ACCURACY INVESTIGATION FOR A LARGE SCALE GIS

T. M. Celikoyan [a*], M. O. Altan [a], G. Kemper [b]

[a] ITU, Civil Engineering Faculty, 80626 Maslak Istanbul, Turkey - (celikoyan,oaltan)@itu.edu.tr
[b] GGS, Kämmererstr.14, 67346 Speyer, Germany, kemper@ggs-speyer.de

ThS 9

KEY WORDS Accuracy, Thematic, Geometric, Database, GIS

ABSTRACT:

In this study, the accuracy of a geographical information system is investigated. The mentioned accuracy is not only the geometric accuracy, but also the thematic accuracy. It is known that the standards for geographical information system are already a popular and topic in both practical and academic world. This study gives an example for the investigation of accuracy.

Geometric accuracy can be derived using the errors between the geometric results and the real values. The real values can be calculated precise surveys or taken from maps of larger scale than the scale which is used in the study. This choice depends on the dimensions of the study area and the scale of the study.

Thematic accuracy can be derived from differences of individual defined and to database added objects. These differences are the errors of definitions.

The originality of accuracy investigation within geographical information system is to decide, how many objects from itself are used for within this investigation. Ideally, all of the objects must be used for this thematic accuracy investigation. But it is nearly impossible to do this investigation with all the objects if a big amount of them are used within a GIS project. This disadvantage forces to used sampling methods.

In this study, an accuracy investigation of a MOLAND project for Istanbul has been done. By geometric accuracy investigation, a sampling class size has been defined for all the object types within the project and these samples are chosen from database randomly. Afterwards, the true values are taken from 1:1000 digital maps for referring samples. The errors are calculated and from these errors, a global geometric accuracy values are derived. For the thematic accuracy investigation, sampling class sizes are determined for all different landuse types. This process gives a number of approx. 3000 for sampling class.

1. INTRODUCTION

The International Standardization Organization (ISO) explains the quality in ISO 8402 as "the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs". Quality is defined as "that which makes or helps to make anything such as it is; a distinguish property, characteristics, or attribute; (in logic) the negative or affirmative character of a proposition (Buttenfield, 1993). Another definitions of quality are "the sum of all the factors that enable ownership satisfaction and bring customers back to buy a product or a service again and again" (Radwan et al., 2001) and "the integration of the terms of completeness, consistency and suitability for use, which are statistically measurable" (Beard and Mackaness, 1993). More general definition of quality is "a statistically measurable concept showing the totality of indicators defining the negative and affirmative distinguishing specifications and characteristics in the fields of the satisfaction of the desired conditions for the producer and appropriateness to the aim for the consumer of a product or a service". (Taştan, 1999)

If the quality is dealt with, accuracy must be defined, which affects the quality directly. The term "accuracy" is the degree of conformity with a standard. It relates to a quality of the results and is distinguished from precision, which relates to a quality of the operation by which the result is obtained (Glossary of Cartographic terms). Another definition of accuracy is "closeness of agreement between a test result and the accepted reference value" (ISO/TC 211 N 1034, 2001).

In geoinformatic, data, which is processed, is mostly spatial data. Data about positions, attributes and relationships of features in space are termed spatial data (Morrison, 2001). Combining the definitions of quality and spatial data, the definition of spatial data quality occurs as "a statistically measurable concept showing the totality of indicators defining the negative and affirmative distinguishing specifications and characteristics in the fields of the satisfaction of the desired conditions for the producer and appropriateness to the aim for the consumer of data about positions, attributes and relationships of features in space" (Celikoyan, 2004).

2. BRIEF INFORMATION ON THE PROJECT AND ITS RESULTS

2.1 Data

After investigating the usable historical data, the historical time periods for landuse detection occur as 1940'ies, 1968. These years were the only periods, in which aerial photographs of Istanbul had been taken. For further years, 1988 and 2000 has been selected. Technical data contains topographical maps from scales of 1:25000 and 1:5000, IKONOS, IRS 1C/D images for

2000, KVR-1000 images for 1988 and airborne photographs for 1968 and 1940'ies. The first time period is mentioned as 1940'ies, because the airborne photos for whole study area had not been taken in the same year. They had been taken in different periods of 1940'ies between 1942 and 1949.

2.2 Study Area

Study area has been selected not only as the city center of Istanbul, but also containing the surrounding territories of Istanbul, which can affect the landuse and its change as well as their interpretation and analysis. The study area contains two zones, the center and buffer zones, as illustrated in Figure 1.

Figure 1. Center (grey) and buffer zone (light grey) with water surfaces (dark grey)

2.3 Data Preparation

Data preparation has been begun with the year 2000, which will be mentioned as reference year. Map to image method was selected for georeferencing the satellite image. All maps have been scanned with 600x600 dpi geometric and 24-bit radiometric resolution using a HP 4C/T DIN A4 flatbed scanner. Georeferencing has been done by using TopoL-GIS

software, coordinate transformations by an Excel calculation. The first georeferencing of maps is a normal Affine-transformation to place the scanned maps to the position of the maps origin coordinate system to enable a good orientation and to detect more easily the displacement in the next step. That was only the first and approximate transformation and done by using the map corners. Afterwards, all the maps are transformed exactly to the corresponding coordinate system using all the grid points on them. With this partial transformation, the maps are not only georeferenced also the errors are reduced. With this transformation, perfect neighborhoods and a mistake minimization of scanning and paper sheets (and sometimes also of the print) effects have been got. Then a third Affine-transformation, to move the image to the other coordinate system (from one to another strip or ellipsoid), had to take place. The correct coordinates have been calculated by the map-corner coordinates or by the grid on the map-sheets.

After georeferencing the topographical maps, satellite images were georeferenced using image to map method. In that way, all the satellite images were ready to digitizing process. Following process was the creation of orthophotos for 1968 and 1940'ies. Totally 94 orthophotos for 1968 and 114 orthophotos for 1940'ies were produced.

2.4 Map Production

Manual method was selected because the richness of legend, in which 62 different landuse types was defined in total four levels. This caused a very long and hard work in map production. Each of four landuse detection step has three layers. The first one is deals with linear objects such as transportation facilities and rivers where the second one deals with areas and the third with 3D areas such as bridges, viaducts or tunnels. As an overview, Figure 2 shows the manual digitizing and classification results for the first level of legend.
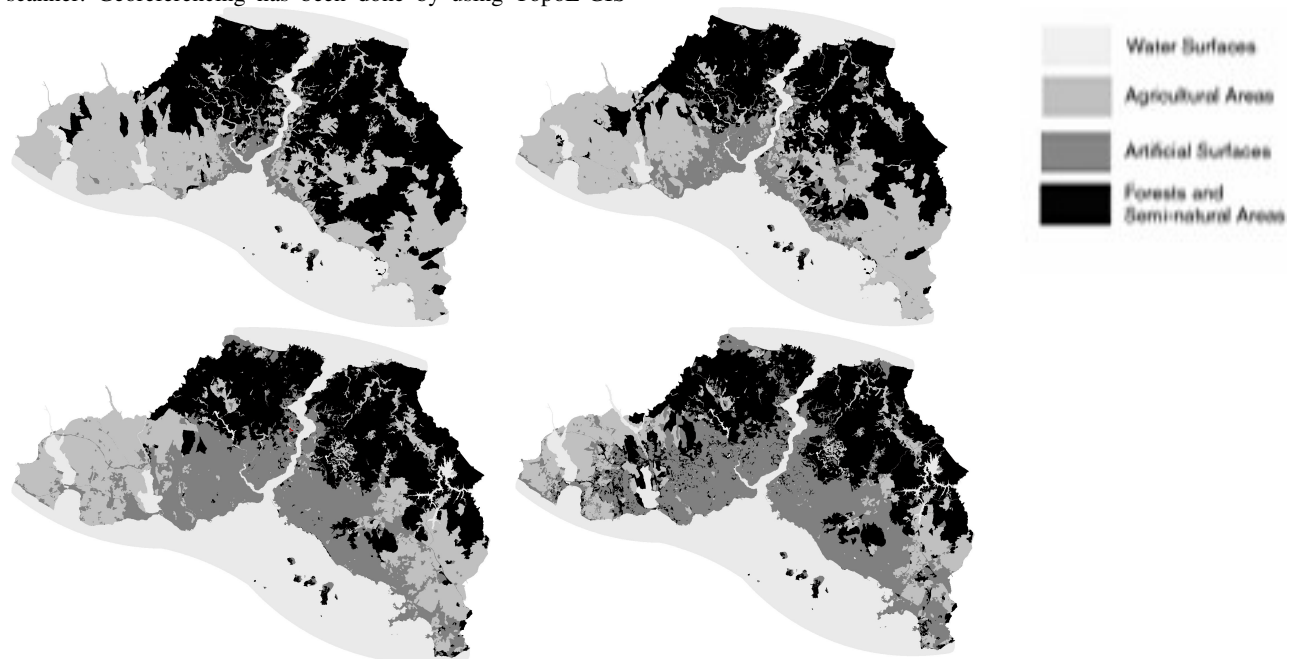

Figure 2. Results of the digitalization and manual classification for all four evaluation years. (left upper: 1940'ies; right upper: 1968; left lower: 1988; right lower: 2000)

3. ACCURACY ASSESSEMENT

The accuracy assessment has two components, which are geometric and thematic accuracies. These two components are different from each other basically but they act an integrating role for each other.

3.1 Geometric Accuracy

The components of geometric accuracy of the results are;
- Accuracy of hardcopy reference maps (1:25000)
- Accuracy of scanning
- Accuracy of map georeferencing
- Accuracy of satellite image georeferencing
- Resolution of satellite images
- Errors done by operator

In praxis, it is not possible to compute an exact geometric accuracy value for the results. But some approximation can be done using these components.

More simple assessment on the geometric accuracy of the results was done by determining the residuals. For this purpose a sampling method was used. A stratified random sampling design was implemented to digitizing results.

For the geometric accuracy of the points, 60 samples were selected randomly from database, where the samples defines 30 lines overall. This means, 60 points are selected so that they are beginning and ending points of 30 lines. This is done in that way in order to make the accuracy investigation easier for lines. Coordinate residuals of these points was calculated using referring point from 1:1000 digital maps. In this step, coordinates from 1:1000 maps were taken as "true values".

The deviations between the sampling points and corresponding points from 1:1000 maps have been derived and extreme values are given in Table 1.

[m]	Observed Coordinates		Coordinates from 1:1000 map		Deviations	
#	x	y	X	Y	ε_ξ	ε_ψ
10	4548627	407046,7	4548627	407046,7	0,0	0,0
2	4553876	415854,4	4553883	415870,6	-7,4	-16,2
54	4541529	434594	4541518	434571,8	11,2	22,2

Table 1. Observed and true coordinates of sampling points with their deviations (Extreme values)

Where;

$$\varepsilon_{x_i} = x_i - X_i \text{ and } \varepsilon_{y_i} = y_i - Y_i$$

With the deviations, the following values about geometric accuracy have been calculated.

Theoretical standard deviations are,

$$\sigma_{\varepsilon_x} = \sqrt{\frac{\sum \varepsilon_{x_i}^2}{n}} = 5.027\,m.$$

$$\sigma_{\varepsilon_y} = \sqrt{\frac{\sum \varepsilon_{y_i}^2}{n}} = 6.187\,m.$$

$$\sigma_c = \sqrt{\frac{1}{2}\frac{\sum \varepsilon_{x_i}^2 + \sum \varepsilon_{y_i}^2}{n}} = \sigma_c = \sqrt{\frac{\sigma_{\varepsilon_x}^2 + \sigma_{\varepsilon_y}^2}{2}} = 5.637\,m.$$

where σ_{ex} and σ_{ey} are the standard deviations in X- and Y directions and σ_c is the circular standard deviation. Further values on accuracy were calculated as follows:

Circular probable error:
$$CPE = \sqrt{1.39\sigma_c^2} = 1.1774\sigma_c = 6.646\,m.$$
Mean square positional error
$$MSPE = \sqrt{2\sigma_c^2} = 1.4142\sigma_c = 7.971\,m.$$
Circular map accuracy standard:
$$CMAS = \sqrt{4.610\sigma_c^2} = 2.146\sigma_c = 12.102\,m.$$

Circular near certainty error:
$$CNCE = 3.5\sqrt{\sigma_c^2} = 3.5\sigma_c = 19.728\,m.$$

After determining positional accuracy of points, accuracy of lines has been derived using these values. A positional accuracy of a line can be shown using the accuracy of the midpoint of it. Using the deviations of the beginning- and endpoints, Table 2 gives the positional accuracy of the sampling lines.

[m]	Deviations		[m]	Deviations	
#	ε_x	ε_y	#	ε_x	ε_y
1	3,742	8,667	16	1,742	4,084
2	0,764	1,974	17	0,994	3,028
3	1,46	0,268	18	1,123	2,658
4	2,861	0,965	19	4,998	1,586
5	0,4	0,824	20	4,463	5,817
6	1,042	0,831	21	3,02	3,408
7	0,342	0,089	22	1,249	1,063
8	0,56	2,221	23	3,465	4,702
9	2,014	1,672	24	5,865	6,573
10	0,888	1,077	25	6,607	2,948
11	2,356	2,246	26	5,032	10,011
12	0,246	1,002	27	7,189	14,072
13	1,284	0,295	28	8,232	0,213
14	4,663	2,787	29	3,73	2,538
15	0,92	2,418	30	2,854	1,828

Table 2. Deviations of the sampling lines

Using the deviations in Table 2, a general positional accuracy can be calculated as follows;

$$\sigma_{\varepsilon_x} = \sqrt{\frac{\sum \varepsilon_{x_i}^2}{n}} = 3.554\, m.$$

$$\sigma_{\varepsilon_y} = \sqrt{\frac{\sum \varepsilon_{y_i}^2}{n}} = 4.375\, m.$$

$$\sigma_c = \sqrt{\frac{1}{2}\frac{\sum \varepsilon_{x_i}^2 + \sum \varepsilon_{y_i}^2}{n}} = \sigma_c = \sqrt{\frac{\sigma_{\varepsilon_x}^2 + \sigma_{\varepsilon_y}^2}{2}} = 4.712\, m.$$

Error band illustrates the accuracy for the lines in much more understandable and interpretable way. In order to draw these error bands for all the lines in sampling group, proportions are selected as ¼, ½ and ¾. The deviations on these portions of the sampling lines are calculated as in Table 3.

[m]	Deviations ε_x			Deviations ε_y		
#	Portion ¼	Portion ½	Portion ¾	Portion ¼	Portion ½	Portion ¾
1	1,953	3,742	5,586	6,180	8,667	12,231
2	0,848	0,764	0,860	2,167	1,974	2,246
3	2,014	1,460	1,130	0,382	0,268	0,182
4	1,438	2,861	4,290	1,444	0,965	0,495
5	0,601	0,400	0,200	1,236	0,824	0,412
6	0,609	1,042	1,530	1,240	0,831	0,431
7	0,421	0,342	0,339	0,130	0,089	0,053
8	0,800	0,560	0,382	3,167	2,221	1,519
9	2,521	2,014	1,946	2,493	1,672	0,880
10	0,763	0,888	1,179	0,697	1,077	1,554
11	1,632	2,356	3,349	3,332	2,246	1,230
12	0,187	0,246	0,340	1,499	1,002	0,512
13	0,834	1,284	1,852	0,364	0,295	0,293
14	4,763	4,663	5,627	1,796	2,787	4,023
15	1,034	0,920	1,022	2,794	2,418	2,609
16	1,323	1,742	2,415	3,917	4,084	5,134
17	0,997	0,994	1,214	3,788	3,028	2,928
18	1,390	1,123	1,105	3,817	2,658	1,758
19	5,024	4,998	6,100	1,222	1,586	2,191
20	6,114	4,463	3,523	7,483	5,817	5,349
21	4,258	3,020	2,163	3,077	3,408	4,424
22	0,673	1,249	1,856	1,190	1,063	1,188
23	4,634	3,465	2,922	2,926	4,702	6,834
24	8,752	5,865	3,067	9,850	6,573	3,317
25	7,689	6,607	7,071	2,420	2,948	3,984
26	6,899	5,032	3,965	11,393	10,011	10,988
27	7,288	7,189	8,722	14,101	14,072	17,212
28	8,678	8,232	9,702	0,143	0,213	0,306
29	5,573	3,730	1,928	2,398	2,538	3,217
30	3,413	2,854	2,951	1,360	1,828	2,551

Table 3. Deviation on predefined portions along sampling lines

Using the deviations given in Table 3, all error ellipses are drawn along the sampling lines in order to show the error bands for all the sampling lines as in Figure 3.
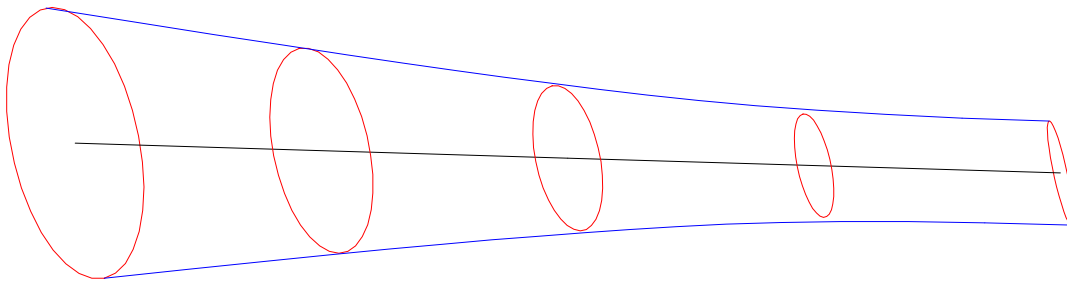

Figure 3. Error ellipses(red) and error band (blue) of a line (black)

A generated accuracy criterion for areas is given in Taştan (1999) as follows.

$$\sigma_F = \sigma_c \sqrt{2F \, \mathrm{Sin}\left(\frac{360}{n}\right)}$$

where F is the value of the area and n is number of edges. Using this formula, geometric accuracy of areas can be given as a graph.

As it can be seen from the formula, accuracy of an area depends on the area and number of edges of itself. Figure 4 shows dependencies of accuracy of area with its area and its number of edges.
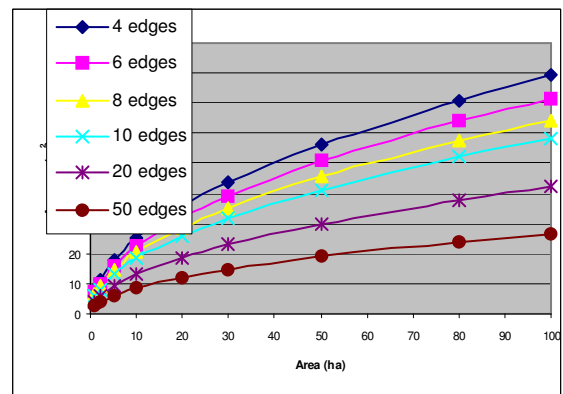

Figure 4 Geometric accuracy changes of areas from different edge numbers and areas

3.2 Thematic Accuracy

For the year 2000, 12816 polygons have been created. That causes the thematic accuracy investigation not applicable for a full investigation. Because of that, sampling groups have been created randomly from database for all the landuse types used in database. The rule for sampling is illustrated in Table 4.

Number of Polygon	Sampling size
30 or more	30
Between 20 – 29	20
Between 10 - 19	10
Between 1 – 9	All

Table 4. Sampling rule for thematic accuracy investigation

Applying the rule to the database for the year 2000, the sampling sizes has been calculated as in Table 5.

Landuse Type	Sampling Size	Landuse Type	Sampling Size	Landuse Type	Sampling Size
1.1.1.1	30	1.2.2.6	30	2.4.4	7
1.1.1.2	30	1.2.2.7	20	3.1.1	30
1.1.1.3	30	1.2.3	30	3.1.2	30
1.1.2.1	30	1.2.4	4	3.1.3	30
1.1.2.2	30	1.3.1	30	3.1.3.1	1
1.1.2.3	30	1.3.2	1	3.2.1	2
1.1.2.4	3	1.3.3	30	3.2.1.1	30
1.2.1.1	30	1.3.4	30	3.2.1.2	30
1.2.1.2	30	1.4.1	30	3.2.1.3	10
1.2.1.3	30	1.4.1.1	30	3.2.4	10
1.2.1.4	20	1.4.2	30	3.2.4.4	2
1.2.1.5	10	2.1.1	30	3.3.1.1	1
1.2.1.6	20	2.1.1.1	30	3.3.1.2	10
1.2.1.7	10	2.1.1.2	5	3.3.2.2	30
1.2.1.8	10	2.1.1.3	3	3.3.3.1	2
1.2.1.9	30	2.1.2	30	4.1.2	2
1.2.1.10	10	2.3.1.2	4	5.1.1.1	6
1.2.2.1	30	2.4.2.1	4	5.1.1.2	8
1.2.2.2	30	2.4.2.2	30	5.1.2.1	10
1.2.2.3	10	2.4.3	4	5.1.2.2	20
1.2.2.4	1	2.4.3.1	30	5.2.3	5
1.2.2.5	1	2.4.3.2	2	**Total**	**1168**

Table 5 Sampling sizes for landuse types

After calculating the sampling sizes, samples have generated using the random property of computer. The sample polygons are identified with their unique number in database.

For all the polygons in the sampling group, a thematic accuracy process has been done manually. This was a very long and monotonous work but in order to make an assessment about the thematic accuracy, it was a must. All the polygons have been controlled manually if they defined correctly by digitalisation process. As a result, some defining errors about the landuse have been detected but the amount of the errors was acceptable. Errors are given in Table 6 with their percentage to the sampling groups.

Landuse Type	Error %	Landuse Type	Error %	Landuse Type	Error %
1.1.1.1	6,7	1.2.2.6	3,3	2.4.4	0,0
1.1.1.2	6,7	1.2.2.7	5,0	3.1.1	0,0
1.1.1.3	6,7	1.2.3	0,0	3.1.2	0,0
1.1.2.1	10,0	1.2.4	0,0	3.1.3	0,0
1.1.2.2	10,0	1.3.1	10,0	3.1.3.1	0,0
1.1.2.3	6,7	1.3.2	0,0	3.2.1	0,0
1.1.2.4	0,0	1.3.3	0,0	3.2.1.1	6,7
1.2.1.1	0,0	1.3.4	10,0	3.2.1.2	6,7
1.2.1.2	13,3	1.4.1	6,7	3.2.1.3	0,0
1.2.1.3	6,7	1.4.1.1	0,0	3.2.4	0,0
1.2.1.4	0,0	1.4.2	0,0	3.2.4.4	0,0
1.2.1.5	0,0	2.1.1	6,7	3.3.1.1	0,0
1.2.1.6	0,0	2.1.1.1	0,0	3.3.1.2	0,0
1.2.1.7	0,0	2.1.1.2	0,0	3.3.2.2	0,0
1.2.1.8	0,0	2.1.1.3	0,0	3.3.3.1	0,0
1.2.1.9	0,0	2.1.2	0,0	4.1.2	0,0
1.2.1.10	0,0	2.3.1.2	0,0	5.1.1.1	0,0
1.2.2.1	0,0	2.4.2.1	0,0	5.1.1.2	0,0
1.2.2.2	0,0	2.4.2.2	6,7	5.1.2.1	0,0
1.2.2.3	0,0	2.4.3	0,0	5.1.2.2	0,0
1.2.2.4	0,0	2.4.3.1	0,0	5.2.3	0,0
1.2.2.5	0,0	2.4.3.2	0,0	**Total**	**3,3**

Table 6 Error percentages of landuse classes

Table 6 gives information about the accuracy of the results. For example, it can be said that the landuse type 1.1.1.1 (Residential continuous dense urban fabric) has been defined with an error of 6.7% where the landuse type 3.1.3 (Mixed forest) has been defined without error.

In Table 6, 48 landuse types have been shown as "totally true". 30 from these 48 landuse types have 10 or less sampling size i.e. they are few in total amount. This causes to define them much more correct than other landuse types. Other 18 landuse types, which are defined as "totally true", have whether specific information about their landuse on 1:25000 maps (e.g. vegetated cemeteries) or they are clearly visible and detectable from aerial or satellite imagery (e.g. industrial areas or seas). Using the additional data such as 1:25000 maps or 1:16000 city plans by digitalisation and defining the areas causes accurate results. Another parameter, which causes the results to be accurate, is the knowledge of operator on Istanbul. Defining the fast transit road and associated land can be a suitable example for this parameter.

In Table 6, it can be seen that a total error percentage is given as %3.3. This value is only a rough result for total accuracy. In order to give more precise result about accuracy, another investigation must be done with the values in Table 6. In this investigation, error percentages for individual landuse types must be taken with their weight. The weight is selected as the percentage of number of landuse type to the total number of defined polygons.

In order to derive a total accuracy result, Bayes theorem must be used for the results.

$$P(F) = \sum \left[P(F/L_i) * P(L_i) \right]$$

where P(F) denotes the percentage of false defined polygons in total area, P(L$_i$) the percentage of the landuse type within the all polygons and P(F/L$_i$) the percentage of false defined polygons within the ith landuse type. By using the formula above, a total accuracy result of %94.31 was determined.

4. CONCLUSION

As in all topics in geoinformatics, accuracy assessment is an integrating part of a GIS. Without accuracy, GIS does only serve for limited aims. In this study, an accuracy assessment for both geometry and consistence was done. For this purpose, an external knowledge for geometric accuracy, an internal knowledge for thematic accuracy was used. The results on both geometric and thematic accuracies shows that the results of the landuse detection study can serve some disciplines such as urban and regional planning, transportation planning and management as well as facility planning and management. By interpreting the accuracy results, it must be taken into account that the scale of the landuse detection project was 1:25000. This will be a factor to accept the applicability of results for further studies. Moreover, this study shows the applicability of accuracy investigation for GIS's, which have big amount of record in database.

Manual digitizing method let the work be harder. Furthermore, in some cases this causes the accuracy to be heterogeneous in whole project. But thinking on the richness and complexity of the legend for landuse classes, any of conventional automatic detection processes was not applicable. Moreover, unplanned and very complex situation in some parts of big cities like Istanbul, let the automatic method for line and texture detection not useful for landuse studies. This is a disadvantage for time, accuracy and cost. It is widely known that automatic methods have very big advantages, if suitable for the aim and if applicable.

REFERENCES

Beard, K., Mackaness, W., 1993. Visual access to data quality in geographic information systems, *Cartographica,* 30-2,3, 37-45.

Buttenfield, B.P., 1993. Representing Data Quality, *Cartographica,* 30-2,3, 1-7.

Celikoyan, T.M., 2004. Mointoring and Analysis of Landuse Changes in Historical Periods for the City of Istanbul by means of Aerial Photography and Satellite Imagery, , *PhD Thesis,* ITU Institute for Science and Technology, Istanbul.

Glossary of Cartographic Terms, www.lib.utexas.edu/Libs/PCL/Map_collection/glossary.html, accessed on 15.06.2003.

ISO/TC 211 N 1034, 2001. Text for DIS 19113, Geographic information – Quality principles, as sent to ISO Central Secretariat for issuing as Draft International Standard, International Organization for Standardization, Oslo, Norway.

Morrison, J., 2002. Spatial data quality, in Spatial databases : with applications to GIS pp. 500-516, Rigaux, P. Scholl, M. and Voisard, A., Morgan Kaufmann Publishers, San Francisco.

Radwan, M.M., Onchaga, R., Morales, J., 2001. A structureal approach to the amangement and optimisation of geoinformation processes*, European Organisation for Experimental Photogrammetric Research*, Official Publication No:41, Frankfurt am Main, Germany.

Taştan, H., 1999. Data quality in geographic information systems, *PhD Thesis,* ITU Institute for Science and Technology, Istanbul. (in Turkish).