

## **Attribute Accuracy**

### **Objectives (Entry)**

This basic concept of attribute accuracy has been introduced in the unit of quality and coverage. This unit will teach a basic technique to quantify the attribute accuracy. At the end of the unit the student will be able to assess the attribute accuracy based on simple error matrix.

### **Attribute Accuracy (Clarification)**

#### **1. Introduction to Attribute Accuracy**

Attribute accuracy indicates the attribute attached to the points, lines and polygons features of the spatial database, which are how reliable and reasonably correct or free from bias.

Qualitative attribute accuracy refers to whether nominal variables or labels are correct, such as assigning wrong land use type in the land use map.

Quantitative accuracy refers to the level of bias in estimating the values assigned such as estimated values of pH in a soil map.

Attribute errors can be caused by doing inventory work from aerial photography and misinterpretation of images. Moreover, it may often occur because base maps are relied on too heavily. Simple typing error may cause attribute error such as wetland paddy fields may appear on hilltops. More complex attribute errors may be due to the sampling strategies that produced the original data from image or aerial photo or field. Although error of sampling technique may exist, such information is seldom included in the GIS database.

A spatial database consists of digital version of real world objects, e.g. power pole, digital version of artificial map features which are fictitious and do not exist in the real world, e.g. contours, and artificial objects created for the purpose of the data base, e.g. pixels.

Phenomena or characteristics of real world objects exist everywhere and vary continuously over the earth's surface. These variations are represented in several ways, by taking point sampling, transects, zoning assuming the variable is constant within each zone and drawing contours.

These methods create discrete objects such as point, line and polygon and a raster as point sample where the points are regularly spaced or as zones where the zones are all the same size of resolution or raster as continuous surface such as elevation or temperature.

These methods approximately measure only part of the variation. Point sampling method misses variation between points, transect method misses variation not on the transects, zoning method pretends that variation is sudden at boundaries, and there is no variation within zones and contours miss variation not located on contours.

Moreover the phenomena can be observed in spatial, temporal and thematic dimensions. Spatial dimension deals with variation from place to place; temporal dimension deals with that of time-to-time and thematic dimension deals with that of one characteristic to another. Attributes are captured at the thematic dimension by defining different characteristic of object at a particular place and at a particular time.

Therefore attribute accuracy depends on the method of measurements to represent the real world object, time of measurement (phenology, crop calendar) and place.

The attribute errors can be occurred due to multiple representation of same phenomenon in different ways in different data models (vector, raster and TIN etc.) at different scales. Moreover it is difficult to convert from one representation to another.

## **2. Simple Error Matrix and Kappa Value**

Error matrix and Kappa Index is normally used traditionally in order to test the qualitative attribute accuracy. It is site-specific accuracy testing.

1. The following simple matrix is constructed between the reference data and map data. F represents forest, W represents water and U represents urban area. The matrix can be interpreted as follow. Total sample points of forest are 30 in reference data. 28 sample points out of 30 forest sample points is truly classified as forest. 1 forest sample point is classified as water and 1 forest sample point is classified as urban. Other samples can be interpreted in a similar way.

		Reference Data			
		F	W	U	Row Total
Map Data	F	28	14	15	57
	W	1	15	5	21
	U	1	1	20	22
	Columns Total	30	30	40	100

### Attribute Categories

F = Forest

W = Water

U = Urban

Overall Accuracy =  $(28 + 15 + 20) = 63\%$

Producer's accuracy

F =  $(28/30) * 100 = 93\%$  W =  $(15/30) * 100 = 50\%$  U =  $(20/40) * 100 = 50\%$

User's Accuracy

F =  $(27/57) * 100 = 49\%$  W =  $(15/21) * 100 = 71\%$  U =  $(20/22) * 100 = 91\%$

2. Calculate the q, the number of cases expected in diagonal cells by chance.

$$q = n(\text{row}) * n(\text{col}) / N$$

$$F = (57 * 30) / 100 = 17.1$$

$$W = (21 * 30) / 100 = 6.3$$

$$U = (22 * 40) / 100 = 8.8$$

$$q = 17.1 + 6.3 + 8.8 = 32.2$$

### 3. Calculate the Kappa

d = diagonal total of cells = 28 + 15 + 20

N = total of columns or rows = 100

Kappa = (d-q) / (N-q)

Kappa = (63 - 32.2) / (100-32.2) = 0.4543

The kappa value does not approximate to 1. Therefore, the attribute accuracy is not very high.

Overall accuracy incorporates only the diagonal cell values and excludes the omission and commission errors or off-diagonal cell values.

KHAT accuracy index indirectly incorporates off-diagonal cell values as the product of the row and column totals.

### 3. Number of Samples per Class for Accuracy Testing

The enough number of samples that represent the thematic classes and ensure good distribution across the map is important to test the attribute accuracy.

Rule of thumb is 50 samples per map class or can be derived using the multinomial calculation or simplified multinomial calculation.

From Tortora (1978), the Multinomial Calculation is

$$n = B \prod_i (1 - \Pi_i) / b_i^2$$

n = sample size

B = upper ( $\alpha/k$ ) \* 100<sup>th</sup> percentile of the  $\chi^2$  with 1 degree of freedom

$b_i$  = absolute precision of each cell

$\Pi_i$  = proportion of class i in the map

This equation is simplified as follow.

$$N = B / 4 b^2$$

For example, there are 17 categories in the map attribute. Therefore k is 17. The desired confident level is 95% and that the desired precision error tolerance is 5%. The value for B must be determined from a Chi square table with 1 degree of freedom and 1-a /k. In this case appropriate value for B is  $X^2(1, 0.997) = 9.3535$ . Therefore, calculation of sample size is as follow.

$$n = 9.3535/4(0.05)^2 = 936$$

Therefore approximately 55 samples per class or 936 total samples would be required.

### Calculate the following (Act)

1. Calculate the sample size to evaluate the attribute accuracy for a map that has 15 Categories.
2. Calculate the overall, producer and user attribute accuracy of the following data set.

Land cover	A	B	C	D	E	F
A: Mixed Forest	8	0	1	0	2	0
B: Paddy	0	10	0	2	0	1
C: Plantation	3	0	21	0	4	0
D: Corn	0	0	0	9	0	3
E: Pine	4	0	5	0	23	0
F: Sugarcane	0	2	0	3	0	14

### Examples of Attribute Values (Look)

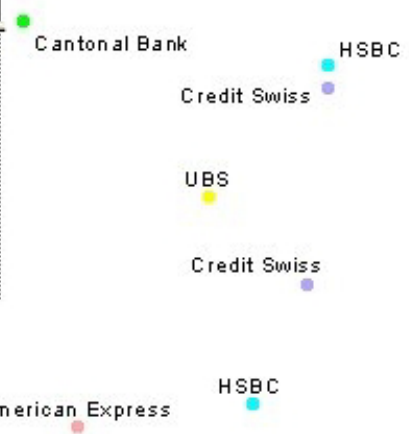
#### 1. Nominal

It is any name or numbers. Numbers merely establish identity. It allowed logical operations, classification and identification.

**Concept :** Nominal Attributes

<i>Banks_id</i>	<i>Branch</i>	<i>Bank_num</i>	<i>Year_estab</i>
538	UBS	2116	1938
539	UBS	2116	1909
542	Credit Swiss	20147	1990
543	Credit Swiss	20147	1934
549	Cantonal Bank	2155	1800
553	American Express	26382	1985
555	HSBC	867	1982
556	HSBC	867	1891

- American Express
- Cantonal Bank
- Credit Swiss
- HSBC
- UBS



Banks\_id, Branch, Bank\_num attributes are Nominal Attribute. Nominal numbers do not indicate any order or relative value. It only identifies the property.

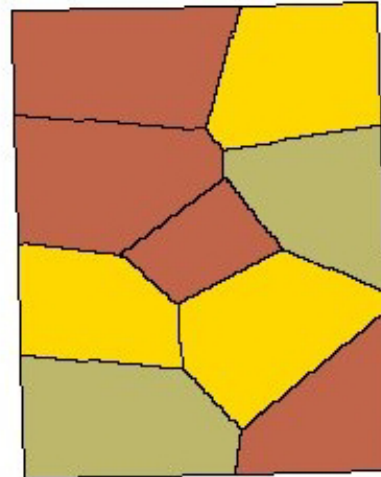
## 2. Ordinal

Ordinal attribute numbers establish order only. Functions on logical and ranking operations and comparison of magnitude operations are allowed.

**Concept :** Ordinal Attributes

<i>Shape</i>	<i>Id</i>	<i>Suit</i>	<i>Suitclass</i>
Polygon	1	1	Best
Polygon	2	2	Medium
Polygon	3	1	Best
Polygon	4	3	Worst
Polygon	5	1	Best
Polygon	6	2	Medium
Polygon	7	2	Medium
Polygon	8	3	Worst
Polygon	9	1	Best

Suitability ranking of forest plantation



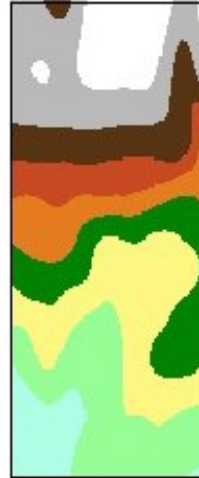
Suitability attributes is ordinal scale that establish the order of suitability.

### 3. Interval

On the interval scales, the difference interval between numbers is meaningful.

**Concept :** Interval Attributes

Interval scale of elevation in meter



The interval is 4.765 meter.



END



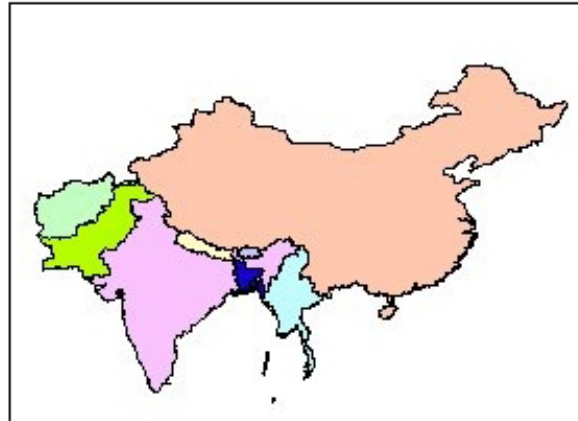
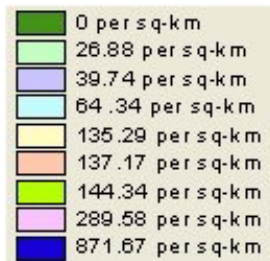
#### 4. Ratio

On a ratio scale, measurement has an absolute zero and the difference between number is significant and division makes sense. Population density map is the example of ration scale measurement.



**Concept :** Ratio Attributes

Population density map of Hindu Kush Himalayan Countries



**Concept :** Ratio attributes

<i>Cntry_name</i>	<i>Sovereign</i>	<i>Pop_cntry</i>	<i>Sqkm_cntry</i>	<i>popdensity</i>
Afghanistan	Afghanistan	17250390	641869.188	26.88
Bangladesh	Bangladesh	120732200	138507.203	871.67
Myanmar (Burma)	Myanmar (Burma)	43099620	669820.875	64.34
Bhutan	Bhutan	1586631	39927.012	39.74
China	China	1281008318	9338902.000	137.17
India	India	894608700	3089282.000	289.58
Nepal	Nepal	19927280	147292.594	135.29
Pakistan	Pakistan	126693000	877753.375	144.34

The population density attribute (*popdensity*) is ratio attribute, which is derived based on Population of country (*pop\_cntry*) and Area of country (*sqkm\_cntry*).

END

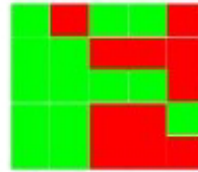
## 5. Boolean


Boolean attribute has 0 and 1 to indicate presence and absence or Yes and No. It is useful for logical and indicator operations such as truth versus falsehood.

**Concept :** Binary Attributes



Presence and Absence of Marmota marmota



 1 Presence of Marmota marmota

 0 Absence of Marmota marmota

The attribute contains 0 and 1 values. 0 represents absence and 1 represents presence.



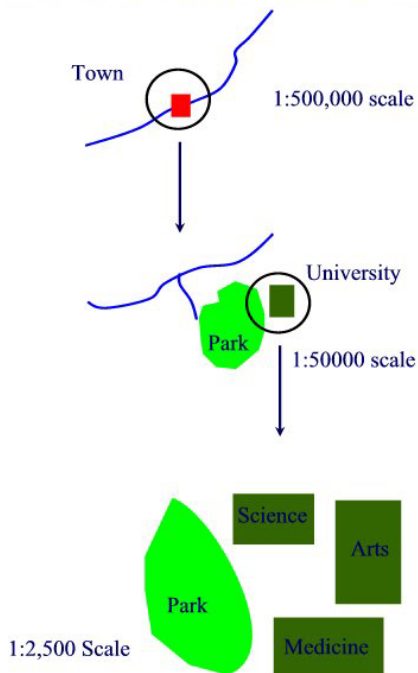
END



## 6. Multiscale Representation

A phenomenon or characteristics or attribute of a real world objects represented in different ways at the different scale.

Attributes of real world objects in different scale



### Exercises (self assessment)

1. Review the facts that influence the attribute accuracy of real world objects and phenomenon.
2. The 1990 land use map is 1:250000 scales. The 2000 land use map is 1:50000 scales. Is it logical to overlay the two maps to analyse the land use changes. Explain your opinion.