# DATA MINING APPLIED IN LAND USE CONTROL IN CITY-COUNTRY COMBINATIVE AREA

**Shuliang Wang, Xinzhou Wang**
National Laboratory for Information Engineering in Surveying Mapping and Remote Sensing
Wuhan Technical University of Surveying and Mapping, Wuhan City, Hubei Province, P.R.China 430079
sl wang@rcgis.wtusm.edu.cn xzwang@wtusm.edu.cn

Work Group II/3

**KEY WORDS:** Data Mining, Data Cleaning, Vectorizing Character-string and Matching Number, Rough Set, Land Use Control, Photogrammetry, City-country Combinative Area,

## ABSTRCT

This paper studies the concepts of city-country combinative area (CCCA), and how to apply data cleaning, and data mining in land use control in CCCA. A new data cleaning method named as Vectorizing Character-string and Matching Number (VCMN) is advanced, which can integrates multi-origin data. It is suggested that the control points be acquired with GPS-supported aerial triangulation, CCCA images be dealt with digital photogrammetry system before image databases come into being. Then all related land databases can be rebuilt up a data warehouse after data cleaning. In the network, online data mining is carried out in the light of constraint conditions. Finally, some Wuhan City CCCA knowledge is discovered with rough set.

## 1 INTRODUCTION

With Chinese development, many cities become bigger, which causes lots of land problems in CCCA. Here, urban-construction land and plantation-protection compete with each other acutely. Though the valuable land in CCCA is the basis of peasants' life and citizens' vegetables, many bad phenomena on land use are happening, such as occupying land in disorder, land misuse, lawless land exchange and wasting land. So revised "Land Management Laws of P.R.China" stresses on land use control, strengthens plantation-protection and controls urban-construction land, in order to ensure the dynamic balance of plantation gross. However, land use is complex, whose change and process are mostly non-linear with layer upon layer acute conflicts. Hence, land use control is a synthetic engineering, including economic, lawful, technical and administrative measures. It is violently advised to monitor land utilization and alteration, carry out control rules in force, perfect land market, strengthen macroscopical accommodation and control, and deepen the system reform. And it is obvious that all measures are based on land information and information processing technology. Besides making full use of those methods, technical measures should be paid more attention to, for they provide the essential information. Take the following for example. Land planning ahead can give the methods on how to control land use in CCCA, and lead to reasonable land use. Dynamic land monitoring may acquire the newest land information instead of manpower, and avoid some land problems in time. The visual land crisis map will forbid man to cherish land consciously. Therefore, after studying the concepts of CCCA, this paper discusses how to acquire and utilize land information with photogrammetry, GPS-supported automatic aerial triangulation, data cleaning ----VCMN, data mining, rough set etc.

## 2 LAND USE CONTROL IN CITY-COUNTRY COMBINATIVE AREA (CCCA)

City-country combinative area (CCCA) is a bridge between city and country, a front position of country urbanizing and modern agriculture. As its development of society and economy is peculiar and active, CCCA owns both better transportations, infrastructures as a city zone and fresher natural environments like a country zone. However, there are a lot of acute land use problems in CCCA. The conflicts between construction and agriculture are violent. Two different administrative systems in CCCA (city and country) make it weak to plan and administer land, which leads to the imbalance of land structure. Moreover, excessively exploiting and occupying lands destroy the original balance of the nature, which lessens its function of absorbing city wastes. It is the main problems in CCCA that the benefits of land use are often lower. For example, in the core zone of Guangzhou City, the building density is more than 52.2%， but in CCCA, it is only 20% or so. Moreover, constructions generally occupy high quality land. Therefore, the people have to control land use in CCCA in order to ensure the sustainable land use.

Land use control in CCCA is that the state makes and releases land planning, carve up land use subarea, in order to assure that land resources are utilized reasonably, which keeps the economy, society and environment developing harmoniously. Land should be planned ahead. It is suggested to check and activate existed land, limit land increment in order to avoid city expanding excessively. Then, the land utilization limitation should be worked out. For example, land use alteration regular includes permission, confined permission, impermissibility. Unfortunately, it is difficult for only manpower to control land use duly, for land information change is mostly non-linear, with layer upon layer acute conflicts. Photogrammetry can be used to acquire land data of space, time and attribute in CCCA. After they are put into the computer, those data can be shared on the network in the scope of laws. With land data cumulating, kinds of databases come into being, which can be built into data warehouse after data cleaning. There is unknown knowledge hidden in data, which can be used to monitor, control and perfect land use, The knowledge from online analytical mining (OLAM) will not only benefit land use control in CCCA, but also accelerate man to protect plantation forwardly and utilize land economically.

## 3 PHOTOGRAMMETRY AND LAND USE CONTROL IN CCCA

Photogrammetry acquires CCCA images without touching objects, which makes land geometry and attribute information extracted from the images more reliable and impersonal. As we know, land is not able to reproduce and reverse use. However, land resources in CCCA are facing serious crisis, and the changes of land information are intricate. Sometimes there are man-made errors in land statistic for some reasons. At the same time, not all ground control points in CCCA could see each other because of artificial and natural objects on the ground, especially in the area with intensive high buildings. Fortunately, GPS-supported automatic aerial triangulation has come into being [Deren Li,1997]. It acquires aerial control points with GPS carried in a plane instead of ground control points or with less ones. Because the GPS points are of high grade, and whose error is evenly distributed, we can take advantage and high relative precision of position and pass points, which can match the requirement of precision and density.

DPS including both analytical photogrammetry and digital image manipulation functions, can be used to extract land information such as feature and topographic essentials. Here is the procedure of its application in land use control in CCCA: Put the basic parameters into the system, and calculate the orientation parameters including interior orientation, relative orientation and absolute orientation; Aerial triangulation; Array stereo-images in the direction of epipolar line; Images matching; Post-manipulating and editing; Make landscape map on land dynamic use in CCCA ; Based on digital

image, get buildings information on the ground with computer auxiliary survey. After all operations are finished, many products on land use control in CCCA can be automatically obtained such as imagine map (Figure 3), cadastral map, section map etc. All of the information should be put and stored into land information system (LIS). As the spatial information in land management is horizontal, the products are easier to be stored and utilized.

With the software help of DDKIN, VirtuoZo and WuCAPS$_{GPS}$, GPS-supported automatic aerial triangulation with high quality takes less time, manpower and expense than traditional aerial triangulation. Moreover, it is easy to refresh land information. However, if we can make full use of three-dimensional information to monitor dynamic land, the spatial forecasting analysis of CCCA land will truthfully and deeply show the people land crises with the technologies of virtual realism, three-dimensional landscape and so on. That will oblige them to work out land planning ahead, use land economically and protect plantation consciously. Hence, the method is welcome in land use control in CCCA, especially for land monitoring in time.

## 4 DATA MINING

Land use control in CCCA is a systematic engineering. It utilizes not only the information from photogrammetry, bot also lawful, social and economical information from land general survey and land detailed survey. Obviously, this is the integration of heterogenous data. When all of the databases meet with each other in LIS, data cleaning and mining becomes necessary.

### 4.1Data Cleaning with Vectorizing Character-string and Matching Number (VCMN)

In order to prepare for data mining well, data cleaning should be done first. If the data integration improved by 4% in a big corporation maybe leads to more than one million US$ profits [Innovative Systems, Inc.,1997]. Land information in CCCA used for land use control is multi-origin, there are kinds of problems in original data such as incomplete data, inaccurate data, ambiguity data, conflicting data, repetitive data. If they are permitted entering CCCA data warehouse, the result of data mining will be incorrect.

The matching and amalgamation of heterogenous data is an important content of data cleaning. It may be abstracted as the clustering analysis. And this paper put forward a method named as Vectorizing Character-string and Matching Number (VCMN) to satisfy the integrating demands of multi-origin data. Suppose that attribute 'A' of the same land 'L' has n($n \geqslant 1$) values,$\{V_1,V_2,...,V_n\}$, which come from different data origins,$\{ O_1,O_2,...,O_n \}$. The credibility degree of $V_i$ is defined as R(L, A, $V_i$), equation (1).

$$R(L,A,V_i)=[W(Q_k)\cdot R_0(L,A,V_k)]\cdot\frac{1}{n}\sum_{j=1}^{n}S(V_i,V_j) \quad (1) \qquad S(\vec{V}_i,\vec{V}_j)=\frac{\sum_{m=1}^{p}v_{im}\cdot v_{jm}}{\sqrt{\sum_{m=1}^{p}v_{im}^2}\cdot\sqrt{\sum_{m=1}^{p}v_{jm}^2}} \quad (2) \qquad S(V_i,V_j)=\exp\left[-\left|\frac{V_i-V_j}{\max(V_1,V_2,\ldots,V_n)}\right|\right] \quad (3)$$

Where, $R_0(L,A,V_j)\in[0,1](j=1,2,...n)$ is the original value of R(L,A,$V_i$)， which mainly aims at data cleaning for many times. It can be given as one for the first times. $W(O_j)\in(0,1](k=1,2,...,n)$ is the creditable weight of data origin $O_j$, which can be given by users. $W(O_k)\cdot R_0(L,A,V_k)= \max\{W(O_j)\cdot R_0(L,A,V_j)\ |V_j=V_i, j=1,2,...,n\}$. $S(V_i,V_j)\in[0,1]$ is the similar degree between $V_i$ and $V_j$. Equation (2) is for character string values, and equation (3) is for numeric values. In equation (2), $V_i$, $V_j$ are the character strings. And $\vec{V}_i\vec{V}_j$ are their vector denotation. $\vec{V}_i$ =($v_{i1}$,$v_{i2}$, $v_{i3}$, ..., $v_{iP}$) and $\vec{V}_j$ =( $v_{j1}$,$v_{j2}$,

$v_{j3},...,v_{jP}$). The vector dimension points to a word, whose value is the times of the word appearing in land 'L'. Especially, if the value of $V_i$ is empty then $M(V_i,V_j)=0$, $R(L,A,V_i)=0$; and if $V_i=V_j$, then $R(L,A,V_i)=R(L,A,V_j)=1$.These satisfy the function definition. Finally, $V_z$ with the biggest $R(L,A,,V_z)$ value of $V_1,V_2,\ldots,V_i\ldots,V_n$ is selected to be the value of attribute 'A' of land 'L', while those empty values or repetitive values ($V_1,V_2,\ldots,V_i\ldots,V_n$ ; $V_i \neq V_z$) are eliminated.

Take it for example. There is a parcel of land (Table 1) in Jiang'an zone CCCA of Wuhan city. On the basis of VCMN,

| Data origin | Owner name | Land area ($m^2$) | Reliable weight |
|---|---|---|---|
| Photogrammetry | Wuhan A B C D E Corporation | 330.24 | 0.4 |
| Land survey | Wuhan A B C D F G Corporation | 332.00 | 0.5 |
| Land statistic | Wuhan A B C F G Corporation | 284.67 | 0.2 |
| RESULTS | Wuhan A B C D F G Corporation | 332.00 | |

Table 1. Data cleaning on a parcel of land in Jiang'an zone CCCA of Wuhan city

let us do data cleaning on its land use change between 1988 and 1995. Owner name attribute is character string. First , the measure-space dimension should be extracted from the attribute 'owner name' origins, eg. (A, B, C, D, E, F, G). Then $V_1$, $V_2$ "Wuhan A B C D E Corporation" and "Wuhan A B C F G Corporation", are changed into $V_1$="A B C D E", $V_2$="A B C F G". $\vec{V}_1$ =(1,1,1,1,1,0,0) and $\vec{V}_2$ =(1,1,1,1,0,1,1). Hence, $S(\vec{V}_1,\vec{V}_2)$ =0.730 according to equation (2). At the same time, land area is numeric value, so $V_1$ and $V_2$, $S(V_1,V_2)$=0.995 due to equation (3). The same with other attributes. At last, we get the results as table 1, Obviously, there are man-made errors in land statistic (Perhaps it is for the owner wants to turn in less land tax.), while the information of land survey is the most reliable.

After land data cleaning, different databases from different data origins have become clean, complete and re-engineering. They can be built up data warehouse on land use control in CCCA for data mining

### 4.2 Data Mining Based on Rough Set

If the data warehouse is able to detect the data world successfully and effectively, it is the focus for data mining to discover useful information. Data mining is to find knowledge in data warehouse, which is concealed, potential, effective, useful, and understood. There are many kinds of data mining methods. Here, minimal decision based on rough set and data auditing is studied to discover land use control knowledge in CCCA.

**4.2.1 Principles of Rough Set** A database is a special case of a knowledge representation system. Let S={U,C,D,V,f} be a knowledge representation system, where U is a nonempty set of objects (i.e., U= $\{u_1, u_2,\ldots, u_n\}$), C is a nonempty set of conditional attributes, and D is a nonempty set of decision attributes. A($A=C\cup D$) is the set of all attributes and $C\cap D=\phi$ .Let $V=\cup\{ V_a | a\in A\}$, where $V_a$ is a finite attribute domain and the elements of Va are called values of attribute a. f is an information function such that $f(u_i,a)\in V_a$, for every $a\in A$ and $u_i\in U$. Every object that belongs to U is associated with a set of values corresponding to the condition attributes C and decision attributes D.

Suppose B is a nonempty subset of A, $u_i$, $u_j$ are members of U, and R is an equivalence relation over U($R=U\times U$). Define a binary relation, called an indiscernibility relation as $IND(B)=\{(u_i, u_j)\in R \mid u_i, u_j\in U, \forall a\in B, f(u_i,a)=f(u_j,a)\}$. It is believed that $u_i$ and $u_j$ are indiscernible by a set of condition attributes B in a KRS iff $\forall a\in B$, $f(u_i,a)=f(u_j,a)$. The indiscernibility relation partitions U into equivalence classes. Equivalence classes of the relation R are called elementary sets in an approximation space APR=(U,R). For any object $u_i\in U$, the equivalence classes of the relation R containing $u_i$ are denoted $[u_i]_R$. If $X\subset U$, then the lower approximation, the upper approximation, are respectively

$\underline{APR}(X)=\{u_i \in U \mid [u_i]_R \subseteq (X\}$, $\overline{APR}(X)=\{u_i \in U \mid [u_i]_R \cap X \neq \phi\}$, X is rough with respect to IND(B) iif $\underline{APR} \neq \overline{APR}$. And a subset (eg.X⊂U) defined with the lower approximation and upper approximation is called Rough Set. Table 2 shows an example of a knowledge representation system, which is land use change in CCCA of Wuhan city between

| U | ZO | TO | PL | VE | FO | GR | BI | TR | WA | VI |
|---|---|---|---|---|---|---|---|---|---|---|
| $U_1$ | JA | -1 | -1 | 0 | -1 | 0 | 3 | 1 | -2 | -1 |
| $U_2$ | JH | -2 | -2 | 0 | -1 | 0 | 1 | 1 | -1 | 0 |
| $U_3$ | QK | -1 | -1 | 0 | 0 | 0 | 1 | 0 | -1 | -1 |
| $U_4$ | HY | -1 | -1 | 0 | -1 | 0 | 1 | 0 | -1 | -1 |
| $U_5$ | HS | -1 | -3 | -1 | -1 | 0 | 3 | 1 | -2 | -1 |
| $U_6$ | WE | -1 | -3 | -1 | -1 | 0 | 3 | 1 | -2 | -1 |
| $U_7$ | CD | -2 | -3 | -1 | -1 | 0 | 3 | 0 | -2 | -1 |
| $U_8$ | JX | -1 | -2 | -1 | -1 | -1 | 2 | 1 | -1 | -1 |

Table 2. Knowledge Representation System

| U | ZO | PL | BI | WA |
|---|---|---|---|---|
| $U_1$ | JA | -1 | 3 | -2 |
| $U_2$ | JH | -2 | 1 | -1 |
| $U_3$ | QK | -1 | 1 | -1 |
| $U_4$ | HY | -1 | 1 | -1 |
| $U_5$ | HS | -3 | 3 | -2 |
| $U_6$ | WE | -3 | 3 | -2 |
| $U_7$ | CD | -3 | 3 | -2 |
| $U_8$ | JX | -2 | 2 | -1 |

Table 3. Reduction table

| U | ZO | PL | BI | WA |
|---|---|---|---|---|
| $U_1$ | JA | -1 | 3 | -2 |
| $U_2$ | JH | -2 | 1 | -1 |
| $U_3$ $U_4$ | QK HY | -1 | 1 | -1 |
| $U_5$ $U_6$ $U_7$ | HS WE CD | -3 | 3 | -2 |
| $U_8$ | JX | -2 | 2 | -1 |

Table 4. Decision table

1988 and 1995. The original data have been not only cleaned by VCMN, but also transformed with the method of histogram equilibria. Where, ZO is the zones of Wuhan City, and JA, JH, QK, HY, HS, WE, CD, JX are respectively Jiang'an, Jianghan, Qiaokou, Hanyang, Hongshan, West & East Lake, Caidian, Jiangxia. TO, PL, VE, FO, GR, BI, TR, WA, VI are respectively the change areas of total land, plantation, vegetable land, forest land, grass land, transportation land, urban-rural building & industrial-mineral land, water, virgin land. -3,-2,-1, 0, 1, 2, 3 respectively denote big down(-8~-12km$^2$), down(-2~-8km$^2$), small down(0~-2km$^2$), unchanged(0km$^2$), small up(0~+2km$^2$), up(+2~+8km$^2$), big up(+8~+12km$^2$). U= $\{u_1, u_2, \ldots, u_8\}$ Each object is described by a set of condition attributes C={TO, PL, VE, FO, GR, BI, TR, WA, VI}, with attribute values VTO={-1,-2}, VPL={-1,-2,-3}, VVE={0,-1}, VFO={0,-1}, VGR={0,-1}, VBI={1,2,3}, VTR={0,1}, VWA={-1,-2},and VVI={0,-1}.The set of values VZO= {JA, JH, QK, HY, HS, WE, CD, JX} of the decision attribute D represents the set of concept descriptions which are to be learned based on the attribute values of C.

**4.2.2 Elimination of Superfluous Attributes** In the data collection, all the features believed to be useful and relevant are collected into CCCA databases. In a database system, we describe each object by the attribute values of C. Very often it turns out that some of the attributes in C may be redundant in the sense that they do not provide any additional information about the objects. Thus it is necessary to eliminate those superfluous attributes to improve learning efficiency and accuracy.

Suppose B⊂C, then $POS_B(D)=\{\underline{B}(X) \mid X \in IND(D)\}$. An attribute a∈C is superfluous in C with respect to D if $POS_C(D)= POS_{C-\{a\}}(D)$, otherwise a is indispensable in C with respect to D. If an attribute is superfluous in the information system, it should be removed from the information system without changing the dependency relationship of the original system. For example, TO, VE, FO, GR, TR, VI are superfluous attributes in Table 2. Table 3is obtained by removing them from Table2. As we can see, Table 3 is simple but has the same discernibility as Table 2. A rule is a combination of values of some condition attributes such that the set of all objects matching it is contained in the set of objects labeled with the same concept, and such that there is at least one such object. Traditionally, the rule is denoted as an implication:

$$(C_1=V_{i1}) \wedge (C_2=V_{i2}) \wedge \ldots \wedge (C_m=V_{im}) \rightarrow ((D=V_d) \quad (4)$$

Where, $C_1, C_2, \ldots, C_m$ are the condition attributes and D is a decision attribute.

The process by which the maximum number of condition attribute values of a rule are removed without decreasing the classification accuracy of the rule is called Value Reduction, and the resulting rule is called minimal decision rule. Thus, a minimal decision rule is optimal in the sense that no condition could be removed without decreasing the classification accuracy of the rule. The computing of minimal decision rules is of particular importance with respect to data mining, since they represent the most general patterns existing in the data. For example, Table 4 depicts a decision matrix

obtained from the KRS given in Table 4. From Table 5, we can get the following minimal decision rules for the ZO.
(PL=-1) ((BI=3) ((WA=-2) → (ZO=JA); (PL=-2) ((BI=1) ((WA=-1)→ (ZO=JH);
(PL=-1) ((BI=-1) ((WA=-1)→ (ZO=QK,HY); (PL=-3) ((BI=3) ((WA=-2) → (ZO=HS,WE,CD);
(PL=-2) ((BI=2) ((WA=-1)→ (ZO=JX).
The knowledge shows that plantation decreased most in Hongshan zone, West & East Lake zone, Caidian zone. Urban-rural building & industrial-mineral land increased most in Jiang'an zone, Hongshan zone, West & East Lake zone, Caidian zone. Water area decreased most in Hongshan zone, West & East Lake zone, Caidian zone. Urban-rural building & industrial-mineral land occupied lots of land in Wuhan city CCCA, especially to plantation and water area. Obviously, Wuhan city should strengthen land use control and plantation protection in its CCCA, and the strictest measures is advised to be implemented in Jiang'an zone, Hongshan zone, West & East Lake zone.

In the light of Modern City development, city and country should blend each other in function and configuration. The first industry, the second and the third should develop harmoniously, so as to form a modern, dispersed and high-effective city-country amalgamation body. At present, there are two development modes on CCCA. One is the service zone of city, the other is the dispersed group. In Wuhan city, small towns around it have not become a certain size, but the main city zone owns a stronger political, economical and cultural attraction. And its CCCA develops a characteristic l and use mode, city and country fusing with each other (Figure 1), which can be seen from Wuhan City images of satellite remote sensing (RS) (Figure 2). So Wuhan City should develop and improve 9 satellite cities such as Houhu, Qingshan, Baisha and so on, in order to set up a compound ecosystem.

### 4.3 Controlling and Analyzing Land Use Dynamically

In order to do daily land use control well in a part of CCCA, a car-borne land information acquirement and update system (CLIAUS) is suggested, which is the integration system of LIS, GPS and charge-coupled device (CCD) camera are set up in a car (Figure 3). When the car is driven in the monitoring area, the land change data can be acquired. The changed data are all put into land basic databases. In the whole CCCA, we may integrate LIS, GPS, CCD camera, RS and DPS into a real-time land use control system in CCCA (Figure 4). After CCCA images of satellite, aviation and unmeasured camera are dealt with by DPS, the images are set up a database. At the same time, all related land data such

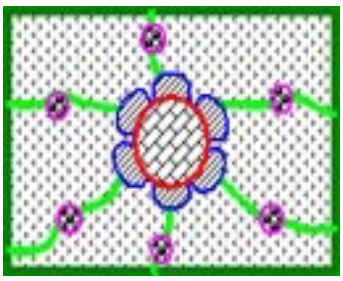

Figure 1. Fusion Mode

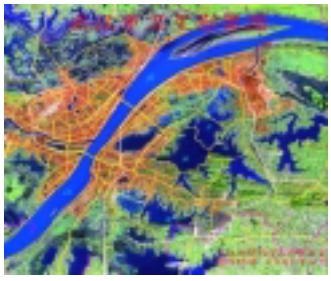

Figure 2. Wuhan RS Image Map

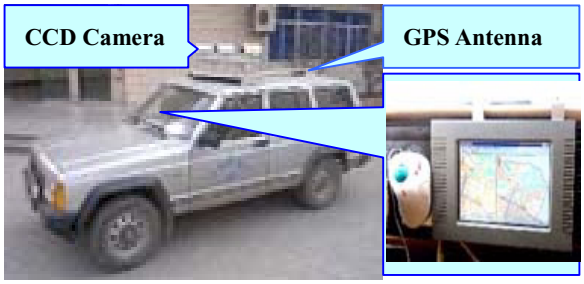


Figure 3. CLIAUS

Where, Main City Zone Service Zone Town Road Ecological Agricultural Land

as images, laws, economy, population etc. are also built up their own databases respectively. Then these databases are not only put into LIS, but also rebuilt up a land data warehouse after data cleaning with VCMN.
In order to make full use of information in the network, OLAM (Online Analytical Mining) is suggested here. Employing the user interfaces, we can let users construct data warehouses, select the desired sets of data, perform constraint-based interactive online analysis processing and mining, visualize and explore the results. OLAM takes advantage of widely available, comprehensive information processing infrastructure. An efficient OLAM should use exiting and real-time infrastructure rather than construct everything from scratch. Furthermore, OLAM provides an

exploratory data analysis environment. It becomes possible to mine different subsets of data and at different levels of abstraction by drilling, pivoting, filtering, slicing, and dicing a multidimensional database and the intermediate data mining results. And it facilitates online, interactive selection of data mining functions and interestingness thresholds. Performing these functions interactively and viewing the results with data/knowledge visualization tools will greatly enhance the power and flexibility of exploratory data mining. With mining constraint, the land use control knowledge can be discovered with OLAM. At the same time, the knowledge should also set up knowledge database in order to make full use of them when data mining. With the help of LIS, the knowledge can represent as imagine map (Figure 2), tridimensional map, virtual forecasting map, cadastral map, land section map etc.

## 5 CONCLUSION

As a new data cleaning method, VCMN can do word analysis and numeric matching together, which will do good to the amalgamation of heterogenous multi-data. Based on rough set, the knowledge from online data mining and social information do greater help for land management in Digital Earth era[Shuliang Wang, 1999]. Because land becomes less and less, it is a continuous project for the people to study how to control land use in CCCA dynamically and intelligently.

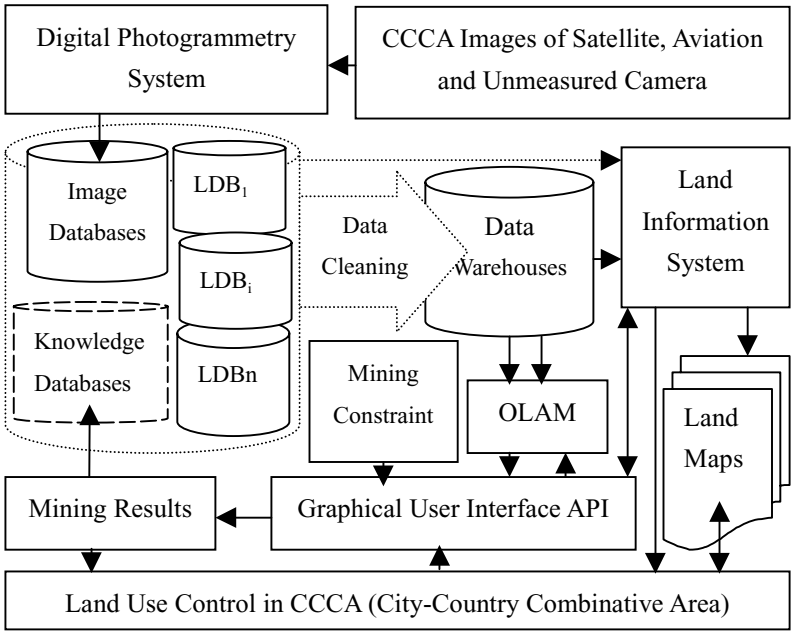


Figure 5. Land Use Control in City-Country Combinative Area

## ACKNOWLEDGEMENTS

This project is supported by both National Natural Science Foundation of China (49874002) and Ph.D. Foundation of Chinese National Ministry of Education (98049801).

## REFERENCES

Deren Li etc, GPS-Supported Automatic Aerial Triangulation. Proc.of Chinese cof. on Remote Sensing, Qingdao, China. 1997.72～79

Kaichang Di, the Theory and Methods of Spatial Data Mining and Knowledge Discovery. Thesis. WTUSM, 1999

Innovative Systems,Inc. "The Magic 4%, The Impact of Data Integrity on Your Data Warehouse", white paper 1997.

Shuliang Wang. Cadastral Management in the context of Digital Earth.Towards Digital Earth--Proceeding of the International Symposium on Digital Earth.Beijing, P.R.China.1999 (12): 250-254