Digitizing - Time Estimation Using A Linear Regression Model

by :
Mohamad Nor Bin Said
Ghazali Bin Desa
Center For Geographic Information & Analysis (CGIA)
Fakulti Ukur

Abstract

From data perspective, GIS can be regarded as having four subsystems namely data input, data storage and management, data manipulation and analysis and data output. Focusing on data input, manual digitizing has become the commonest practice but there is no standard way of quantitatively estimating the consumed time. This paper reports the results of an experiment to relate digitizing time to the total polygon perimeter contained in a map. A linear model was proposed and a linear regression technique was applied to test the model and to investigate the extent of its validity.

1.0 Introduction

Data input is one of the primary functions in any Geographic Information System (GIS). Inputting spatial data can be done via a number of ways such as manual digitizing, automatic scanning, entry of coordinates using coordinate geometry (COGO) and conversion from previously automated information but the first could be considered as the commonest in practice (Dangermond, 1990). Among the advantages that attract GIS users to consider manual digitizing are : low capital cost, low-cost labor and great flexibility and adaptability. While it is a time-consuming procedure, the technique can be taught to users within hours, and with modern database error checking software, the quality of information can be quite high. Interactive entry and editing can be done while users work on cartographic data; errors on the basic map can be easily discovered and updated while in the process of entering the information; and digitizing devices are very reliable. For these reasons, the majority of actual cartographic data entry is still performed using manual digitizing.

Manual digitizing involves the use of an electromagnetic, electrostatic or electromechanical device called a 'digitizer'. Typically a map is affixed to a tablet or table underlain with a current-carrying wire grid. A 'point locator' or better known as a cursor is moved along points or lines defining features on maps. These movements are converted into electrically identified locations which are read directly into computer (Ashdown and Schaller, 1990; Cimon et. al., 1990).

Projecting the workload associated with digitizing is an important step in formulating GIS projects. Time estimation is always required especially for budget calculation. A classical approach commonly used to provide this estimate is by taking the average measured time for digitizing a sample of maps, charts or photographs typical of the project at hand. However Cimon et. al. (1990) have introduced a more rigorous approach by using a linear regression model. In their study, a functional linear model can be developed using the relationship between total perimeter length of polygons contained in a map and digitizing time. The relationship is modeled as

$$t_i = a + mp_i \quad \ldots\ldots 1$$

where;

t_i = digitizing time
p_i = total perimeter length
a = a constant
m = the time-perimeter gradient

The quantity **p** can be obtained from the following relationship :

$$p_i = n_i \sqrt{A_i / n_i} \quad \ldots\ldots 2$$

or

$$p_i = \sqrt{A_i n_i} \quad \ldots\ldots 3$$

where ;

A_i = total area of polygons on map i
n_i = total number of polygons on map i

Although Cimon et. al., introduce a rigorous approach to determining digitizing time, their study has not been appropriately accepted because the relationship between digitizing time and total perimeter length may not necessarily be linear in nature. Experiments have been conducted at the Center for Geographic Information & Analysis (CGIA), Fakulti Ukur, University Teknologi Malaysia to test the model and to investigate the extent of its validity. Therefore, the purpose of this paper is to report the results of the experiment.

2.0 Method

The variables for the model are the digitizing time (t) and the total perimeter length (p). By taking redundant observations for t and p, the two coefficients can be solved using linear regression method. Six map samples, at the same scale of 1 : 500, were used. Each sample map was deliberately drawn to contain irregular polygon features. The number of polygons (n) were then counted. The total area of polygons (A) can be estimated in a number of ways. In this study, the 'counting squares' method was used. Knowing A and n, the total perimeter length of polygons p can be determined either by using equation 1 or 2 above.

The digitizing time for each map has to be considered as the result of contribution from six related processes : registration, transformation, line tracing, labeling, error checking and correction and topological formation (see Table 1).All of these processes need to be performed using a GIS software system. In this study, a Pc ARC/INFO was chosen because of its availability at the Center for Geographic Information & Analysis.

The two coefficients, i.e. the gradient m and the constant a, will provide initial indication with regard to the validity of the model. For any situation, a should always give a positive value because it indicates the time for registration and transformation. And it is independent of the total perimeter length of polygons in a map. The gradient m, on the other hand, indicates the relationship between digitizing time and the total perimeter length. Some of the processes that contribute to digitizing time may make the model behaving non-linearly. In this study, the processes for error checking and correction and topological formation were regarded as those that may contribute to this behavior.

Once a model is obtained, it can be used to calculate offsets (v) which is the difference between the calculated and observed digitizing time. In order to have these offsets determined, four more maps with different number of polygons and total area size are digitized. Two tests are performed : (1) using model containing all digitizing processes and (2) using the model without the time for error checking and correction and topological formation.

3.0 Results And Analysis

3.1 Test Using All Time Components

Table 2 shows the data obtained from the experiment. Based on these data, the two coefficients are determined. Figure 2 shows that the gradient (m) is found to be 0.001691 while the constant (a) is -0.041. The statistical test gives the R-squared value of 0.638. The R-squared value is very low, i.e. only 63.8% of the observations fall into the same population. This may be due to some inputs to the total observed digitizing time that have contributed a random behavior to the functional model. Beside the low R-squared value, the resulting constant (a) is found to be negative. This should not be accepted as the time for registration and transformation will never be having a negative value.

Figure 2 shows that there are some points having large variation from the gradient. From Table 2, a conclusion can be made that these variations are caused by too much time have been taken for the processes of checking and correcting errors. Therefore, these (error checking and correction) need to be treated independently so that the model may remain linear.

The offsets obtained by making comparison between the calculated and observed digitizing time is shown in Table 3. They vary from 13.2% to 44.1% and these may be considered unacceptable. The reasons for the large offset is mainly due to the non-linear behavior of the model.

Process	*Activity*
Registration	This includes fixing the map on the tablet and registering the control (Tic) points. A minimum of four points are used, i.e each map is registered using four control points throughout the test.
Transformation	The time observed for transformation process is only for the computation to convert digitizer's coordinates into the ground coordinates. This does not include the time for computing and preparing the chosen control points. It is assumed that these values are readily available. The Root Mean Square (RMS) is always noted to ensure the transformed coordinates to be within an acceptable accuracy.
Line Tracing	An arc-node tracing is adopted to digitize the whole polygons on each map. Nodes are defined at every intersection to ensure easier future error corrections. For an arc, the tracing is started at a node followed by recording the points along the arc (vertices) and ends at the other node of the arc. The interval between vertices are varied according to the irregularity of the arcs, i.e more points are captured at curved arcs compared to ones which are more linear. Since the software system is fully topological, the common boundaries are only digitized once.
Labelling	The labelling process is only to register label points to each polygon to enable them to be linked to their attributes file. For the entire observation, each polygon is given different identification numbers and is done in auto-increment mode.
Error Checking And Correction	The errors are checked right after each digitizing process (for each map). The observations include the most common ones such as node matching (undershoots and overshoots), extra and missing label points as well double lines at common boundaries (Figure 1). The corrections, on the other hand are made either by : a) moving the node to the intersection points, for both undershoots and overshoots. b) splitting the arc at the intersection point, for overshoots problem c) removing the whole arc in error and carry out re-digitizing d) removing extra label points and adding the missing ones, if there are any e) cleaning the entire map coverage by assigning a new error-cleaning (fuzzy) tolerance but this had to be carefully done because the process may degrade the positional accuracy and may also introduce new errors.
Building Topology	The topological relationship between polygons are created by the software system, i.e after all the errors are removed

Table 1

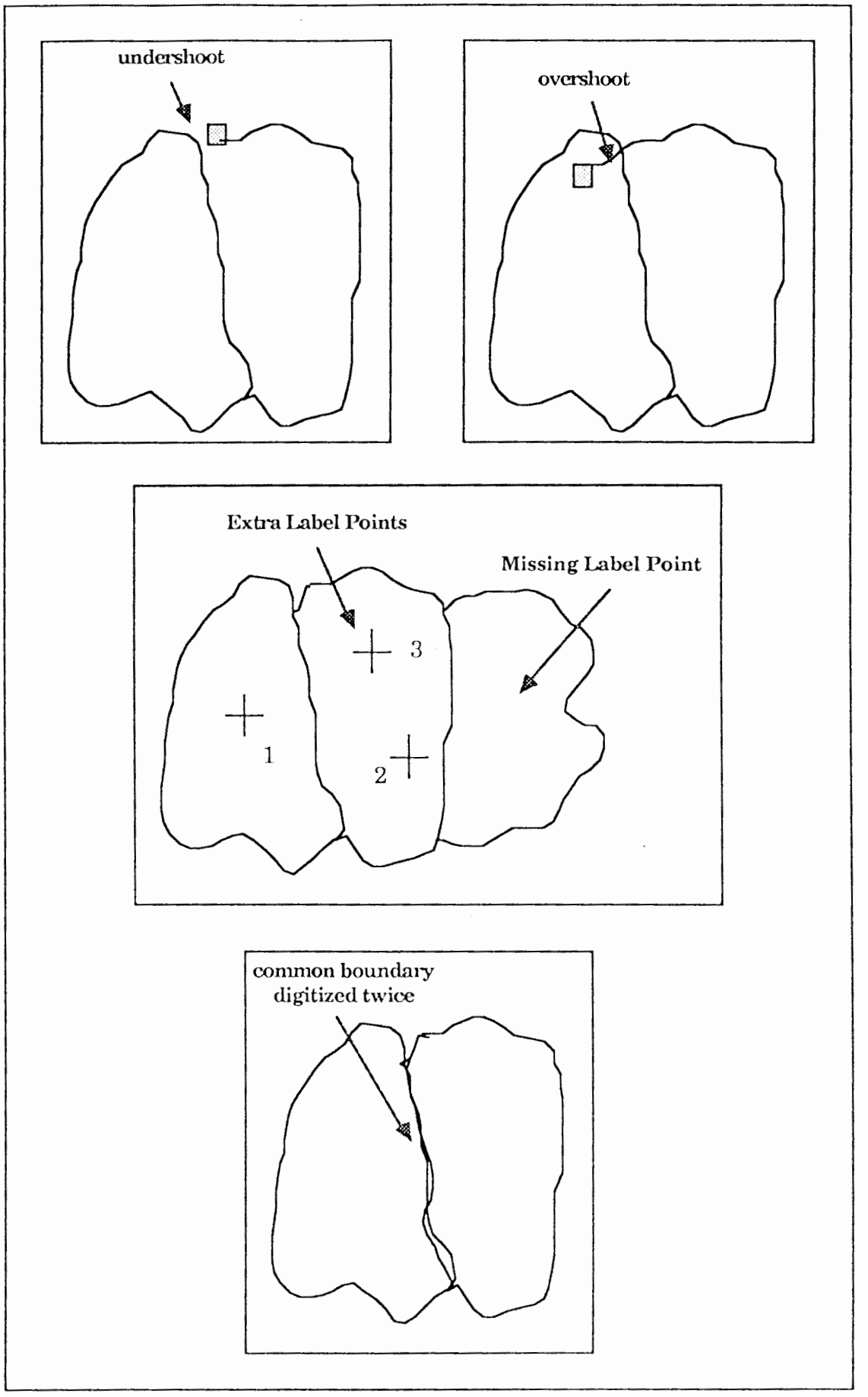
undershoot
overshoot
Extra Label Points
Missing Label Point
1
2
3
common boundary
digitized twice

Figure 1

3.2 Test By Excluding The Processes For Error Checking and Correction And Topology Construction

Table 4 shows the data without the time for the processes of error checking and correction and topological formation. When the digitizing time is plotted against the total perimeter length, the new gradient is 0.00083 and the constant is 0.078 (Figure 3). Now the R-squared value is much higher (0.936) indicating that the resulting model is 93.6% reliable. Table 5 shows the offsets by making comparison between the calculated digitizing time (based on the new model) and the observed one. The offsets show only a small variation, i.e. between 0.5% to 17.3%. This is much more acceptable compared with the first test made earlier.

4.0 Discussion

The regressed equation obtained in this test is only valid to digitize polygons of irregular shape. A different method or analysis has to be made for polygons of linear boundaries such as cadastral parcel, building outline, etc. because the boundaries of this type of polygons can be defined by digitizing using node-to-node technique. This is obviously faster compared to those of irregular shapes where every vertices along the arc have to be picked during the digitizing process.

Furthermore, the following conditions should be met when applying this technique :

a) The coefficients as obtained in the above test were valid for map of scale 1 : 500 only. A different regression has to be made for maps of other scales.

b) Again the above regression results could only be applied to estimate digitizing time for processes other than the error checking and correction and the topological formation.

b) The digitizing operator should be well trained with the digitizing software used.

It has also to be noted that, the digitizing operations depends on some other factors such as :

a) The software capability, for example a software without topological features will need the digitizing to be done in different way which will probably take a longer time to digitize the same map.

b) The speed of hardware used also contribute certain offset in time. For example, the above test was carried out using a 386-PC which is obviously slower if done on a workstation or a mainframe machine.

c) The skill and experience of the operator is another factor that determines not only the speed but the amount of errors introduced throughout the process. Furthermore

Map ID	Perimeter			Time Observation						
	Number of Poly-gons (n)	Total Area (A) (m)	*Total Peri-meter(p) (m)*	Regis-tration (hour)	Trans-formation (hour)	Line Tracing (hour)	Label-ling (hour)	Error check & correc-tion (hour)	Building topology (hour)	*Total Time (hour)*
IP10	10	2132	146.01	0.0183	0.0550	0.1453	0.0056	0.0964	0.0144	0.3350
IP15	15	5652	291.17	0.0186	0.0522	0.2375	0.0058	0.1558	0.0211	0.4853
IP20	20	11311	475.63	0.0181	0.0503	0.3583	0.0097	0.1300	0.0153	0.5836
IP30	30	12984	624.12	0.0194	0.0544	0.5389	0.0136	0.2578	0.0172	0.9022
IP35	35	12217	653.91	0.0189	0.0472	0.4650	0.0136	0.1425	0.0181	0.7056
IP40	40	13716	740.70	0.0198	0.0481	0.6675	0.0192	0.9247	0.0183	1.7003

Table 2

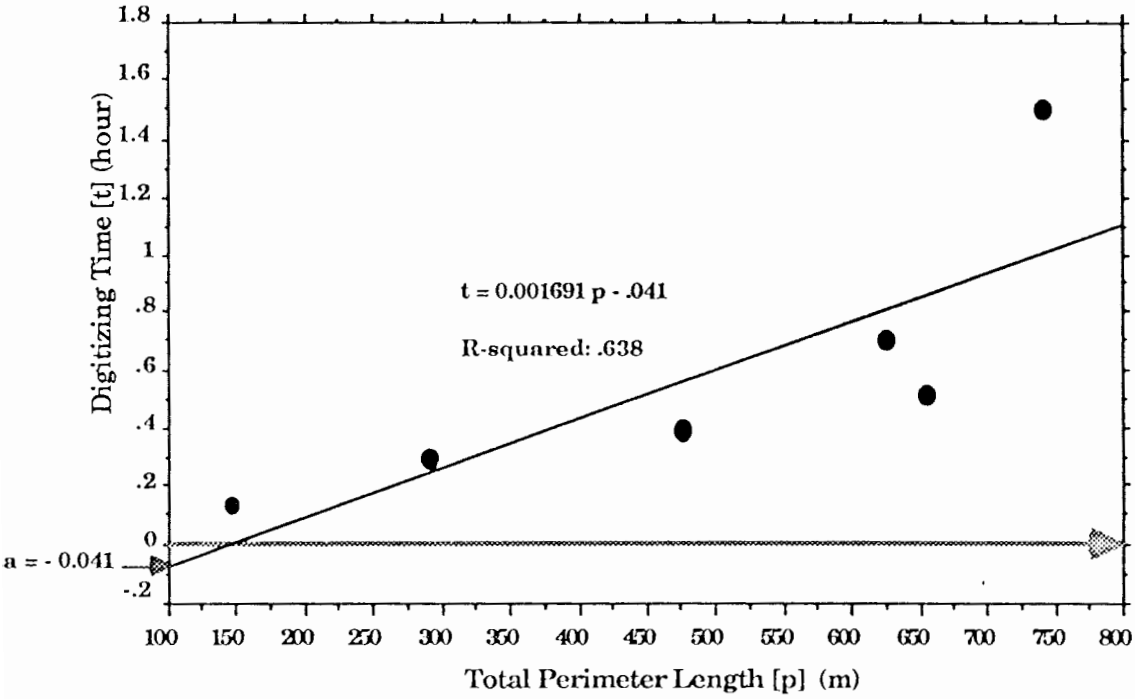

Figure 2

Map ID	Perimeter Estimate			Digitizing Time			
	Number of Poly-gons (n)	Total Area (A) m	*Total Peri-meter (p) m*	Calculated t = 0.001691 p - 0.041 (hour)	Observed (t) (hour)	Offset (v) (hour)	% Offset = (v/ t x 100)
TM10	10	12377	351.81	0.5539	0.4731	0.0808	17.1
TM15	15	10711	400.83	0.6367	0.4494	0.1872	41.7
TM20	20	8336	408.31	0.6494	0.5739	0.0756	13.2
TM25	25	10815	519.98	0.8383	0.5817	0.2567	44.1

Table 3

Map ID	Perimeter Estimate			Time Observation				
	Number of Polygons(n)	Total Area(A) (m)	*Total Perimeter(p) (m)*	Regis-tration (hour)	Trans-formation (hour)	Line Tracing (hour)	Labelling (hour)	*Total Time(t) (hour)*
IP10	10	2131	145.98	0.0183	0.0550	0.1453	0.0056	0.2242
IP15	15	5652	291.17	0.0186	0.0522	0.2375	0.0058	0.3142
IP20	20	11311	475.63	0.0181	0.0503	0.3583	0.0097	0.4364
IP30	30	12984	624.12	0.0194	0.0544	0.5389	0.0111	0.6264
IP35	35	12217	653.91	0.0189	0.0472	0.4650	0.0136	0.5447
IP40	40	13716	740.70	0.0197	0.0481	0.6675	0.0192	0.7628

Table 4

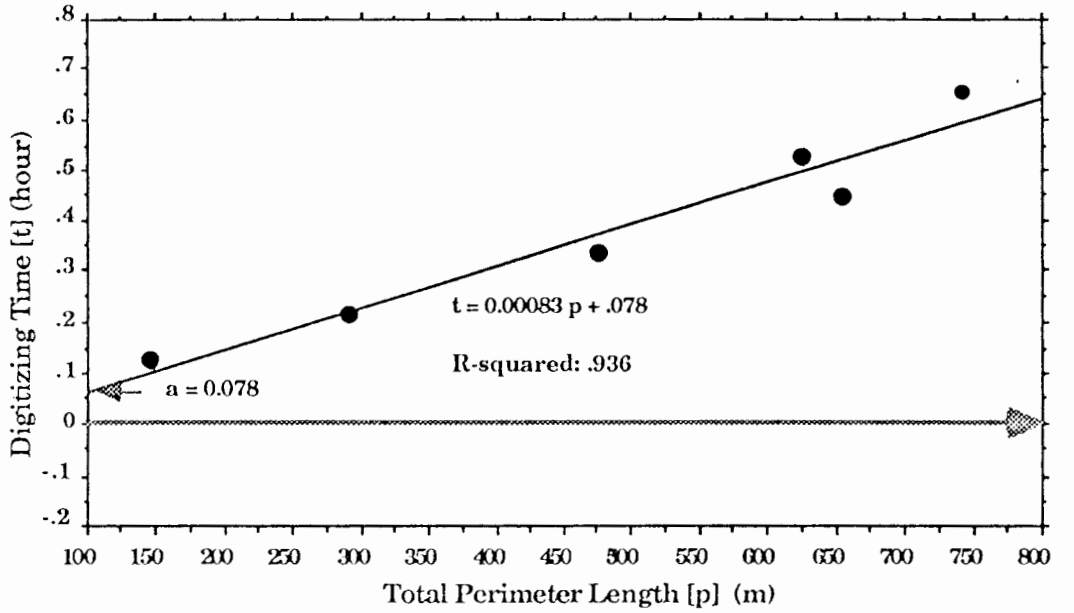

Figure 3

Map ID	Perimeter Estimate			Digitizing Time			
	Number of Poly-gons(n)	Total Area(A) (m)	*Total Peri-meter(p) (m)*	Calculated t = 0.00083p + 0.0777 (hour)	Observed (t) (hour)	Offset (v) (hour)	% Offset = (v/ t x 100)
TM10	10	12377	351.81	0.3697	0.3289	0.0408	12.4
TM15	15	10711	400.83	0.4108	0 .3608	0.0494	13.7
TM20	20	8336	408.31	0.4106	0.4264	0.0017	0.5
TM25	25	10815	519.98	0.4844	0.5092	0.0247	4.9

Table 5

the most important is the level of awareness of the spatial accuracy of the digital map being created. Some operators never care about what transformation's RMS means or what fuzzy tolerance indicates. For example, a larger fuzzy tolerance is given during the error cleaning process for the sake of speeding up the corrections without considering the casualty of ground accuracy.

5.0 Conclusion And Recommendation

Based on the results of this test, a conclusion was made that the linear model to estimate time as a function of total polygon perimeter is applicable to certain extent. This includes the time for map registration, transformation, line tracing and labeling operations. The occurrence of errors during digitizing process varies randomly no matter how many polygons contained in a coverage. Therefore the checking and removal time have to be treated separately from the other parts. However the model could probably be modified by adding some more terms or totally changed if the whole process is to be estimated using one single function.

Acknowledgment The author is very grateful to Dr. Azahari Bin Hussein for providing statistical software package used in this study.

References

Cimon,N. , Coe, P.K. and Quigley, T.M., 1990. A Regression Technique For Estimating The Time Required To Digitize Maps Manually. *International Journal of Geographical Information Systems, Vol. 4 Number 1 (1990).*

Dangermond, J., 1990. A Review Of Digital Data Commonly Available And Some Of The Practical Problems Of Entering Them Into A GIS. *Introductory Readings In Geographic Information Systems (Taylor & Francis), 1990.*

Ashdown, M. and Schaller, J., 1990. Geographic Information Systems And Their Application In MAB Projects, Ecosystem Research And Environmental Monitoring. *UNESCO - Man And The Biosphere Program, 1990.*

[redacted]
Law Lecturer
M. C. Law (I.I.U.)
Advanced D.I.L. (ITM)