographic map	49	28	21			
	Lake and Reservoir	49	48	1			
	Dried river	49	19	30			
	Swamp	36	19	17			
	Shore	46	17	19			
	Sluice	50	33	17			
	Dam	47	25	22			
	Waterfall	46	18	28			
	Dyke	46	24	22			
	Sand beach	47	22	25			
	Fountain	31	12	19			
	Well	28	11	17			
	Hydrology station	31	15	16			
	Ferry	36	29	6			
	Dock	36	27	8			
	Shipping assistant signs	35	28	6			
Pipe and barrier	High-volt electricity line 110KV in populated area	39	35	4			
	High-volt electricity line 35KV in remote area	37	25	12			
	Communication line	44	16	28			
	Other pipes	45	13	22			
	Fence	28	1	27			

Table 1. To be continued

Class	Feature	Reponses	Accept	Reject	Alt 1	Alt 2	Alt 3
Terrain	Contour and height points	51	51	0			
	Cliff	47	22	25			
	Desert	53	44	9			
	Dune	53	44	9			
	Slide	53	34	19			
	Mud flooding	51	31	20			
	Volcano	53	30	23			
	Snow mountain	53	44	9			
	Cave	32	20	12			
	Peak	31	21	10			
	Fathom line and point	31	15	16			
	Other features	30	10	20			
Building	Industrial well	48	39	9			
	Observation station	46	36	10			
	TV tower	51	45	6			
	Satellite receive station	46	36	10			
	Stadium	43	28	15			
	Temple	46	35	11			
	Water tower, chimney	35	30	5			
	Church	46	34	12			
	Mongolia tent	46	21	19			
	Ancient relic	45	36	9			
	Electricity station	45	36	9			
	Water factory	47	30	17			
	Oil station	47	30	17			
	Water house	46	16	30			
	Grave	46	34	12			
	Thresh field	46	16	30			
	Independent feature	40	28	12			
	School and hospital	49	31	18			
	Other features	49	11	38			
Boundary	Current boundary	56	56	0			
	Add administrative town boundary	56	8	48			
	Special region boundary	56	34	18			
Land cover and land use		70	70	0			
Geographic names	Administrative name	58	58	0			
	Geographic name	57	52	5			
	Other names	49	32	17			

the users with respect to inclusion of certain features in a framework data set. In order to somehow determine which features should be incorporated in the framework data, the survey results are analysed based on aggregation of features into different clusters. A statistical cluster analysis was then used to distinct between more or less important details of features. Generally speaking, there are two categories of methods for clustering Tacq (1997), Snedecor and Cochran (1980):

1). Partitioning Algorithms. A partitioning algorithm describes a method that divides the data set into k clusters, where the integer k needs to be specified by the user. Typically, the user runs the algorithm for a range of k-values. For each k, the algorithm carries out the clustering and also yields a "quality index," which allows the user to select the "best" value of k afterwards. Algorithms of this type include k-means, partition around medoids, fuzzy clustering etc.

2). Hierarchical Algorithms. A hierarchical algorithm describes a method yielding an entire hierarchy of clusterings for the given data set. Agglomerative methods start with the situation where each object in the data set forms its own little cluster, and then successively merges clusters until only one large cluster remains which is the whole data set. Divisive methods start by considering the whole data set as one cluster, and then splits up clusters until each object is separate.

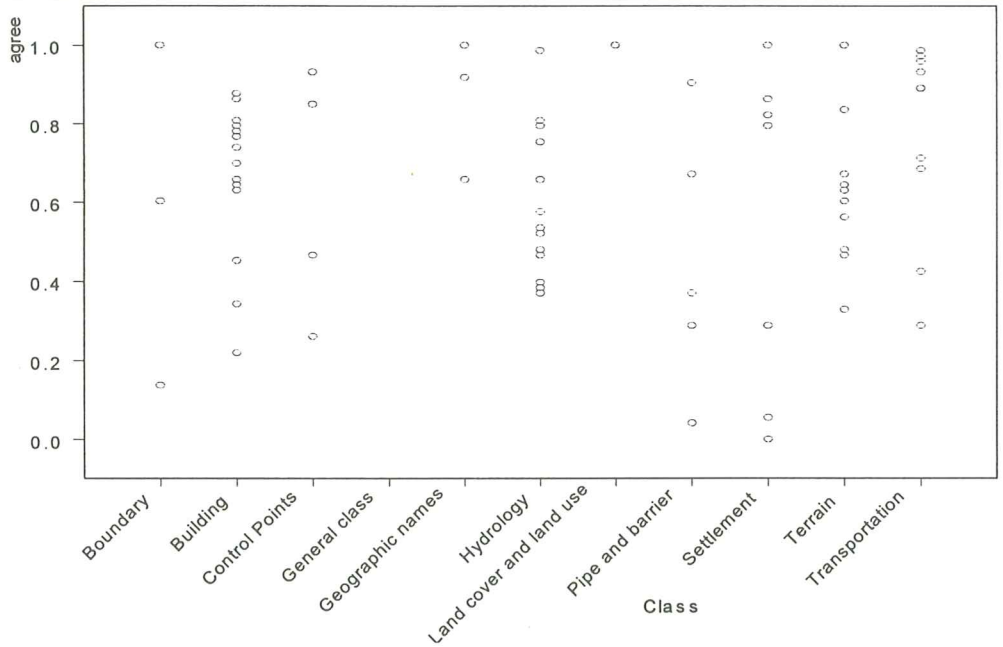

Figure 3. Data distribution based on the feature class

The agglomerative nesting method was chosen to cluster the importance of features. This method is described in Kaufman and Rousseeuw (1990). The dissimilarity between two objects measures "how different" they are, or in the case of this research, how different clusters of important features are. The dissimilarity is described by the following three axioms of a metric (distance functions):

1). d(i, i) = 0
2). d(i, j) ≥ 0
3). d(i, j) = d(j, i)

If a data matrix of survey results is used, the function starts by computing the dissimilarity matrix. Initially (at step 0), each object is considered as a separate cluster. The rest of the computation consists of iteration of the following steps:

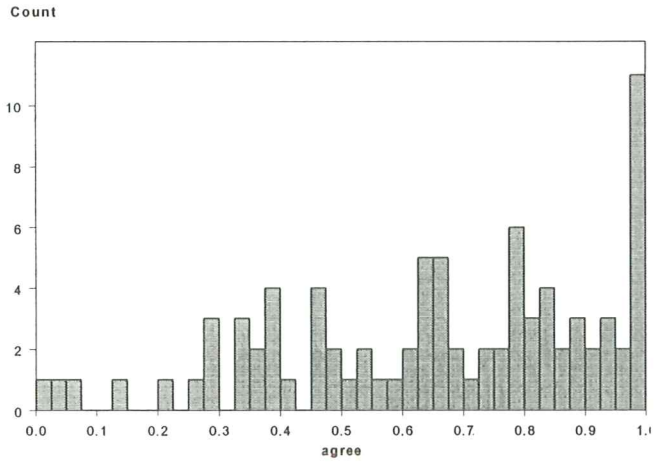

Figure 4. Feature histogram

1). Merge the two clusters with smallest distance - between-cluster dissimilarity;
2). Compute the dissimilarity between the new cluster and all remaining clusters.

The between-cluster dissimilarity can be defined in various ways, notably there are three methods:

1). Group average method

$$d(R,Q) = \frac{1}{|R||Q|} \sum_{i \in R, j \in Q} d(i,j)$$

2). Nearest neighbor method = single linkage method

$$d(R,Q) = \min_{i \in R, j \in Q} d(i,j)$$

3). Furthest neighbor method = complete linkage method

$$d(R,Q) = \max_{i \in R, j \in Q} d(i,j)$$

R and Q represent the clustered groups, and the | | lines indicate the number of elements in either group. The d(i.j) represents the (Euclidian) distance / correlation function between element i of group R and element j of group Q. The hierarchy obtained from this method can be graphically displayed by means of a Clustering tree. This is a tree in which the leaves represent objects. The vertical coordinate of the place where two branches join equals the dissimilarity between the corresponding clusters. For example, in Figure 5 objects 1, 2 and 3 are the leaves, the value of the height stands for the dissimilarity between two groups. In this figure the dissimilarity between object 1 (actually it is a group with one object) and a group composed of objects 2 and 3 is around 0.05.

The clustering process was conducted using the S-plus software, which provides several clustering methods. For the brev-

ity, for any introduction and/or explanation of the software, reference is made to its user manual. One of the main results is clustering tree (Figure 5), which shows the process of clustering graphically. The dissimilarity matrix is calculated by Euclidean distance between features. The between-cluster dissimilarity is calculated based on the group average method that keeps clustering robust and consist [see for method also Kaufman and Rousseeuw (1990)]. In this figure, the number at the bottom represents the feature number, whereas the vertical axis shows the height representing the dissimilarity between the features.

The clustering method does not provide the way for the best decision of number of groups. The major concern is, however, that through the clustering, the relative importance of each feature can be decided. Table 2a and 2b show two samples out of the results based on the cluster boundaries (in fact it is a large table for every feature; here only two small portions are shown).

10-group, 9-group, and etc. refers to clustering the features in 10 groups, 9 groups and so on. A 1-group would mean that all features would be clustered into one group, which is the data itself. The numbers in the table represent the number of times that a feature is included in one of these clustered groups.

On the basis of this table, the following conclusions can be drawn:

- When the numbers of clustering are less than 8, Group 1 in these groups is invariable. This is indicated by the thick line between group 1 and group 2 in table 2a: the feature number in Group 1 remains the same regardless of what the number of clustering should be when the number of clustering is less than 8.
- When the numbers of clustering are greater than 7, there is only one feature in one group (group 2). Group 2 in the grey area indicates this.
- When the numbers of clustering are between 4 and 9, the last group of these clusters is invariable. With reference to the block gray areas, the feature numbers are the same.
- When the number of clustering changes from 3 to 4, the third group of the 3-group is split into two groups, When the clustering number changes from 4 to 5, the second group is split into two groups. When the number of clustering changes from 5 to 6, the fourth group is split into two groups. When the number of clustering changes from 6 to 7, the second group is split up. Figure 6 shows such a process.

V. USE OF SILHOUETTE PLOTS TO SELECT CLUSTERING NUMBERS

An additional tool to distinct between the relative importances of features is the application of so-called silhouette plots. These silhouette plots provide an answer to the best choice for cluster numbers. To decide the number of classes that should be applied in clustering, the following criteria are applied:

1). The features in each group should not be too few when there is a certain clustering. If there are no features or only very few features in a certain group, the class of importance in this group would be not representative, and thus not lead to an appropriate classification.

2). The number of clustered groups should not be so few that the clustering loses any practical meaning. For example, the 2-group clustering is not feasible since the semantic meaning of 2-group would result in "absolutely accept" and "absolutely reject" only, which is too simple a differentiation in practice.

3). All features listed in the tables / questionnaires are currently included in the topographic maps, which contains as many features as possible. It would however be an opportunity to remove some features from these fixed framework data. Splitting up the lower numbered groups could do this.

Based on the first two criteria clustering numbers can be derived ranging from 7-group to 3-group, in line with conclu-

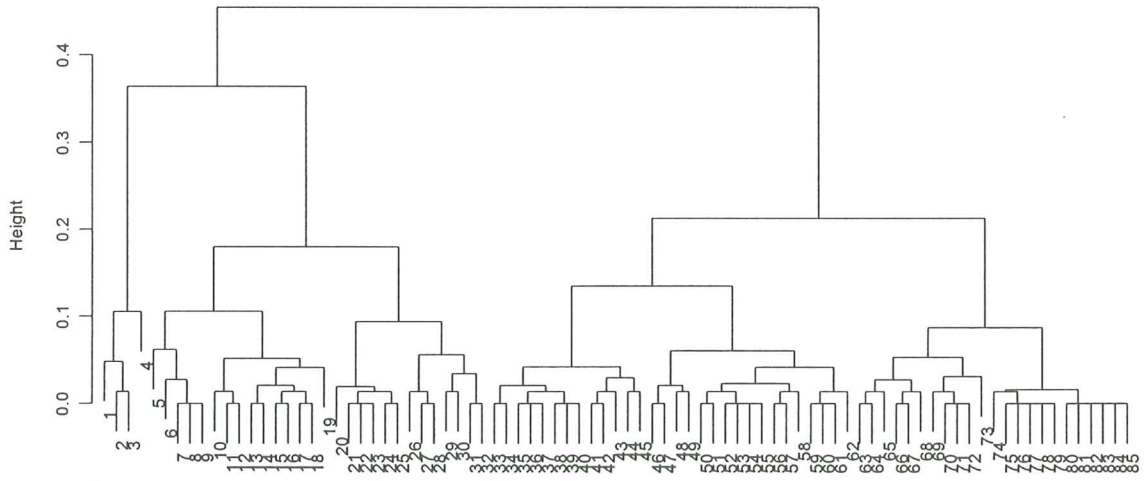

Figure 5. Clustering process

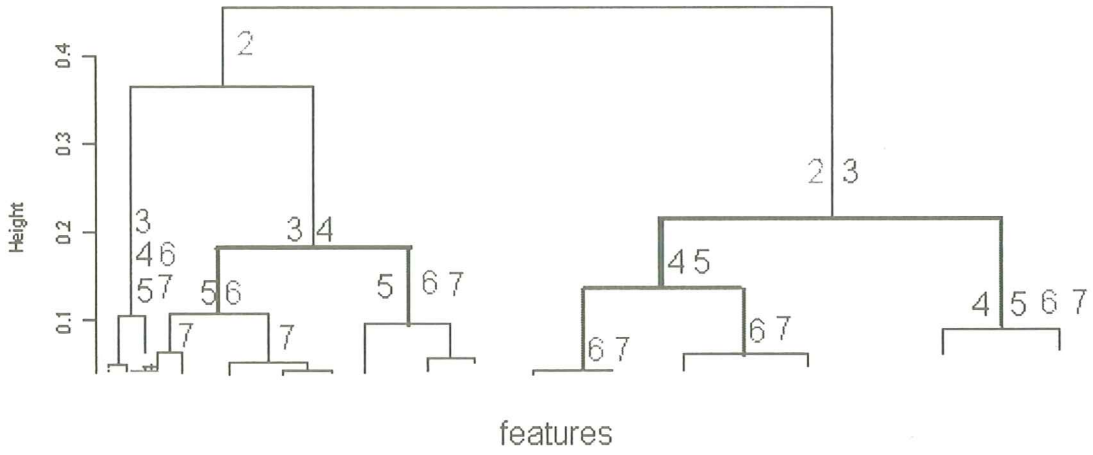

Figure 6. An example of cluster merge

sions 1) and 2) of the paragraph above. Furthermore, looking at conclusion 3) of the paragraph above, the two reasonable numbers of clustering are 5-group and 7-group according to criterion 3. This is because when the number of clustering changes from 4 to 5, the second group of 4-group clustering is split into two groups, and when the clustering number changes from 6 to 7, the second group of 6-group clustering is split up. Therefore if there is a need for a classification based on the assumptions, 5-group clustering and 7-group clustering are the best choices.

The selection of the best clustering numbers can also be verified through other clustering methods. A partitioning algorithm can carry out the clustering and also yield a "quality index," which allows the user to select the "best" value of *k* afterwards. Partitioning around mediods was adopted to verify the reasonability of the number of clustering. The clustering number can be evaluated by means of the *silhouette plot* (Rousseeuw, 1987). For each object *i*, the silhouette value *s(i)* is computed and then represented in the plot as a bar of length *s(i)*. In order to define *s(i)*, A denotes the cluster to which

Table 2. clustering results of 2-10 clustering numbers

Feature	10-group	9-group	8-group	7- group	6- group	5- group	4- group	3- group	2- group
Destroyed settlement	1	1	1	1	1	1	1	1	1
Fence	1	1	1	1	1	1	1	1	1
Settlement type	1	1	1	1	1	1	1	1	1
Administrative town boundary	2	2	2	1	1	1	1	1	1
Second road/railway (can be deleted in density area)	9	9	8	7	6	5	4	3	2
Bridge	9	9	8	7	6	5	4	3	2
High-volt electricity line (110KV)	9	9	8	7	6	5	4	3	2
Geographic name	9	9	8	7	6	5	4	3	2
Triangle points	9	9	8	7	6	5	4	3	2
GPS points	9	9	8	7	6	5	4	3	2
Railway station	9	9	8	7	6	5	4	3	2
Airport	9	9	8	7	6	5	4	3	2
Second road/railway	10	9	8	7	6	5	4	3	2
Main river, canal	10	9	8	7	6	5	4	3	2
Minor river, canal	10	9	8	7	6	5	4	3	2
River, canal (with classes)	10	9	8	7	6	5	4	3	2
Lake and reservoir	10	9	8	7	6	5	4	3	2
Main road/railway	10	9	8	7	6	5	4	3	2
Current boundary (county based)	10	9	8	7	6	5	4	3	2
Administrative name	10	9	8	7	6	5	4	3	2
Land cover and land use	10	9	8	7	6	5	4	3	2
Integrated settlement	10	9	8	7	6	5	4	3	2
Dense tent, cave, ...	10	9	8	7	6	5	4	3	2
Contour and height points	10	9	8	7	6	5	4	3	2

object i belong, and the calculation proceeds as

$a(i)$ = average dissimilarity of i to all other objects of A

Now consider any cluster C different from A and define

$d(i, C)$ = average dissimilarity of i to all objects of C

After computing $d(i, C)$ for all clusters C not equal to A, we take the smallest of those:

$b(i)=\min_{C \neq A} d(i, C)$

The cluster B which attains this minimum, namely $d(i, B)=b(i)$, is called the *neighbor* of object i. This is the second-best cluster for object i.

The value $s(i)$ can now be defined:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

It can be found that $s(i)$ always lies between -1 and 1. The value $s(i)$ may be interpreted as follows:

1). $s(i) \approx 1 \Rightarrow$ object i is well classified
2). $s(i) \approx 0 \Rightarrow$ object i lies between two clusters
3). $s(i) \approx -1 \Rightarrow$ object i is badly classified

The silhouette of a cluster is a plot of the $s(i)$, ranked in decreasing order, of all its objects i. The entire silhouette plot shows the silhouettes of all clusters next to each other, so the quality of the clusters can be compared. The *overall average silhouette width* of the silhouette plot is the average of the $s(i)$ over all objects i in the data set.

Figure 7 shows the obtained silhouette plots with clustering numbers 4, 5, 6, 7, 8, 9 respectively. The average silhouette widths yield 0.57, 0.6, 0.61, 0.61, 0.61,0.59. In general clustering numbers 6, 7, 8 seem appropriate choices. since the average silhouette widths (all are 0.61) are the greater than all others. The best clustering number is 7 according to Figure 2 since all $s(i)s$ are great than zero.

If a 5-group clustering is chosen, it would imply 5 classes with a classification ranging from those features that respondents would absolutely consider necessary as framework data features to those features which respondents do not consider necessary to include as framework data, corresponding to number 5,4,3,2,1 of 5-group clustering in table 2a/b. Similarly, for a 7-group clustering, the classification could be "absolutely necessary to include to not necessary/nontrivial", corresponding to number 7,6,5,4,3,2,1 of 7-group clustering.

Looking at the actual features represented by these cluster, it would mean for the framework in China that:

- The features "destroyed settlement" and "fence" can be removed from the framework data, based on the current 1:50,000 maps. The existing feature "administrative boundary" on that 1:50,000 map apparently provides sufficient information as framework data.
- When the clusters numbers are between 4 and 9, the last group of these clusters is invariable. In other words, this group is quite stable when the clustering numbers change. They can be classified as very important features.
- The features that are clustered as normal should start from Mongolia tent. The cut-off point of features to be included or not included can thus be considered the features "Mongolia tent" and "bus station."

VI. DERIVATION OF ALTERNATIVES

Although a cut-off point of features to be included in framework datasets was defined in the previous paragraph, one could also derive various possible alternatives based on the cluster analysis. Such alternatives could be based on the classifications as derived by the clusters. Using these classifications 3 alternatives were evaluated:

- *Alternative 1*: The framework data containing a minority of clusters of features only, meaning that only those clusters are considered which can be classified as considerable acceptance by respondents.
- *Alternative 2*: The framework data containing the majority of features and disregarding a minority of mostly rejected features.
- *Alternative 3*: The framework data removing the completely rejected features only.

These alternatives are obtained through the above clustering. Obviously other alternatives can be generated based on the number of clusters that one would like to consider. The results are given in the columns alternatives (Alt 1,2,3) of Table 1.

Alternative 1 contains the least features: It contains only 24 features among the total features. Yet these features cover all 10 classes. Alternative 1 provides features that reflect the main terrain features in general. In other words, few kinds of features are basically able to construct the skeleton of the framework data for China. This alternative is almost identical to the current smaller-scales topographic maps. Therefore, it can be applied for general uses for application areas such as macro analysis and planning, education, general navigation, and geographic maps for other thematic features. Alternative 1 would be appropriate for a situation where large-scale investments will not be possible for the generation and maintenance of framework data sets.

The choice for larger number of clusters would lead to Alternative 2. This Alternative 2 is similar to the standard framework data in literature since covers almost all generic features that most users use in their fields. Compared with Alternative 1, there is not much difference in classes 1, 3, 5, 6, and 10. The updated classes such as class 2, 4, 7, 8, 9 in A2 include more features than those in Alternative 1. The change makes the features in the relatively complete system.

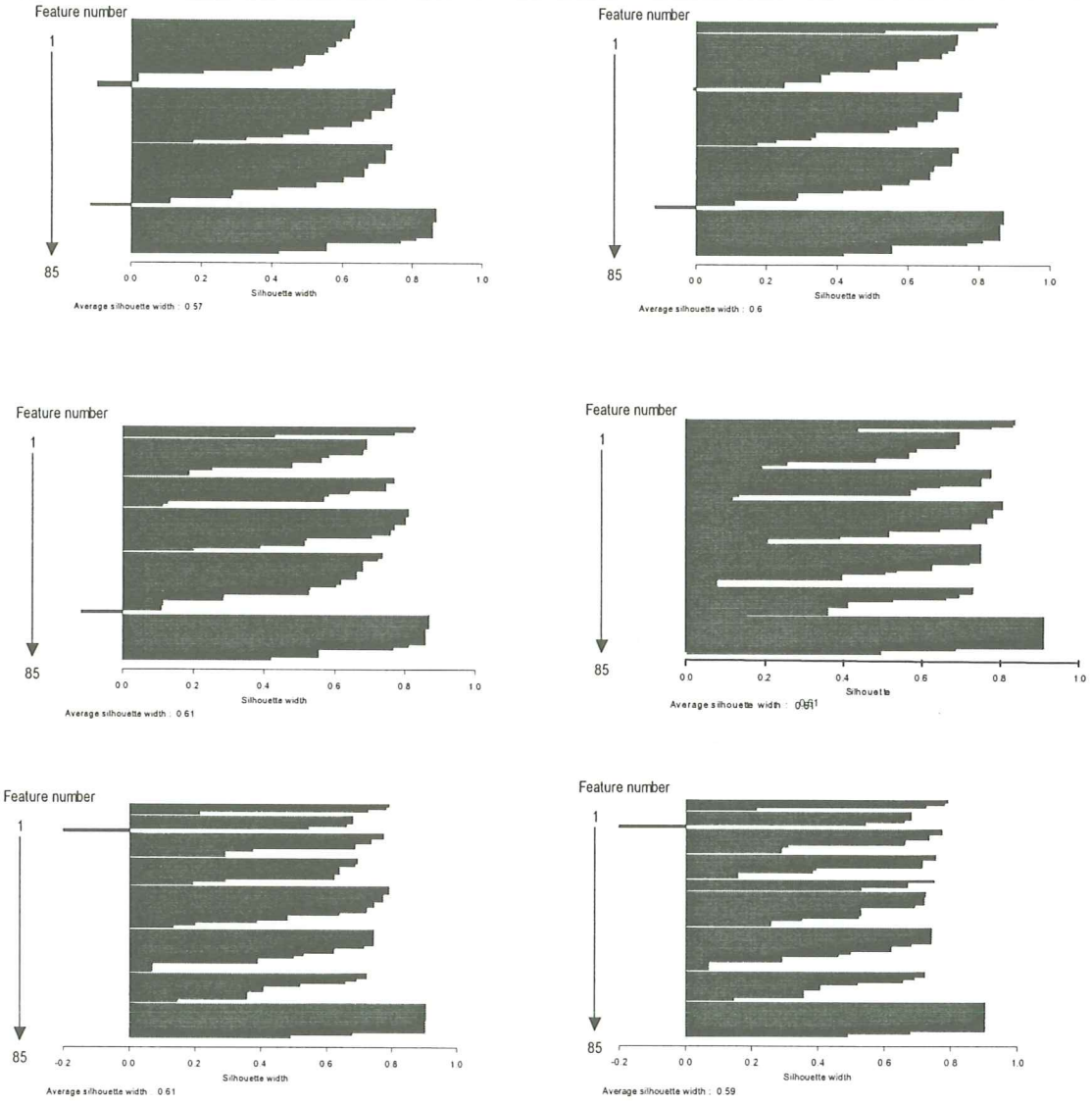

Figure 7. Silhouette plots of cluster numbers 4, 5, 6, 7, 8, and 9

The Alternative 3 is the complete coverage of the framework data (except for the feature destroyed settlements). The difference between Alternative 2 and Alternative 3 lies in that the latter covers more features. However, it would cost much more in terms of data collection, data distribution and maintenance. As a result, this alternative is perhaps less feasible, and it would therefore not be a good choice for the definition of the framework data set.

Based on the above analysis, two alternatives can be drafty selected as the framework data of China: Alternative 1 and Alternative 2, with a slight preference for the latter According to the above analysis, Alternative 2 is a good selection for the framework data of China at scale 1:50,000 all over the country; and Alternative 1 can be adopted for public access because of data security.

IIV. CONCLUDING REMARKS - WHERE TO CONTINUE THE STATISTICAL ANALYSIS

The main reason to apply statistical analysis of user requirements was to derive to a methodological answer of which data are still core, fundamental data to how many users. The analysis and examples showed that the cluster analysis allows for an answer to this, resulting in some features of the current maps to be excluded from core data sets. Based on this result, the various institutions could make individual arrangements of who should continue to produce what, and which information needs to be exchanged or exchangeable for the majority of users.

The analysis should not end with these results. The statistical clustering was largely focused on the analysis of the feature importance according to the user's requirement, but did not specifically take into account the regional or sectoral differ-

ences. It is noted that there are still differences between different organizations at different administrative levels. In order to highlight these differences, a more detailed survey and subsequent statistical analysis would still be necessary. Currently, however, it is considered reasonable to start at a national level, since in China all framework data at 1:50,000 are administrated and maintained in SBSM. The framework data at 1:10,000 are administrated and maintained at the provincial level, and the features could be different between different areas.

Finally, this article focused on the analysis of the content layers only, and did not include the implications on spatial quality, consistency or completeness. If certain features are omitted from the core, fundamental data sets, it may have implications for the adjacent features if these are topologically related in the various systems. The analysis should therefore not stop with the list of features, but would need to consider these consistency problems one-by-one.

ACKNOWLEDGEMENTS

Thanks to Liu Yunfeng, Shanxi Bureau of Surveying & Mapping, who helped collect data.

REFERENCES

[1] Groot, R., 1998, Entering the 21st Century Strategies for National Surveying and Mapping, *Report of the Executive Seminar*, ITC, the Netherlands.

[2] Groot, R. and Mclanghlin, J., 2000, *Geospatial Data Infrastructure Concepts, Cases and Good Practice*. Oxford: Oxford University Press.

[3] Kaufman, L. and Rousseeuw, P. J., 1990, *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.

[4] Masser, I., 1998, *An International Overview of Geospatial Information Infrastructure: Lessons to Be Learnt for the NGDF*, NGDF, UK. URL http://www.ngdf.org.uk/Pubdocs/General/masser.html, last accessed in 2001.

[5] McLaughlin, J. D., 1991, Towards national spatial data infrastructure, In *Proceedings of the 1991 Canadian conference on GIS*, Ottawa, Canada, Canadian Institute of Geomatics, Ottawa, Canada, March, pp.1-5

[6] Onsrud, H. J., 2001, *Survey of National and Regional Spatial Data Infrastructure Activities Around the Globe,* University of Maine, USA. http://www.spatial.maine.edu/~onsrud/GSDI.htm. Accessed Last update 14 October 2001.

[7] Rhind, D., 1992, the Information of Infrastructure of GIS, In *Proceedings of the 5th International Symposium on Spatial Data Handling*, Vol.1, pp1-19. Humanities and Social Sciences Computing Lab, University of South Carolina, Columbia, S.C., U.S.A.

[8] Rousseeuw, P. J., 1987, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53-65.

[9] Snedecor, G. W. and Cochran, W. G., 1980, *Statistical Methods*, Seventh edition. Iowa State University Press.

[10] Tacq, J., 1997, *Multivariate Analysis Techniques in Social Science Research: From Problem to Analysis*. London, Sage.

[11] Wulan, 2002, Methodology for selection of framework data, Case study for NSDI in China, *MSc thesis Geoinformatics*, March 2002.

[12] FGDC, 2002, Framework Introduction and Guide, FGDC, USA, URL http://www.fgdc.gov/framework/faqframe.html, last accessed in 2002.

[13] GSDI Cookbook, Spatial Data Infrastructure Cookbook, 2001, http://gsdi.gov , last accessed in 2001.