TEMPORAL DATA MINING USING GENETIC ALGORITHM AND NEURAL NETWORK – A CASE STUDY OF AIR POLLUTANT FORECASTS

Shine-Wei Lin*, Chih-Hong Sun, and Chin-Han Chen*****

*National Hualien Teachers Colledge, **National Taiwan University, ***Isou University

shine@mail.nhltc.edu.tw

ABSTRACT

Artificial intelligence technology like neural network and genetic algorithm can easily cope with highly complicated and non-linear combined spatial and temporal issues. Therefore this paper integrated genetic algorithm and neural network techniques to build new temporal predicting analysis tools for Geographic Information System (GIS). These new GIS tools can be readily applied in a practical and appropriate manner in spatial and temporal research to patch the gaps in GIS data mining and knowledge discovery functions.

The specific achievement here is the integration of related artificial intelligent technologies into GIS software to establish a conceptual spatial and temporal analysis framework. And, by using this framework to develop an Artificial intelligent Spatial and temporal Information Analyst (ASIA) system which then is fully utilized in the existing GIS package so that it is convenient for the domain experts to work with it and apply it. This study of air pollutants forecasting provides a geographical practical case to prove the rationalization and justness of the conceptual temporal analysis framework.

Keywords: Geographic Information System (GIS), Temporal, Data Mining, Genetic Algorithm, And Neural Network

1. INTRODUCTION

Recently, the information science has been focusing its research in artificial intelligence on the developing of neural networks, fuzzy logic and genetic algorithms. Neural network research shows that mankind faces complicated issues in the aggregation learning method. By simulating mutually integrated neurons we can proceed with the learning behavior of mankind and find the relationships between input influence variables and output environmental related results (Muller et al., 1995). Through imitating the competion and selection process of a living creature, the genetic algorithms make a computer execute designed evolutional regulations. And, it can naturally adjust environmental order and structure and find the optimal solution (Scott, 1990). For this reason we can deal with numerous environmental effect factors in the face of real world spatial and temporal problems by combining neural network and genetic algorithms techniques. These intelligent technologies do not use a linear order to explain the system behavior and can properly integrate them with a GIS to make new spatial and temporal analysis models. These models can simultaneously handle a great deal of spatial and temporal information, and take into account complex relations between factors. Compared with traditional linear statistics models they can more realistically fit the prospective trend towards the complex and dynamic real world spatial and temporal issues. Not only can we free ourselves from searching for all the possible factors, but we also escape from the corner of the shortcomings of GIS spatial and temporal analysis (Frank, 2000; Gahegan, 2000; Openshaw and Openshaw, 1997).

It is because of the geographer who so anxiously expects to solve complex and dynamic spatial and temporal issues that the GIS began to integrate information science in the data mining and knowledge discovery research (Boots, 2000; Fischer, 1997; Leung and Leung, 1993; Marble, 2000). Therefore, at this level it is reasonable to use the computer to deal with geographical issues. The first goal of this research is to formalize the interaction between humans and their environment, and to build an integrated conceptual spatial and temporal analysis framework combining the power of GIS and information science.

Anselin (2000) stated that there are three essential requirements for a well designed GIS integrated information system, including providing a data format that can be convert into other GIS styles, designing reusable components in the 'Windows' programming environment, and having a visual interface platform. Therefore, the second goal in this research is to develop an Artificial intelligent Spatial and temporal Information Analyst (ASIA) package, including the design of a new data format for the converting with other GIS styles, using the C++ program to build an open access artificial intelligent object module, and use the ArcView GIS software for the visualization platform. This package integrates artificial intelligent technology and spatial and temporal data in commonly used GIS software environment. The third goal of this research is proceed with a time trend forecasts of air pollutants as a requirement for the model validation in order to prove the reasonableness and usefulness of the conceptual spatial and temporal analysis framework, and of the integrated artificial intelligent system.

2. CONCEPTUAL GIS SPATIAL AND TEMPORAL ANALYSIS ARCHITECTURE

This research built a conceptual GIS spatial and temporal analysis architecture that conclusively integrated the GIS and artificial intelligence, and the heuristic data mining technology. This conceptual architecture includes a spatial and temporal analysis development stage, an integrated information platform and operational research steps. See figure1.

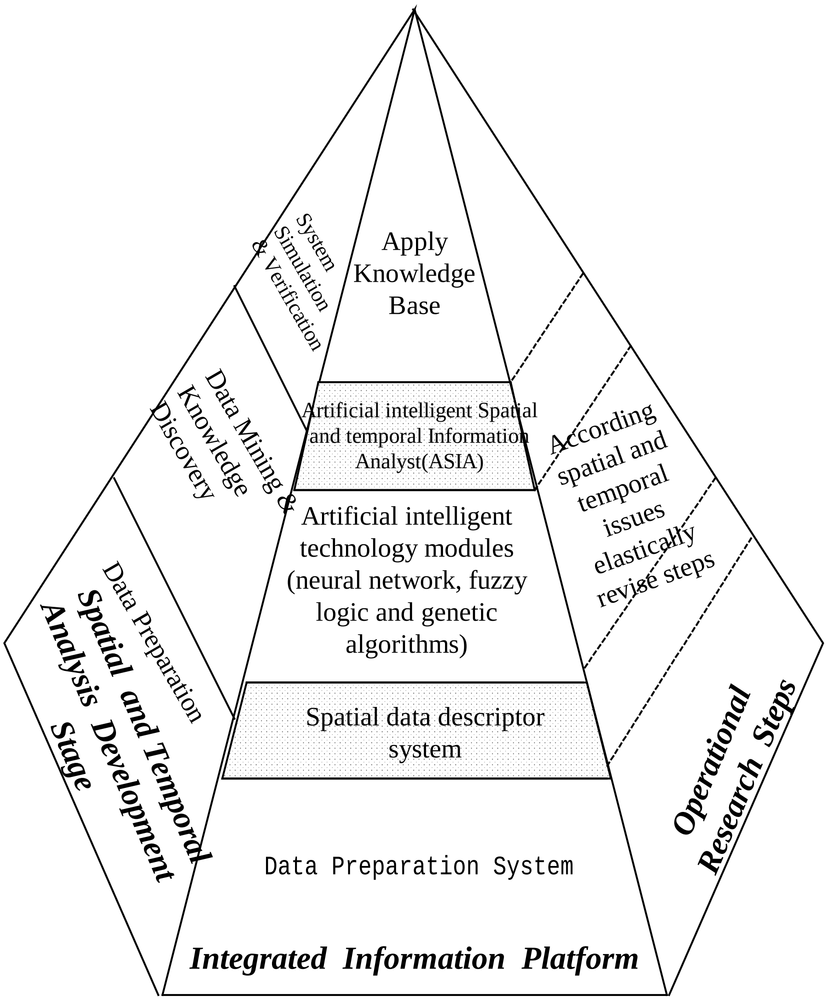

Apply Knowledge Base
Artificial intelligent Spatial and temporal Information Analyst(ASIA)
Artificial intelligent technology modules (neural network, fuzzy logic and genetic algorithms)
Spatial data descriptor system
Data Preparation System
Integrated Information Platform
System Simulation & Verification
Data Mining & Knowledge Discovery
Data Preparation
Spatial and Temporal Analysis Development Stage
According spatial and temporal issues elastically revise steps
Operational Research Steps

Figure1: Conceptual GIS spatial and temporal analysis architecture

The initial level in the ‘spatial and temporal analysis development stage’ is data preparation, which integrates three types of GIS database research methods, including filtering data noise, obtaining hidden information, and building a formalized data format. The kernelled second levels are data mining and knowledge discovery, which are based on artificial intelligent technologies making a connecting neuron network or ruling fuzzy logic knowledge model with genetic algorithms. And, the third and final level are system simulation and verification which use ASIA, the new ArcView GIS extension, to apply the discovered knowledge model allowing the geographers to master complicated spatial and temporal issues.

There are several standalone but mutually joined information systems in an ‘integrated information platform’. The data preparation system integrates the GIS basic ability for display, filtering, extracting, and transformation. The spatial data descriptor system is an information extraction interface between the AI technology and the GIS data, which can appropriately import newly, formalized GIS data into the neural network, fuzzy logic and genetic algorithms, and produce environment related knowledge. As for the ASIA, the ArcView GIS extension, is an AI based GIS knowledge discovery package, and can let geographers apply this knowledge in a practical manner when facing various spatial and temporal issues in spatial analysis researches.

The operational research steps are the final phase in the conceptual GIS spatial and temporal analysis architecture. It can reorganize, regulate, or improve itself in a highly elastic manner when faced with different objective research topics

3. ARTIFICIAL INTELLIGENT SPATIAL AND TEMPORAL INFORMATION ANALYST (ASIA)

The ASIA is an ArcView GIS extension, and a concrete operational system used in previous conceptual GIS spatial and temporal analysis architecture. Users load the GIS map into ASIA, and select the appropriate spatial descriptors to extract the hidden information in the GIS map. This allows the genetic algorithms to choose from the descriptors and transform it into the new GIS2 data format. Finally, select the connecting neural networks knowledge model to explore the weighted matrix knowledge data.

3.1 GIS Data Prepreparation and Spatial Descriptors System

According to GIS spatial or other time-sequence data sources, ASIA can use the menu to perform the data prepreparation system, including features generalize from polygon or polyline, grid description, grid resampling, grid clip by selected polygon theme, grid mosaic, grid topology (transform grid to lattice format), grid linear contrast enhancement, grid covariance, and random split table. And the spatial descriptor system generates simple statistic vectors, rotation invariant topographical feature indexes, Evans surface trend indexes, and surface frequency fast Fourier transforms indexes. See figure2.

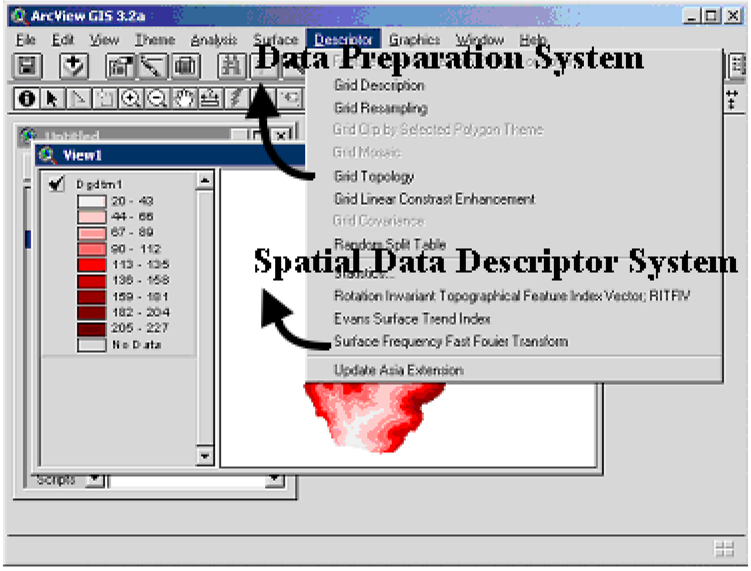


Figure2: ASIA data prepreparation and spatial descriptor system

3.2 GIS2 Format Transformation

The GIS2 data format is a basic data structure in ASIA and the joint connection between other AI spatial and temporal analysis models. It include two pure ASCII data format, as an intermediately file format, can easily communicate with much larger capacity GIS software. First data file is attribute-recording file that in the first line writes down variable and object record numbers and from the second line subsequently writes down every GRID cell values. See table1.

Table1: GIS2 attribute recording file

19	1	739596						
3313	54.6524	97.776	-0.219237	3.60014	-0.00351583	-0.0219008	…	0
3296	63.3855	65.9391	-0.534405	1.21776	-0.0102157	-0.00895094	…	0
…								

Second data file is coordinate and theme recording file that subsequently writes down the coordinate system records, GRID cell size, no data value, and every variable or object corresponding theme. See table2.

Table2: GIS2 coordinate and theme recording file

Number of columns or rows:	cols	862
	rows	848
The original of the coordinate (low and left corner)	xllcorner	260440
	yllcorner	2.67E+06
GRID cell size:	cellsize	40
No data value:	NODATA_value	-9999
Variable numbers:	input	19
Object numbers:	Output	1
Corresponding every variable in the attribute recording file's theme:	Elevation	
	Slope	
	…	
	Observation	

In ASIA the GIS2 format transformation makes through a dialog and includes three steps: First, use the polygon vector theme to assign the transformation boundary. Second, assign the environmental factors theme group (learning sources) and AI spatial and temporal model learning objectives. Third, assign the GIS2 file transformation path. See figure 3.

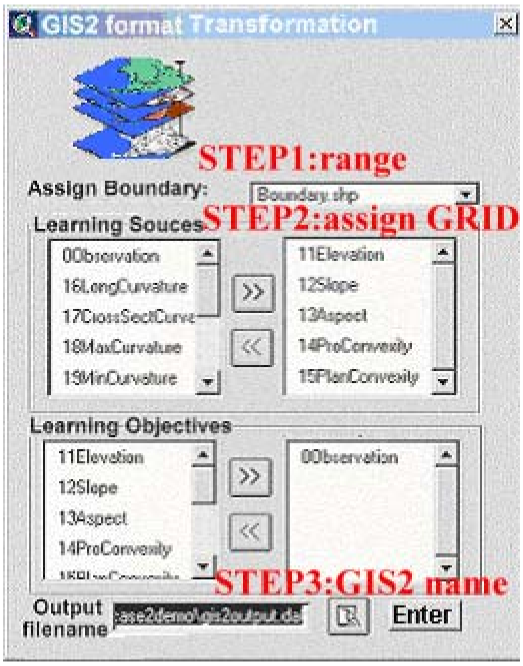
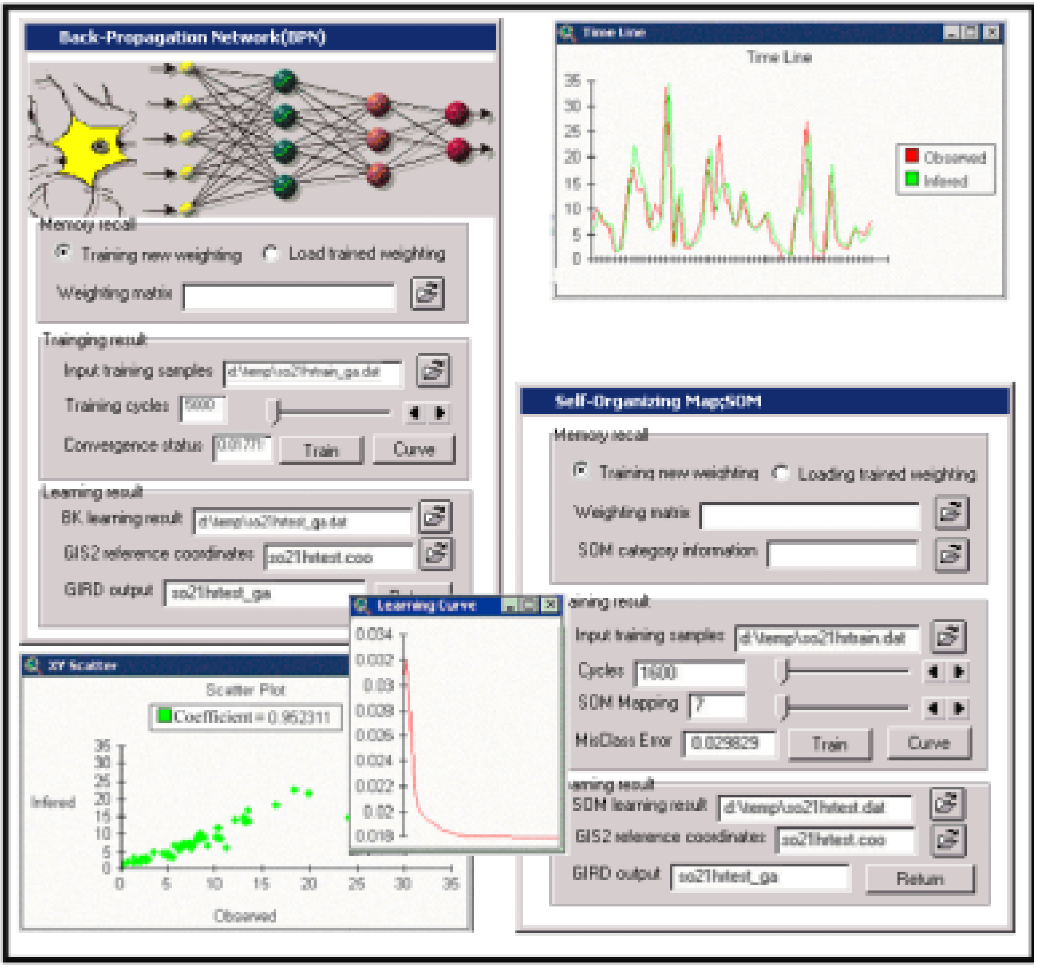


Figure3: Three steps in AISA GIS2 data format transformation

3.3 Supervised and Unsupervised Neural Network Spatial Analysis Model

In “Memory recall”, which is part of the ASIA neural network analysis model, the user should first of all decide whether to train a new or load a trained weighting matrix. If the user selects to train a new weighting matrix, then he should secondly decide the training cycles in BPN or additionally mapping information in the SOM model. In SOM model mapping information “3” means 3*3 nine categories. After selecting “Input training samples” from the GIS2 file source we can press “Train”, and survey the train result according to the “Learning Curve”. On the contrary, if the user selects to load a trained weighting matrix, it means the system had been trained before and can be recalled by assigned filename. The default filename in the back-propagation network (BPN) is “neurowgt.dat”, and in self-organizing maps (SOM) they are “somwgt.dat“ and “sommap.dat”. Thirdly, user should assign the verifying or simulating GIS2 file in order to put the weighting matrix of spatial and temporal knowledge in use in the part of “Learning result”. When faced with a temporal table format, ASIA can draw simple time line or XY scatter map or when faced with a spatial GRID theme format, ASIA can generate a new GRID theme in the result. See figure4.



Memory recall

● Training new weighting ○ Load trained weighting

Weighting matrix:

Trainging result

Input training samples: d:\temp\so21htrain_ga.dat

Training cycles: 5000

Convergence status: 0.01711

Learning result

BK learning result: d:\temp\so21htest_ga.dat

GIS2 reference coordinates: so21htest.coo

GIRD output: so21htest_ga

Memory recall

● Training new weighting ○ Loading trained weighting

Weighting matrix:

SOM category information:

aining result

Input training samples: d:\temp\so21htrain.dat

Cycles: 1600

SOM Mapping: 7

MisClass Error: 0.029829

arning result

SOM learning result: d:\temp\so21htest.dat

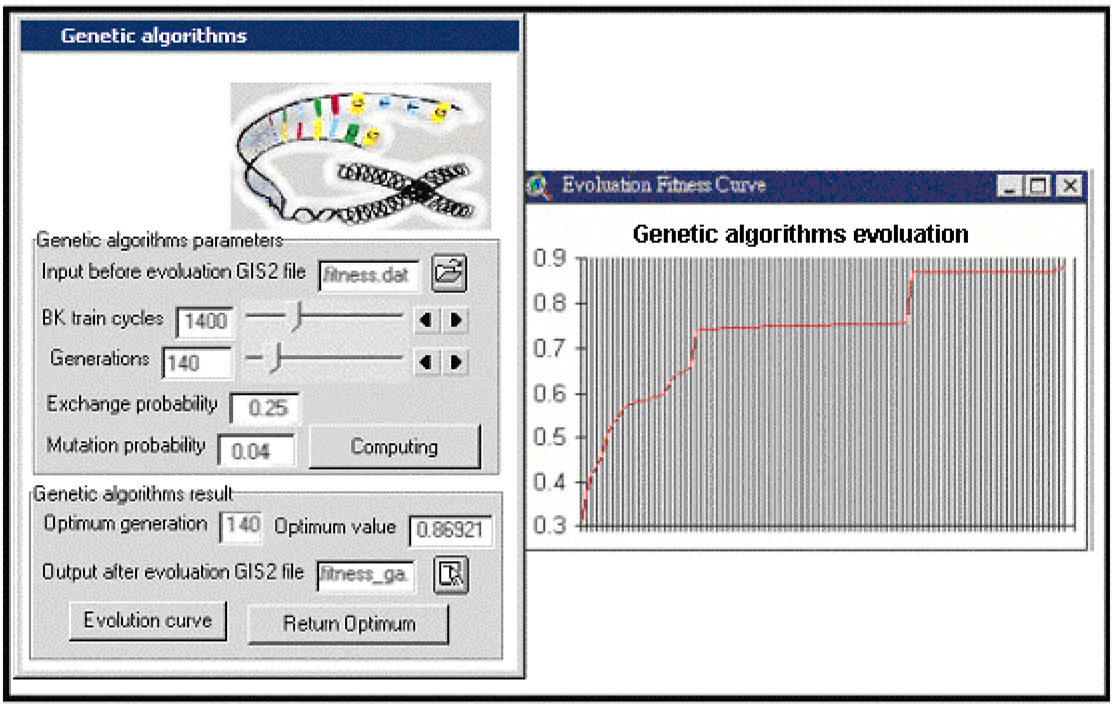
GIS2 reference coordinates: so21htest.coo

GIRD output: so21htest_ga

Figure4: Supervised and unsupervised neural network analysis model

3.4 Genetic Algorithms Optimum Spatial Analysis Model

The genetic algorithms can naturally discriminate which descriptors are important or duplicate from abundant collecting digital data and use the “0” or “1” series to represent proper descriptors. The ASIA optimum genetic algorithms analysis model integrates the neural network learning convergence to evaluate every generation’s propriety as the optimum evolving function. It can keep the evaluation methods identical and ensure the performance and correctness of the selected generation in the neural network predict model. After decides the neural network training cycles, evolving generations, exchanges probability and mutation probability which can result in the optimum generation. After ‘computing’ and from ‘genetic algorithms result’, user can observe the evaluation fitness convergence curve and output the selected optimum descriptors file. See Figure5.



Genetic algorithms

Genetic algorithms parameters

Input before evoluation GIS2 file: fitness.dat

BK train cycles: 1400

Generations: 140

Exchange probability: 0.25

Mutation probability: 0.04

Genetic algorithms result

Optimum generation: 140

Optimum value: 0.86921

Output after evoluation GIS2 file: fitness_ga

Figure5: Genetic algorithms optimum analysis model

However, because the variables of the genetic algorithms first generation are generated randomly, they gradually adjust through exchange and mutation probability. It will mean passing through many generations to obtain an optimum solution if we are facing with a large amount of variables. At the same time, if the records are plenty it will also require a long time to get the convergence status. It will also probably fall into the local minimum, and not the global minimum because of the naturally selected engine, which is not able to guarantee the same optimum result in every selected cycle. So the ASIA system in genetic algorithms provides the elasticity to adjust the evolution in the small range of control of the user's specialized domain knowledge to ensure effectively fit the optimum solution.

4. EXAMPLE ON TEMPORAL DATA MINING CASE STUDY

This research selected an hourly air pollution observation station in Ban-Chiao City and uses 1996 data as the ASIA neural network data-mining basis to develop a forecast model of the air pollutants concentration.

4.1 The Case Study and the Conceptual GIS Spatial and Temporal Analysis Architecture

The conceptual GIS spatial and temporal analysis architecture can deal with complex geographic issues, and it includes the spatial and temporal analysis development stage that is the directing side, the data preparation system, which is the integrated system, and the operational research steps, which can elastically adjust, in different situations and under different circumstances. See table3.

Table3: Relationship between the case study and
the conceptual GIS spatial and temporal analysis architecture

Spatial and temporal analysis development stage	Integrated information platform	Operational research steps
Data Preparation (database extraction, manipulation, and management)	Data Preparation System	Step1: Air pollutants station data search and collect.
		Step2: To decide the air pollutants forecasting target (SO_2, CO, O_3, PM10, NO_2) and period (between 1 to 3 hours).
	Spatial data descriptor system	Step3: To filter data noise and deal with the data in advance if needed. (Obvious wrong and unavailable)
Data mining and Knowledge discovery		Step4: To transform all the variables and goals to GIS2 file format (pollutants on the right time, accumulate before 24 hours, the change on the forecast period, and the change rate).
	Artificial intelligent technology modules	Step5: Select, using genetic algorithms and neural network as the model of air pollutants forecast
	Artificial intelligent Spatial and temporal Information Analyst (ASIA)	Step6: Adjust the parameters of the genetic algorithms (the neural network training =1400 cycles, the evolve =140 generations, the exchange probability=0.25, and mutation probability=0.04)
		Step7: Output the optimum generation and output to GIS2 file format. E.g. corresponding with the input variables $f_{so2:}$(1100011001000011010101111101001)
System simulation and verification		Step8: To build neural network learning (before the 25th day of every monthly) and verifying (after the 26th day of every monthly) datasets
		Step10: Use co-variance to check neural network learning and for verifying the result.
	Apply knowledge base	Step11: Use the neural network weighting matrix to establish the air pollutants forecast model

4.2 Research results

4.2.1 Filter data noise and adjust the time lag phenomenon

In the research of data mining and knowledge discovery we always face lots of original digital data without any calibration or examination. For example, the thickness of the air pollutants is important to the physics factor in the movement of the atmosphere. Supposing that a typhoon or a front happened to pass by the station it would tend to leave the recordings to be very low. Obviously, it is important to eliminate this data, but exclude the restrictions of the model itself. In this case study the data will irregularly ascend or descend sharply in some particular period because of the instability of the auto-recording instrument. For example, the SO_2 forecast for the days of 3/31 and 7/27, the observed data have unstable situations at non-rush hour at the end of the day. And it will seriously affect the predicting accuracy. In this study case the data preparation system automatically deleted the unavailable data and used adjacent records, filtering obviously wrong data. See figure6.

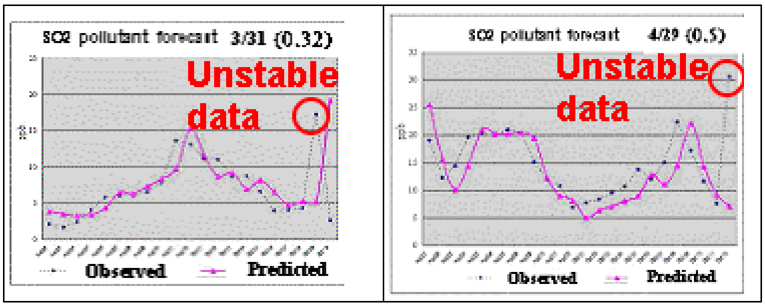


Figure6: SO_2 pollutant data noise

Besides the data noise, the time forecast research also has a time lag phenomenon. This is because the database does not have the relative data prior to the next time data appearance. This time lag phenomenon can be improved by searching the time cycle phenomenon of the database. This study case used the accumulation of the prior 24 hours as the neural network input variables. This apparently decreased the forecast delay phenomenon from 0.67 to 0.88(in the example of SO_2 pollutant). See figure7.

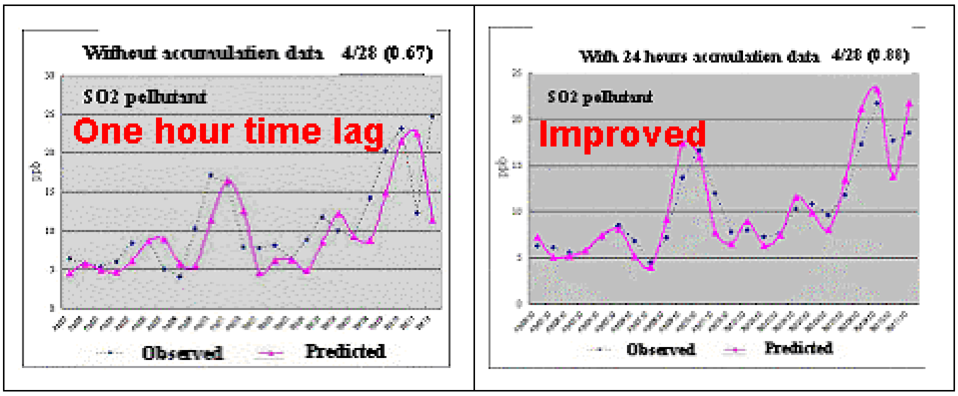


Figure7: The improvement of the time lag phenomenon

4.2.2 Genetic algorithms evolution process

Before input the genetic algorithms had 32 variables including the pollutants deepness and the conditions accumulation of the surrounding atmosphere quantity of change and the rate of change. See table4. Figure8 is an example of the SO2 evoluate function convergent process.

Table4: 32 genetic algorithms evoluation variables

Variable ID	Variable Contents
1	The forecast objective pollutant deepness
2-16(the 24 hours accumulations)	Sulfur dioxide, carbon monoxide, ozone, minute particles, nitrogen oxide, nitrogen monoxide, nitrogen dioxide, hydrocarbon, non-methane hydrocarbon, atmospheric temperature, dew point, ground surface temperature, atmospheric pressure, ultraviolet radial and methane
17-31(variables changed quantity in a particular time period)	
32	The objective pollutant change rate

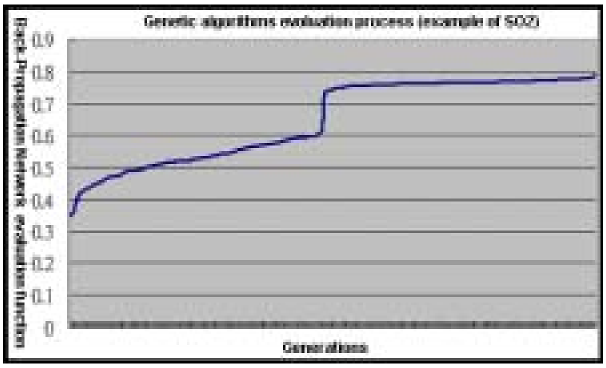


Figure8: The SO2 evoluate function convergent process

The evoluate function convergent process can get the best value and find the best variables combination. In genetic algorithms “0” means not to choose and “1” means it is one of the best variable selections. ASIA can directly convert the optimum result into a new GIS2 file. See table5.

Table5: The genetic algorithms optimum variable selection

Fso2=(1,1,0,0,0,0,1,1,0,0,1,0,0,0,0,1,1,0,1,0,1,0,1,1,1,1,1,0,1,0,0,1)
F_{co}=(1,1,1,1,1,1,1,0,1,0,1,1,1,0,0,1,1,1,1,1,0,0,1,0,1,0,1,1,1,1,1,1)
Fo3=(1,0,1,1,1,1,0,1,0,0,1,0,1,0,0,0,0,1,1,0,1,1,1,0,1,1,0,1,0,1,1,1)
Fpm10=(1,0,0,1,1,1,1,1,0,1,0,1,0,0,1,1,0,1,0,1,0,0,0,1,1,0,0,0,1,0,1,1)
Fno2=(1,1,1,1,0,0,1,1,0,0,1,0,1,0,1,0,1,0,0,0,1,0,1,1,1,1,0,0,0,1,0,1)

4.2.3 neural network forecast result

4.2.3.1 the result between 1 to 3 hours periods forecast

This research used the correlation index to examine the predicted results from the neural network. The two data groups were the observing data and the verifying data. We used observing data to feed into the neural network and obtained a weighting matrix. And we use this weighting matrix in the verifying data to compare the estimated values and the real values.

In table 6 it shows the 1 to 3 hours period forecast evaluation. For the five pollutants the 1-hour period forecast correlation index average is 0.94, the 2 hours period forecast is 0.77, and the 3 hours period forecast is 0.67. In the one hour predict O3, although the lowest is also almost 0.9, PM10 is the highest at almost 0.97. The non-linear model can be faced with a turnabout, a whirl, or a duplicate time and it will vary its curves, However, it is only suited for short time analysis and prediction because of its long acting feedback characteristics, which are always hard to control. The same quality of firm forecasts as made for the short time period cannot be guaranteed for the long time period. Even though the case study prediction had a high accuracy on the 1-hour air pollutant forecast it could not extend that same accuracy to the longer time period prediction. Obviously, the non-linear model is more suitable for short time period forecasts and decreases progressively as the time period increases.

Table6: The neural network result between 1 to 3 hours periods fore

Forecast object Forecast time period	SO_2	CO	O_3	PM10	NO_2	Average
1 hour forecast correlation	0.952	0.947	0.899	0.969	0.952	0.94
2 hour forecast correlation	0.793	0.764	0.677	0.857	0.749	0.77
3 hour forecast correlation	0.703	0.620	0.520	0.753	0.622	0.64

4.2.3.2 One-hour daily forecast result

For the one hour daily forecast neural network predict correlation index, almost all the pollutants in all days are above 0.9. Each month, after the 26th day the data is verified, and there are a total of 63 days. For SO2, 50.8% of the days were above 0.9 and 88.9% were above 0.8. For CO 68.3% of the days were above 0.9, and all the other days were above 0.8. For O3, 47.6% of the days were above 0.9 and 90.5% were above 0.8. For PM10, 60.3% of the days were above 0.9 and 92.1% were above 0.8. For NO2, 65.1% of the days were above 0.9 and 96.8% were above 0.8. See table 7.

Table7: air pollutants, daily predict achievement

Correlation index	SO_2	CO	O_3	PM10	NO_2
Above 90%	50.8%	68.3%	47.6%	60.3%	65.1%
Above 80%	88.9%	100.0%	90.5%	92.1%	96.8%

5. SUMMARY AND CONCLUSIONS

In the multi-variable analysis of statistics it is not easy to find notable variables in complex geographical problems. This research used genetic algorithms to precede naturally evolution in select variables. This was for the sake of convenience, because then all the possible variables can be used and there is no need to worry about duplication, representation or mutuality problems. In this case study we listed all of the variables recorded by the air pollutants station by itself, and so the genetic algorithms can properly select variables through natural selection. If it is needed the user can also join or delete variables depending on there specialized domain knowledge.

This case study forecasted the air pollutants thickness for the next time period. Corresponding to the need of the researcher it can also expand to forecast the changing rate, the changing patterns…etc. Compared with spatial analysis tools the time trend analysis tools are insufficient in the GIS. This research established artificial intelligent analysis tools in data mining and knowledge. They not only apply in other spatial analysis issues but also can appropriately strengthen the inadequacies of the GIS in time trend analysis ability.

We didn’t start from the traditional environmental engineering point of view, but from the data mining and knowledge discovery point, so as to analyze a vast amount of instant digital data in the forecast issues. Inevitably, the data mining research approach is restricted by the quality of the data itself.. It could not greatly expand itself around the pollutant measuring station area, because of the spatial variations. In conclusion, without fluid mechanic dynamics it is much cheaper for data collection, it is faster in computation, and it is easier to get the next time period pollutants thickness.

ACKNOWLEDGMENTS

This research was supported by Dr. C. H. Chen kernelled AI C++ program library.

REFERENCES

Anselin, L. (2000) “Computing Environments for Spatial Data Analysis”, *Journal of Geographical Systems*, 2:201-220

Boots, B. (2000) “Using GIS to Promote Spatial Analysis”, *Journal of Geographical Systems*, 2:17-21

Fischer, M. M. (1997) “Computational Neural Networks: A New Paradigm for Spatial Analysis”, *Environment and Planning A*, 29:1873-1891

Frank, A. U. (2000) “Geographic Information Science: New methods and technology”, *Journal of Geographical Systems*, 2:99-105

Gahegan, M. (2000) “On the Application of Inductive Machine Learning Tools to Geographical Analysis”, *Geographical Analysis*, 32(1): 113-139

Leung, Y., Leung, K.S. (1993) “An Intelligent Expert System Shell for Knowledge-based GIS: 1. The Tools”, *International Journal of Geographical Information Systems*, 7(3): 189-199

Marble, D. F. (2000) “Some thoughts on the integration of spatial analysis and geographic Information Systems”, *Journal of Geographical Systems*, 2:31-35

Muller, B., Reinhardt, J. and Strickland, M. T. (1995) *Neural Networks: An Introduction Physics of Neural Networks Series*, Berlin: Spring

Openshaw, S., Openshaw, C. (1997) *Artificial intelligence in Geography*, John Wiley and Sons LTD

Scott, A. (1990) “An Introduction to Genetic Algorithms”, *AI Expert*, 4(3): 49-53