Guobiao Code Extensions:
A Method for Geocoding Sub-county Chinese Data

Merrick Lex Berman
Diamond Bay Research
http://www.dbr.nu

1. What are Guobiao Codes and why should we use them?

The Guobiao (GB) Codes, issued by the Chinese "State Bureau of Technological Supervision"[1] in Beijing, are unique identification numbers assigned to the administrative divisions in the People's Republic of China (PRC). The GB Codes are arranged hierarchically to distinguish three major levels of administrative regions:

the ***provincial*** level, ***ADM1*** (including provinces, autonomous regions, and municipalities under direct central government control);

the ***prefectural*** level, ***ADM2*** (including prefectures, autonomous prefectures, leagues, and municipalities under direct control of the provincial government);

the ***county*** level, ***ADM3*** (including counties, autonomous counties, districts, banners, and municipalities under the direct control of the prefectural government). At any of these three levels, the areas being assigned GB codes include all of the land area claimed to be part of the territory of the PRC, and therefore the GB Codes are an exhaustive indirect spatial referencing system for geocoding this land area.

GB Codes are useful for their complete coverage of China's land area, their heirarchical structure, and the fact that they are perfect for use as primary keys in databases, allowing for cross-referencing and merging data that pertains to any particular administrative division. Being exhaustive, GB Codes can be seen as "containers" for manipulating data related to the regions they identify. Being hierachical, they allow for easy searching at the desired administrative level. And as keys for merging and matching records in unrelated databases, GB Codes have proven to be the best available standard for indirect spatial referencing of China data.

the same Postal Codes are used for several adjacent towns, districts, *banshiqu*, or other local units without being contained within a defined borderline. This ambiguity argues against the adoption of Postal Codes, in their present form, as the optimal indirect geospatial references for Chinese data. On the other hand, thanks to the (more or less) clearly defined boundaries of all the regions identified by GB Codes, and their utility in matching up with statistical data aggregated at the provincial, prefectural and county levels, it makes sense to choose GB Codes for geocoding.

GB Codebooks have been issued in 1980, 1982, 1984, 1986, 1988, 1991, 1995, and 1999. The first six editions are referred to as: GB 2260-80, GB 2260-82, etc. Beginning in 1995, the document number format was changed to GB/T 2260-1995. Although the current GB/T 2260-1999 reflects the latest administrative changes in the P.R.C., it has been recommended that ECAI standardize on the GB 2260-91, current to 01/01/1992. The reason for this is to avoid sorting problems associated with the reclassification of many counties as cities or municipal districts, (and their associated name changes), beginning with the 1995 edition. Also, the 1991 edition was the last national GB Code list used by Prof. William Lavely in the compilation of the CITAS Project datasets.[2]

2. Limitations of GB Codes for the purposes of geocoding

One important feature of GB Codes is that they are not static, they usually are changed whenever an administrative unit changes in status, name, or hierarchical structure. Boundary changes, on the other hand, might occur without corresponding changes in the GB codes; for example, when part of a county is placed under the jurisdiction of another, but both keep their original GB Codes. Since 1980, there have been eight official publications of GB Code Tables, each one containing a certain number of such changes. As long as a look-up table is maintained, we can keep track of these historical relationships. Nonetheless, it is essential for us to date the records in our datasets as accurately as possible, so that we can establish the correct

Another limitation of GB Codes is that they do not currently extend below the county level, despite the fact that there are administrative divisions, such as *zhen* and *xiang* which can be delineated within counties. If the GB Codes included another hierarchical layer for the *zhen* and *xiang*, then we would have clearly defined sub-county "containers" to keep our local data in. These are the smallest administrative units that exhaust all the land area of the PRC, so we might as well use them.

In all likelihood, future editions of the GB Codes will include a sub-county layer. In the meantime, this paper will propose a method for extending the existing GB Codes to the sub-county level, as well as a scheme for further extension to identify point and polygon features within the sub-county regions. If at some point in the future a GB Code table with sub-county polygons is officially promulgated we won't have wasted our time because we can use a look-up table to integrate the GB Code extensions to the new official GB Codes.

3. Standard Syntax of GB Codes

xx yy zz

xx first pair = province

yy second pair = prefecture

zz third pair = county

The GB Codes are six digit numbers. The first two digits identify the ***ADM1*** level units: provinces, autonomous regions, and provincial level municipalities. All ADM1 level units will have zeroes in the second and third pair positions: ***xx0000***. So if we want to geocode any of our records as relating to entire provinces, these are the geocodes to mark them with. Incidentally, this is the sequence of provinces that you will find in the table of contents of most atlases published in the PRC since 1990.

The last two digits of standard GB Codes identify the ADM3 level units: counties, autonomous counties, county-level cities and *shixiaqu*, as well as banners. Sub-classifications for the ADM3 level are:

xxxx01 to xxxx18 = (a) ADM3 level municipalities and shixiaqu
(b) shixiaqu directly under ADM2 level municipalities

xxxx21 to xxxx80 = ADM2 counties and banners

xxxx81 to xxxx99 = ADM2 municipalities (directly under ADM1)

Note that *shixiaqu* as subdistricts of an ADM2 level municipality, are always numbered, ***01***, ***02***... using the third pair of the GB Code. These should be considered sub-districts within ADM2 level municipalities, rather than equivalent to the other ADM3 level units.

4. Sub-County Extensions of GB Codes

Up to this point, we've been looking at the official Guobiao Codes issued by the State Bureau of Technological Supervision (according to the June 21st, 1995 standard #GB/T 2260-95). Now let's take a look at my proposal to extend the GB Codes below the county level to incorporate all the data available to us about *zhen*, *xiang* and settlements. For convenience, we can refer to these sub-county regions as ***ADM4*** level units, and to the new codes as ***DBR Codes*** (to make sure they not confused with GB Codes). This system was devised in an attempt to assign unique identifiers to populated places in Yunnan, with considerable help and criticism from Prof. Lawrence Crissman (director of ACASIAN).[3] The syntax described is suitable for any province in China, and will be used for geocoding of sub-county data in Sichuan, Anhui, Gansu, and Tibet during the coming year.

The syntax for this type of GB code extension was pioneered by Prof. William Skinner, who added a fourth pair to the GB codes to identify sub-county units as part of his macro-regional analysis project.[4] The units being identified were:

xxxxxx00	**=**	***primary cities***
xxxxxx01, xxxxxx02, …	**=**	***zhen***
xxxxxx49	**=**	***shixiaqu chengguan*** ***(not listed as zhen in the census)***

Prof. Lawrence Crissman utilized these extensions to digitize of all the *zhen* in China into GIS. In the resulting dataset of Skinner/ Crissman *zhen* IDs, we have the syntax:

xxxxxx00	**=**	***primary cities located within ADM3 divisions***
xxxxxx01, 02, 03	**=**	***ADM4 zhen***
xxxxxx49	**=**	***ADM3 municipal districts (with an administrative seat distinct from the central city point)***[5]

The numbering sequence of the resulting dataset of Skinner/ Crissman point IDs follows the order of the *zhen* as they are listed in the census volumes, EXCEPT when the *zhen* that is the *chengguan*, (*ie*, the location of the county administration), was NOT listed first. In other words, if the *zhen* that is *chengguan* appears in the list after some other *zhen*, it still gets ***01*** in every case, then the remaining zhen are numbered in sequence--***02***, ***03*** … --beginning from the top of the list (and skipping the one already numbered ***01***).

This system is satisfactory for identifying *zhen*. However, in order to identify all the **sub-units** at the ADM4 level, both *zhen* and *xiang*, in addition to populated places within these sub-units, I needed to make at least some modifications for the proposed

1) For easy identification, there must be **separate** ranges of ID numbers for all ADM4 level sub-units: *zhen*, *xiang*, etc; and we must use a consistent source for their sequencing.

2) There must be a way to clearly differentiate points from polygons, but still preserve a relationship between the two codes if they are referring to the same object. We must provide for an expandable range of populated place IDs within each ADM4 level sub-division, that utilizes the same point-polygon conversion process.

3) For any additional objects that we want to identify within an ADM division, we need a system to allow for infinite expansion. Furthermore, there should be a way to clearly distinguish actual administrative division units, along with their subordinate populated places, from these other objects.

First, I had to distinguish *zhen* from *xiang*, and to determine the sequence of the IDs being assigned to them. In an earlier draft, an attempt was made to add two digits to existing GB Codes to represent the ADM4 level units. However, when applying that system to geocoding of *xiang* in Sichuan Province, it became apparent that the range of 48 open digits was simply not enough to account for all the *xiang* in several counties. Even though we can account for all the *zhen* in Chinese counties using a range of 48 digits, we expect *xiang* will continue to be promoted to *zhen* in the future, and we will be forced to rescale the system sooner or later. For this reason I advocate the use of three digit identifiers for ADM4 level sub-units, which offers a comfortable degree of scalability for the foreseeable future.

For *zhen*, I propose using: ***001*** to ***399***. For *xiang*, I propose using ***400*** to ***799***. The range ***800*** to ***998*** are reserved for other cases, such as the *shixiaqu chengguan* mentioned earlier, military bases, special industrial sites, and so forth, as long as they are **necessary elements** to exhaust the space within the next higher ADM3 unit.[6] ***999*** is reserved for the special case of assigning subordinate objects to ADM3, ADM2, or ADM1 units, explained in the following section.

Regarding the sequence used to assign DBR codes to ADM4 units, the common practice is to follow the order listed in the census, as mentioned above. However, in

publications are by no means widely available, and are not consistent with each other. They are not really suitable for use together as authoritative standards.

An alternative listing is found in the 1989 Index of Chinese Postal Codes [ICPC].[7] The ICPC list contains all of the *zhen* and *xiang* listed separately by county, for the year 1989. These can be checked against the 1991 GB Code tables for changes in jurisdiction, and can provide a standardized source for a sequence of ADM4 polygon units. For this reason, DBR codes always follow the sequence found in ICPC, keeping in mind that *zhen* and *xiang* are listed in seperate ranges, ***001 - 399*** and ***400 - 799***, respectively. The exception found in the Skinner / Crissman scheme--to always give the *zhen* that is *chengguan*, ***01***, regardless of its appearance in the census sequence--is **not preserved**, since any transfer of this status from one seat to another would require renumbering the entire list.

Relying on the ICPC, which is both **static** and to a large degree **exhaustive**, means that we can use it for sequencing once, and then leave it alone. The only changes needed will be to promote *xiang* to *zhen*, or to create new units. In the former case, we will need to retire the previously used numbers (which will identify those objects in their appropriate historical instance); in the latter case, we need only make use of the next available number; and in either of these cases, the use of three digits will provide adequate leeway.

For the second condition, to distinguish points from polygons, Prof. Crissman and I came up with a scheme to use a dot "." to indicate points. The dot "." placed between standard six-digit GB Codes and the proposed three digit extension, identifies the object as a point, not a polygon. Likewise, by removing the dot "." we can take a point feature and reclassify it as a polygon feature. In this way, we can take Shilin, Lu'nan: ***530126.002*** as point, and quickly reclassify it: ***530126002***, as a polygon. The key to maintaining the integrity of the system is to assign the ID ***002*** (within ADM3 polygon ***530126***) **only once**, to identify Shilin. In addition, we can use a hyphen "-" to

we need to have an expandable series of IDs, and we simply begin numbering ***1, 2, 3 …***

Our two points in Shilin, then, would be: ***530126002.1*** and ***530126002.2***. Likewise object number ***.2*** can be converted to a polygon: ***5301260022***. We can go on to assign more subordinate points: ***5301260022.1*** in order to identify a neighborhood, street, census tract, etc, within an administrative division… as long as they are elements of administrative division or populated place.[8]

When assigning IDs to populated places within ADM3 divisions, we **must not** confuse them with ADM4 level sub-units, and we **must not** give them two unrelated IDs. Therefore I propose that all populated places, which are not part of the exhaustive administrative division hierarchy, **always** be given IDs in the numbering sequence below ADM4. For example, a military garrison in Shilin, which is not a unique and independent polygon at the ADM4 level, would get the next available number in the polygon: ***530126002.3***, even though it is administratively subordinate to the prefectural capital (at the ADM2 level),

In order to accommodate objects that **physically** straddle the boundary between two or more ADM divisions, we must assign an ADM4 level ID as a placeholder for the higher level ADM units. I propose assigning the number ***999*** for this purpose in every case. If, for example, our garrison happened to overlap part of Yingpan *zhen* ***533325003*** and neighboring Lajing *zhen* ***533325002***, we would have to back up to the ADM4 level ***533325***, add on ***999***, and assign the ID: ***5333259991*** (polygon) and ***533325999.1*** (point). The reason for reserving ***999,*** is to prevent double-numbering.

Let's say that another garrison#2 actually straddled the boundary between Kunming municipality and Dongchuan Municipality. We are forced to use the ADM1 level for our base code: ***530000***. The problem is that no matter what number we assigned to this, if not part of our ***001-999*** reserved range, would cause a potential double-numbering later on. However, by using ***999*** in every case, we not only avoid

set. (This is useful if we have boundary changes to make). Therefore, the Kunming - Dongchuan garrison can be numbered: ***530000999.1***. If a different garrison#3 overlapped Kunming's Wuhua *qu* ***530102*** and Panlong *qu* ***530103***, we could go up one level, ***501000*** and add ***999***: ***530100999.1***. Even if we convert both garrison#2 and garrison #3 to polygons: ***5300009991***, ***5301009991***, we have no double-numbering problems.

I suggest that we **never** truncate the first six-digits of DBR Codes in order to assign IDs to subordinate populated places of ADM1 or ADM2 level units. In other words, *banshiqu* in Kunming ought to be ***530100999.2*** and ***530100999.3***, rather than ***5301.2*** and ***5301.3***. The primary reason to adopt this measure is to **always preserve** a direct relationship between official six-digit administrative division GB Codes and DBR Codes for the same object It also avoids the problem of having a truncation for Yunnan province, ***53*** combined with a higher number in the sequence, such as ***2100***: ***53.2100*** (point) and ***532100*** (polygon)--the problem being that the ID ***532100*** is already being used as the official GB code for Zhaotong Prefecture. By disallowing truncation of six digit GB Codes, we eliminate this problem.

What do we do with objects that are not necessary elements to exhaust ADM space or their subordinate populated places?

To provide for the third condition, geocoding of any other objects, I propose using an alpha-numeric sequence, as follows: any object that we want to identify, is assigned a number: ***1, 2, 3…*** , and **at the same time**, given a prefix to indicate which feature type (polygon, point, or line) is being identified. I propose using the prefixes:

a = polygon, ***b = point***, ***c = line***.

It is essential that each number in the sequence refers to a **unique object**. For example, if we want to identify a reservoir in Shangyong *xiang* as a point we would

numbers within Kunming: ***530100a1*** and ***530100a2***.[9] To go further, let's say that within the first watershed, we have a few habitat zones to identify: ***530100a1a1, 530100a1a2, 530100a1a3***. And in the second watershed we want to identify the point where a factory drain pipe empties into the stream: ***530100a2b1***. As long as an authority list is maintained, we will never run out of available numbers for assigning new features.

The advantages gained by reserving alpha-numerics for non-administrative objects is to **clearly differentiate** them from administrative divisions. Any DBR Code containing an alphabetic character **must** be non-administrative, in the sense that it identifies an object that is not part of the administrative division hierarchy, or a subordinate populated place.

4. Concluding Remarks

Let's take a look at the DBR Codes in a nutshell:

Six digit GB Code

PLUS ***three digit ADM4 unit extension (zhen 001-399, xiang 400-799, other 800-998)***

OR ***(populated place ID, 999… that overlaps more than one ADM level division)***

OR ***subordinate other feature above ADM4 level (a1, a2, b1, b2)***

DBR Code

PLUS ***subordinate populated place (.1, .2, .3)***

OR ***subordinate other feature (a1, a2, b1, b2)***

Other types of objects, whether they be man-made, natural, historical, or mythical, can be assigned IDs that are immediately differentiated from the ADM hierarchy by use of alpha-numeric combinations. These can be sub-divided and numbered without any limitations. All populated places are given IDs following the three-digit extensions (**001-999**), beginning ***.1, .2, .3*** … , allowing for quick sorting and querying functions and preventing truncation of traditional GB Codes.

After the animated discussion of the proposed DBR codes system at the PNC meeting in Berkeley, January 2000, I believe that we should not inject subject matter classification schemes into DBR codes. They should be plain vanilla assignments of ID labels to points and polygons ONLY, and leave all other forms of sub-classification of features for separate fields in our database. The DBR codes should be purely and simply: "indirect spatial references," and we can leave the sub-classification schemes to our the Dublin Core and cataloguing specialists.

Adopting an indirect geocoding system for China, such as the DBR codes presented here, will give us an excellent tool for geocoding the site-specific and disaggregate data that we are all beginning to stockpile in digital form. The use of geocodes as meta-tags will allow us to query and cross-tabulate information in ways that we never thought of before, providing new perspectives and new, intriguing problems. For some researchers who deal with very site-specific data, the availability of DBR codes on the settlement point level, as well as the flexibility to create new sub-unit point and polygon codes whenever necessary, will be useful to differentiate villages that are located in the same *zhen* or *xiang*, and to incorporate data from extremely small scale maps or other multi-dimensional spatial data needs to be captured. In this way, I hope that our disparate data sets can be integrated and cross-referenced, opening up new and exciting possibilities for future research.