MINING LABOUR MARKET INFORMATION FOR USE IN DEVELOPING AN IT MANPOWER PREDICTIVE MODEL

Gar-yun Garry Suen
Information Technology Training and Development Centre
Vocational Training Council, Hong Kong

Index: Data Mining, Artificial Neural Network, Statistical and Mathematical Technique, Labour Market Information, Manpower Predictive Model

Abstract: To improve the information content of the data and to empower knowledge workers of today and tomorrow, the latest "hot" technologies that have emerged on the client/server arena are focused on filtering unnecessary data and presenting the valuable information in a user-friendly, intuitive, and easy to understand way. One of these technologies is data mining. This paper will highlight the significant of discovering meaningful new correlations, patterns, and trends by digging into (mining) large amounts of labour market information (LMI) stored in databases, using artificial-intelligence (AI) and statistical and mathematical techniques. For experimental purpose, a manpower predictive model will be built upon the LMI to forecast the manpower requirement of the IT sector. The IT manpower predictive model is essential for planners in business, government and academia to derive an appropriate vocational education and training plan to meet the needs of the industry.

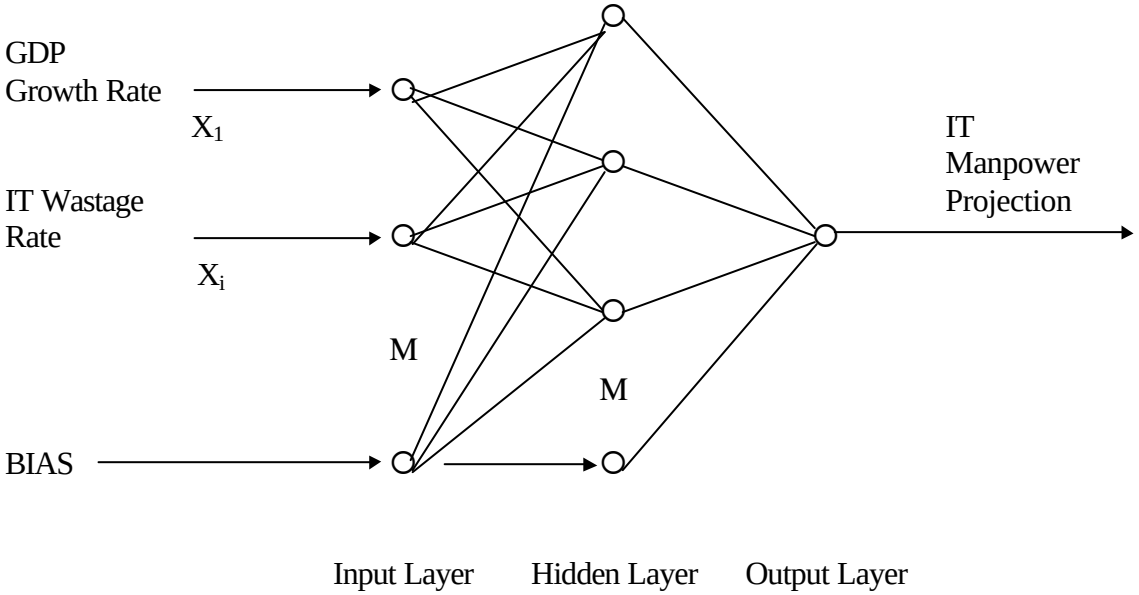
INTRODUCTION

How does data mining work? It isn't magic. Instead it works the same way a human being does. It uses historical information (experience) to learn. However, in order for the data mining technology to pull the "gold" out of your database, you do have to tell it what the gold looks like (i.e., what business problem you would like to solve). It then uses the description of that "gold" to look for similar examples in the database and uses these pieces of information from the past to develop a predictive model of what will happen in the future.

As IT is a rapidly changing and developing field, an accurate manpower predictive model is required for vocational training and educational planning purposes. Since it takes a number of years to educate and train people for the IT jobs, the industry needs must be anticipated sufficiently far in advance to allow time for that training. Failure to anticipate the needs and to develop the required high level IT manpower may seriously impede economic progress.

This paper demonstrates on how to develop a manpower predictive model using feed-forward artificial neural network (ANN) (as shown in Figure A) by going through the process of data analysis, parameters setting, and model training and testing. The historical information (such as IT vacancy rate, IT wastage rate, unemployment rate, emigrant growth rate, GDP growth rate, wages/earnings growth rate, consumer price index, and labour force participation rate) which might have an influence on the IT manpower requirement will be used for developing the predictive model.

Figure A Feed-Forward Artificial Neural Network



ARTIFICIAL NEURAL NETWORKS (ANNs) MODEL

ANNs are often represented as automated data mining techniques. While they are very powerful at building predictive models, they do require significant data preprocessing and a good understanding and definition of the predictive target. Usually normalizing predictor values between 0.0 and 1.0 and converting categorical to numeric values is required. The networks themselves also require the setting of numerous parameters that determine how the neural network is to be constructed (e.g., the number of hidden nodes).

An effective modeling tool (that discovers relationships from a database of examples) can generate a compact module that captures from simple linear to complex, non-linear relationships in a form that is readily executable. It excels at modeling systems such as fuzzy estimation, probabilities, expert experience and judgment, process sensor measurements and control strategies.

Experiments

Setting

The following procedures will guide us through a demonstration of modeling tool's capabilities on creating a data matrix of historical LMI and a number of predictive models based on the initial setting of training parameters.

Entering the Dataset

In order to create an ANN model, it is necessary to load the dataset into the modeling tool. Importing is one method to fill in the data matrix (ODBC.DM) required to build the model.

Factors/variables displayed as columns are presented in a spreadsheet format with socio-economic data collected by the Census and Statistics Department and the IT manpower data collected by the Vocational Training Council, and each observation from 1985 to 1997 in quarterly basis is displayed as row. The intent of the data capture is to see if any of the process factors/variables and performance indicators of the data matrix can be used appropriately for the development of manpower predictive model of the IT sector.

The parent data matrix can be apportioned into training and test matrices. The dataset from 1985 to 1994 (with total number of observations N = 40) is used to train the ANN model. Other portion of parent data matrix from 1995 to 1997 (with total number of observations N = 12) is reserved for testing the accuracy of the trained model by considering the difference between measured and predicted outputs or the model error. The column (factor/variable) names of the data matrix are described in Table A.

Table A Factors/Variables of the Data Matrix (ODBC.DM)

Column (Factor/Variable) Name of the Data Matrix	Column (Factor/Variable) Description
IT/LABR	IT manpower as a % of total labour force
IT-GRW	IT manpower growth rate
IT-WTG	IT manpower wastage rate
IT-VAC	IT manpower vacancy rate
IT/ESTAB	Average No. of IT manpower per establishment
PC-GRW	Terminals/PCs/Workstations/Servers growth rate
SYS-GRW	Computer System (mainframes, mini-computers, super-micros) growth rate
NET-INST	% rate of installation with network
GDP-GRW	GDP growth rate
IMP-GRW	Imports quarterly quantum index growth rate
EXP-GRW	Domestic exports quarterly quantum index growth rate
REEXP-GR	Re-exports quarterly quantum index growth rate
TOTEXP-G	Total exports quarterly quantum index growth rate
LABR-GRW	Labour force growth rate
EMP-GRW	Total employment growth rate
UMEMPLOY	Unemployment rate
UNDEREMP	Under employment rate
MANU-EAN	Earnings growth rate by manufacturing sector
SERV-EAN	Earnings growth rate by services sector
CPI(A)GR	Consumer Price Index (A) growth rate
CPI(B)GR	Consumer Price Index (B) growth rate
CPI(C)GR	Consumer Price Index (C) growth rate
CCPI-GR	Composite Consumer Price Index growth rate
EMIG-GRW	Emigrant growth rate
TOTAL-MP	Total IT manpower requirement

Data Analysis

After the dataset has been loaded into the modeling tool, the data matrix is checked to see that only good data has been imported. Outliners, typing mistakes and scaling problems may be seen statistically or graphically as shown below.

Basic Statistics

Statistics report which gives an overall picture of the data matrix (ODBC.DM) is shown in Table B.

Table B Basic Statistics of the ODBC.DM

Factor/Variable	N	Mean	Std Dev	Minimum	Maximum	Sum Sq
IT/LABR	52	0.013154	0.004633	0.006000	0.020000	0.684000
IT-GRW	52	0.134923	0.107106	0.042000	0.371000	7.016000
IT-WTG	52	0.090058	0.045007	0.039000	0.178000	4.683000
IT-VAC	52	0.066192	0.014343	0.045000	0.086000	3.442000
IT/ESTAB	52	0.261538	0.156974	0.070000	0.590000	13.600000
PC-GRW	52	0.581308	0.543715	0.025000	2.605000	30.228000
SYS-GRW	52	0.087827	0.023062	0.062000	0.134000	4.567000
NET-INST	52	0.461500	0.175080	0.220000	0.693000	23.998000
GDP-GRW	52	0.057269	0.040625	-0.029000	0.177000	2.978000
IMP-GRW	52	0.149058	0.096678	-0.041000	0.362000	7.751000
EXP-GRW	52	0.024019	0.093322	-0.103000	0.270000	1.249000
REEXP-GR	52	0.219212	0.141061	-0.055000	0.500000	11.399000
TOTEXP-G	52	0.143808	0.095161	-0.025000	0.371000	7.478000
LABR-GRW	52	0.016404	0.015731	-0.015000	0.059000	0.853000
EMP-GRW	52	0.017769	0.016794	-0.012000	0.062000	0.924000
UNEMPLOY	52	0.022500	0.006160	0.013000	0.035000	1.170000
UNDEREMP	52	0.014212	0.005207	0.006000	0.025000	0.739000
MANU-EAN	52	0.038481	0.036497	-0.057000	0.135000	2.001000
SERV-EAN	52	0.044481	0.024125	-0.020000	0.088000	2.313000
CPI(A)GR	52	0.074885	0.026806	0.025000	0.131000	3.894000
CPI(B)GR	52	0.075962	0.025667	0.029000	0.124000	3.950000
CPI(C)GR	52	0.083538	0.024770	0.034000	0.122000	4.344000
CCPI-GR	52	0.077500	0.025461	0.031000	0.123000	4.030000
EMIG-GRW	52	0.077788	0.230971	-0.300000	0.579000	4.045000
TOTAL-MP	52	25620.750	12186.646	9378.0000	44847.000	1332279.0

Model Building

In order to derive an appropriate ANN model for the IT manpower forecast, it is necessary to study a number of trained ANN models using different number of hidden neurons from 1 to 9. Other initial parameters controlling the way on how the modeling tool trained on the pattern data are set up as follows:

Max Hidden:	(from 1 to 19)	Learning Rate:	0.75	Seed: 15
Eon:	100	HLearning Rate:	0.75	
Max Training:	200,000	TLearning Rate	0.75	
Hidden Freeze:	0.75	IO Learning Rate	0.75	
Error Tolerance:	1.e-003	Alpha:	0.8	
Good RSQ:	0.9999	Theta:	0.5	
Sign Inc:	5.e-002	Random Fact:	0.5	
Tolerance:	5.e-002	Auto Save:	1,000	

Models Analysis

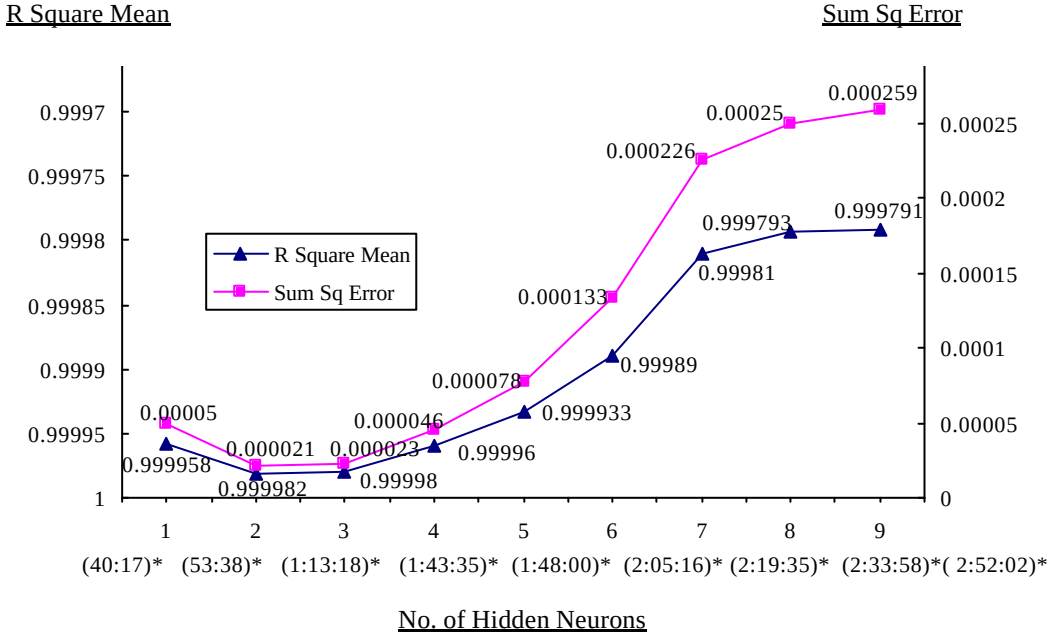
The application of the modeling tool provides two basic methods for analyzing trained models. The first is a numerical or statistical method and the second is a graphical method. Both of these methods can be used on either the training matrix or the test matrix. Also, additional test matrices can easily be imported for analyzing long-term performance.

ANN models have been created and trained on the basis of automatic increment of hidden neurons (by setting the maximum number of hidden neurons from 1 to 9). In this respect, it is necessary to find out the performance of each of these ANN models by including or excluding factors/variables from the input list of the model based on prior knowledge. The performance of all these ANN models are validated by checking the predicted outputs statistically and graphically as shown in the following experimental results:

Result I - Models Performance Analysis based on the Number of Hidden Neurons

Figure B below shows the performance analysis of 9 well-trained NN models. All these models each have the R square mean close to 1.0 and the sum square error close to zero. It is observed that the hidden neuron's learning rate decreases when adding a new neuron in the Automatic Increment training algorithm. This results in the increase of training time in accordance with the addition of hidden neurons as shown in figures in brackets.

Figure B Models Performance Analysis based on Number of Hidden Neurons



* Figures in brackets denote the time taken in hh:mm:ss to train the neural models after 200,000 times the training matrix has been presented before ending the training.

In comparison, it is observed that the model with 2 hidden neurons (with sum square error of 0.000021 and R square mean of 0.999982) performed better than the other models. In order to improve the performance of the ANN model with 2 hidden neurons for projecting IT manpower requirement, further analysis of its factors/variables for inclusion into the input list of the model is essential.

Result II - Sensitivity Analysis of the Output of the Trained Model (with 2 Hidden Neurons) on the Training and Test Matrices

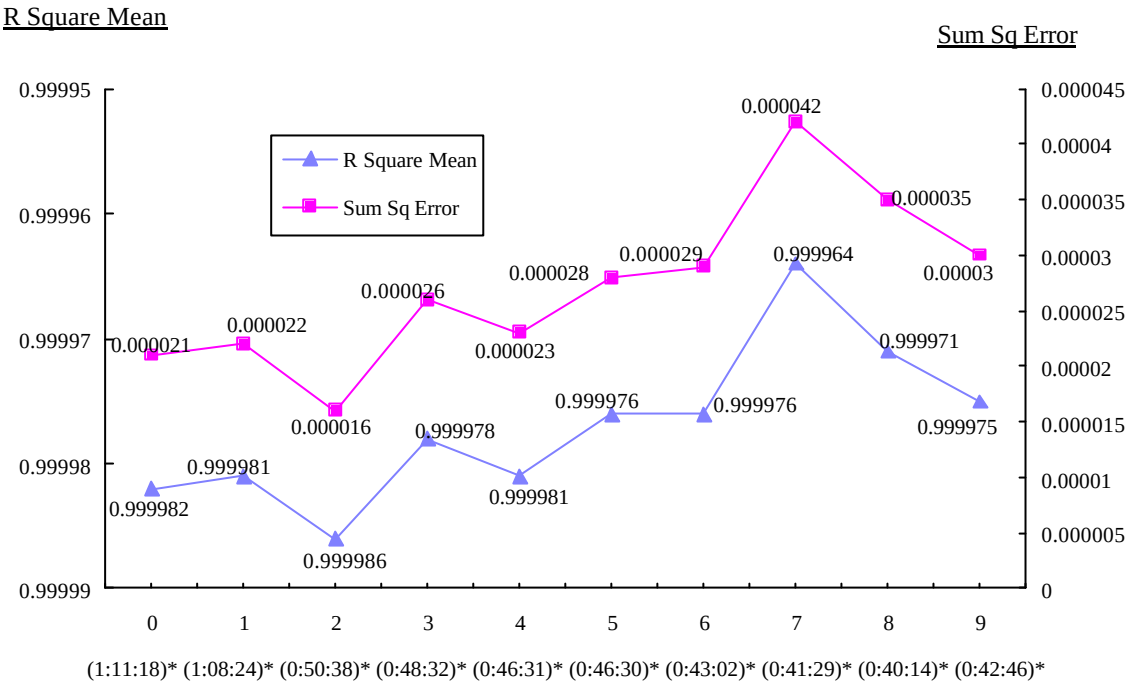
A sensitivity analysis is run to see which factors/variables account for most of the variability of total IT manpower requirement (TOTAL-MP) of the trained model with 2 hidden neurons on the training and test matrices. The result is ranked in order with factors/variables of the most effect at the top of the list as shown in Table C.

Table C Sensitivity Analysis of the Output (TOTAL-MP) of the Trained Model with 2 Hidden Neurons on the Training and Test Matrices

Variable Name	AveAbs Sens	Ave Sens	Peak Sens
NET-INST	0.18175	+0.18175	+0.08609
IT-WTG	0.18055	-0.18055	+0.08944
SYS-GRW	0.09439	-0.09439	+0.03850
IT-VAC	0.07946	+0.07946	+0.04372
CPI(B)GR	0.06373	+0.06373	+0.03104
CCPI-GR	0.05160	-0.05160	+0.03065
CPI(C)GR	0.04370	+0.04022	+0.03096
IT/LABR	0.03570	+0.03570	+0.01722
CPI(A)GR	0.03529	-0.03527	+0.02433
EMIG-GRW	0.03246	-0.03246	+0.02093
IT-GRW	0.02958	+0.02958	+0.01631
REEXP-GR	0.02605	-0.02589	+0.01792
GDP-GRW	0.02274	-0.02207	+0.01392
EMP-GRW	0.02226	-0.02226	+0.01152
EXP-GRW	0.01544	+0.00868	+0.01081
PC-GRW	0.01239	+0.01193	+0.00795
SERV-EAN	0.01218	-0.01218	+0.00698
LABR-GRW	0.01122	+0.01122	+0.00586
IMP-GRW	0.01089	+0.00346	+0.00820
MANU-EAN	0.01002	+0.00990	+0.00630
UNDEREMP	0.00887	-0.00283	+0.00739
TOTEXP-G	0.00771	+0.00754	+0.00667
IT/ESTAB	0.00743	+0.00387	+0.00649
UNEMPLOY	0.00460	-0.00460	+0.00219

In order to pick the best factors/variables for inclusion into the input list of the model with 2 hidden neurons, the last 9 factors/variables of the sensitivity analysis report are each removed one by one from the input list to form different models for further training and analysis. The performance of these 9 trained ANN models based on various input factors/variables on the training matrix is shown in Figure D.

Figure D **Models Performance (with 2 Hidden Neurons) based on Various Input Factors/Variables on the Training Matrix**



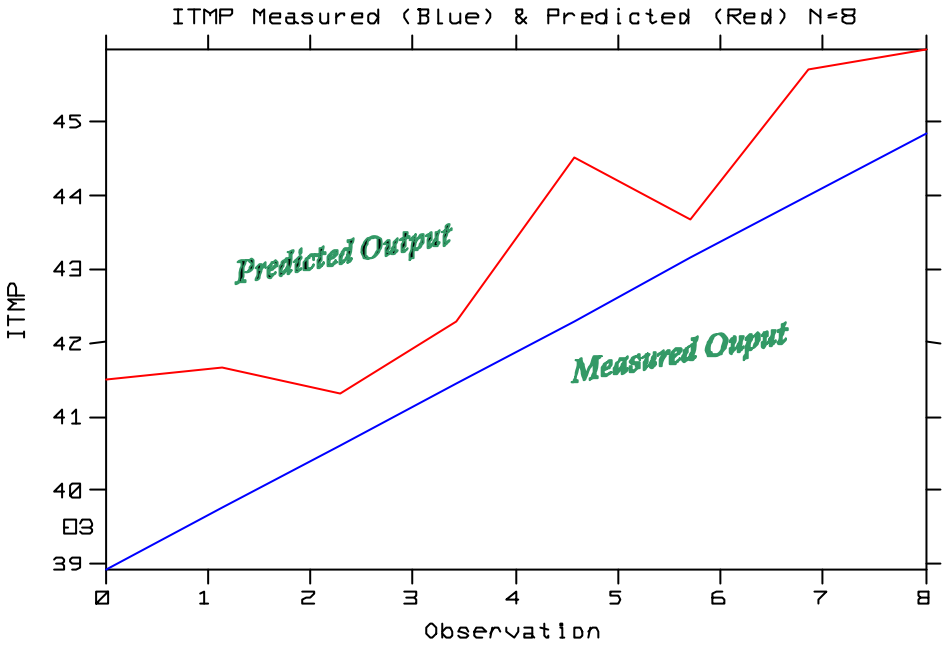
* Figures in brackets denote the time taken in hh:mm:ss to train the neural models after 200,000 times the training matrix has been presented before ending the training.

The resulting graphical display enable us to quickly determine how well all these trained models performed on the training matrix. It is observed that these trained models each have the sum square error close to zero and the R square mean statistic larger than 0.9999 (near 1.0) of the good RSQ training parameter. Among these, model with 2 factors/variables (i.e. IT/ESTAB and UNDEREMP) removed from the input list performed the best (with R square mean 0.999986 and sum square error 0.000016). It is also observed that in general the time taken to train each model is inversely proportional to the number of factors/variables removed from the input list of the model.

Result III – Keep Best ANN Model During Training

Sometimes the best model of a particular process develops somewhere between the first few seconds of training and the maximum epoch allowed. The modeling tool can keep this intermediate model developed during the training session as the final model by selecting either the mean square error or R square as the measurement. A measured and predicted graph with best performance on the test matrix is shown in Figure E.

Figure E Measured and Predicted Outputs of the Model (with 2 Hidden Neurons) on the Test Matrix having 2 Factors/Variables Removed



CONCLUSION

Findings of the Project

Feed forward neural networks with a single hidden layer are statistically consistent estimators of arbitrary measurable, regression functions under certain practically-satisfactorily assumptions regarding target noise, number of hidden units, size of weights, and form of hidden-unit activation function (White, 1990). Unfortunately, the above consistent results depend on one impractical assumption that the networks are trained by an error minimization technique that comes arbitrarily close to the global minimum. Such minimization is computationally intractable except in small or simple problems (Judd, 1990).

ANNs do provide powerful predictive models and theoretically are more general than other data mining and standard statistical techniques. In practice, however, the gains in accuracy over other techniques are often quite small and can be dwarfed by some of the costs because of careless construction or use of the model by non-experts.

Further Development of the ANN Model

When standard neural networks are used to solve problem in projecting IT manpower requirement which relied on complex labour market information, the deficiency of the network architectures become apparent. (i.e. Brute-force learning is time-consuming and the performance is usually poor.) It is suggested that if the problem is sufficiently partitioned by job level (such as IT management, systems development and operative), standard neural networks can be used as building blocks (or functional units) in the construction of a more complicated neural structure. Each of these functional units is responsible for solving a portion of the problem by job level. This, of course, will ease the learning process and improve the accuracy of the functional units. On the other hand, the partitioning scheme depends very much on the designer's knowledge of the problem concerned. These include identified appropriate learning factors/variables by job level for each of the corresponding functional units. Once such scheme is identified, the resulting neural network would consist of an intelligence that is in-built by the designer, and the output of

each functional unit would aggregate to form the total IT manpower requirement. This is the core idea of neural intelligence.

REFERENCES

Consultancy Study on the Manpower and Training Needs of the IT Sector, Education and Manpower Bureau, 2000.

Data Warehousing, Data Mining and OLAP, Alex Berson, Stephen Smith, McGraw-Hill, 1997.

Forecasting Methods & Applications, Spyros Makridakis & Steven C. Wheelwright, A Wiley/Hamilton, 1978.

Fundamentals of Neural Networks, Laurene Fausett, Prentice Hall, 1994.

Increased Rate of Convergence Through Learning Rate Adaptation, R A Jacobs, Neural Networks, Vol 1, No 4, 1988, pp 295 – 307.

Neural Networks Primer, Maureen Caudill, AI Expert, June 1988, pp 53 – 59.

1985 – 2000 Editions : Hong Kong in Figures, Census and Statistics Department.

1985 – 1998 Editions : Manpower Survey of the IT Sector, Vocational Training Council