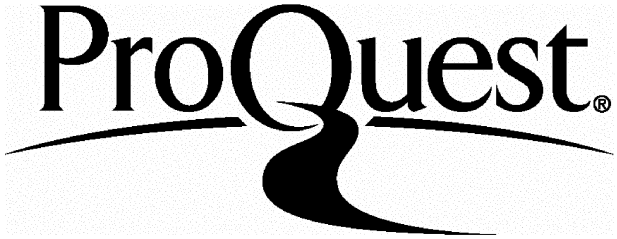
Department of Computer Science
University College London
2004

Intelligent Systems for Modelling Economic Policies

Elpida Makriyannis

Thesis submitted for the degree of
Doctor of Philosophy in Computer Science
University of London



Abstract

This thesis introduces 'intelligent country modelling', an intelligent decision-making tool to analyse macroeconomic structural issues. The Chinese dual-track economic model is selected to investigate the transitions from a centrally planned to a market-oriented economy and from a rural-agricultural to an urban-industrialised country. The model analyses Chinese provinces according to their growth type determining the regions that should adopt more rapid or gradualist policies. It also explores China's trading partner's over- or under-traded behaviour relative to bilateral trade flow knowledge. The understanding accomplished by this intelligent tool provides insight into China's future economic prospective thus assisting policy makers to design alternative strategies for country-specific sustainable equally distributed growth.

The thesis presents the design, implementation and evaluation of a new intelligent systems model, the Growth and Trade Country Analyser (GTCA), to combine the analysis of China's interior growth structure and international trade actions. The technical implications when considering building such a system are significant. Firstly, complex multivariate provincial data and regional growth indicators that occupy insignificant values or different provincial growth descriptions leave the assessed database sparse and incomplete. Secondly, asymmetric bilateral trade flows caused from different magnitude data sets for developed and developing countries make crucial prediction points noisy. These implications are solved by employing the GTCA that applies effective non-trivial detection and translation of explicit knowledge structures where complexity makes it impossible for human observation, statistical analysis or other intelligent methods.

The GTCA integrates two separate stages. The first stage employs Self-Organising Maps (SOMs) which combine clustering and projection algorithms to analyse the complex joint effect of the long-term growth factors and the critical properties of provincial growth structure by visualising and locating individual Chinese provinces on the map according to their growth type at a 2D level. The second stage uses genetic programming to determine the symbolic relationship that predicts over- or under-traded partner behaviour in China's real-world trading environment from interacting sets of bilateral trade flows.

Intelligent country modelling contributes to the detection, extraction and translation of previously unknown potentially useful economic growth and trade data element relationships. It reformulates policy goals through a feedback loop based on the results solving technical implications including noisy asymmetric trade flows and multivariate provincial growth range data. This model tests the SOM's ability to characterise complex provincial growth indicators and visualise important features of different dimensionality. This intelligent decision-making tool is believed to be the first designed to analyse economic policies and also the first intelligent model developed to identify China's economic growth structure and trading environment through observation of its reform period.

Acknowledgements

"Or watch the things you gave your life to broken and stoop and build them up with worn-out tools"
Rudyard Kipling

I would like to thank my supervisor Prof P. Treleaven, for his guidance, understanding, support and belief in me from the beginning and my second supervisor Prof D. Norse for sharing with me his expertise on the Chinese economy. I am indebted to Prof T. Kohonen and Dr J. Vesanto of Helsinki University of Technology, for always providing me with new suggestions and help. It was an honour to have Prof M. Vellasco read my thesis, I would love to thank her for many invaluable comments. I would also like to thank my examiners, Dr P. DeWilde and Dr A. Roe for reading my work and proposing very useful improvements. I would like to thank Prof. Paolo Lisboa for being the first person that believed in my research, Dr B. Langdon for his help with genetic programming and Dr D. Corney for meticulously reading my work as well as Prof A. Venables, Prof S. Broadbent and Prof R. Smith for their invaluable feedback on several economic issues researched in my thesis.

On a more personal note, I would love to thank Ms Anita Majumdar for helping me find the strength when everything went wrong. I would wholeheartedly love to thank Sir Harry Siegruhn for an inspirational conversation. I would like to thank Avy, Mike and Yanni in room 301 for creating the happiest lab environment and my friends Kenny, Maria, Marilena, Nagham and Sofia for keeping our friendship alive and supporting me each in their own way. I would like to dedicate this thesis to my parents and my grandmother. To my mom, Angela, the strongest and most intelligent person I know, who taught me to be caring, independent and fight for my dreams. To my dad, Yiorgo, for supporting me in every way imaginable and for his economic and political insight as well as his fighting for underprivileged people all his life which influenced me to design this tool. Finally, to my grandmother, Elpida, who was more interested in making sure I was eating well and wrapping warm than researching on things with funny names like SOMs and GPs.

Table of Contents

Chapter 4: GTCA part I: Intelligent Growth Policy Mapping

Chapter 5: GTCA part II: Evolving Bilateral Trade Flows

Chapter 6: The GTCA Model: A Case Study

Chapter 7: Conclusions and Future Work

Appendix A

Appendix B

List of Tables and Figures

Chapter 1

Introduction to the Growth and Trade Country Analyser

Chapter 1 presents the motivation behind this thesis to create a Growth and Trade Country Analyser and introduces intelligent country modelling. Established techniques to recent developments are identified in a brief overview of economic modelling in intelligent systems. The research hypothesis and objectives are presented and the research contributions are listed. In the final section, a thesis map organises the rest of the study into chapters.

1.1 Motivation for Intelligent Country Modelling

The research goal of this thesis is to investigate and develop tools for automating the application of intelligent systems for macroeconomic structural issues. In the past three decades the application domain of intelligent systems research has accelerated significantly, gaining acceptance in numerous diverse scientific research fields. There is an increasing objective to develop, evaluate and refine these techniques to produce more accurate and effective solutions to real-world problems. This necessity emerges from the belief that these systems analyse research applications in an abstract manner extracting them from their actual environment, ultimately increasing the possibility of performing in an unstable manner when faced with inevitable deviations. This thesis proposes the design of a decision making tool that aims to prove that intelligent techniques can efficiently and consistently deal with complex real world problems in an automated manner. An intelligent decision support analyser is designed to upgrade an expert's decision making in macroeconomic policies.

Development and information economics question the conventional form of decision making in macroeconomic structural issues - growth and trade policies – arguing that good economic policies have the power to change the lives of people in developing and emerging economies. They stress the need for models that adopt the position of providing explicit knowledge into growth and trade laws to explain the fluctuations, declines and recessions, that influence the development of market economies all over the world [Chom96], [Tod99], [Stig02].

World organisations, including the World Bank and the Asian Development Bank (ADB), emphasise the need for models that pursue a country's autonomous reform in economic policies by proposing alternative views of economic development strategies. The complex multidimensional issue of economic development looks at the social, political and institutional changes necessary to rapidly improve the standard of living especially for poor people in developing countries. This rapid improvement occurs with the reduction and elimination of poverty, inequality, and unemployment within a growing economy and the production of more life sustaining necessities such as food shelter, health care, and the broadening of their distribution [Nas99], [Stig02].

In order for these macroeconomic development policies to be successful, governments are required to play a more active role in shaping markets by promoting new technology and investment policies and guiding the order and pace in which they initiate these reforms with extreme care [Stig02]. An important obstacle to this development is the way in which advanced industrial countries retain their own barriers in primary interest goods while pushing developing countries to eliminate theirs, preventing them from exporting their products and consequently depriving them of export income. A combination of labour and business interests in highly developed countries jeopardises international trade relations and laws of free trade due to stimulation of protectionist interests. Duties of selling products below cost are called dumping duties and are seen as austere protectionism [HoeKo95], [Krug99]. International trade laws examine these legal and institutional aspects of today's global trading system and focus on how countries can conduct trade in goods and services across national borders in a fair manner for all parties concerned. Another important issue in defective use of trade laws is the economic policies imposed by powerful world organisations to developing countries such as the unnecessary cutting out of food subsidies that can be afforded [Klein01].

Figure 1.1, shows the interdependency of the order and pace in growth and trade policies required to initiate successful economic development. Stability and equality are the criteria for pacing and sequencing in growth policy formation, whereas trade policy criteria that promote these features are transparency and fair trade.

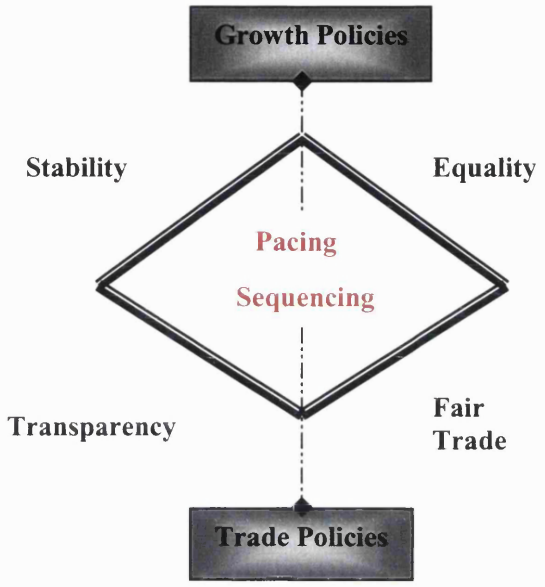


Figure 1.1 Structural Policies Criteria

Asymmetries in information of growth and trade policies between particular regions and country groups can cause erroneous decision-making that is extremely costly for countries and their governments. There does not seem to be a system available that deals with these asymmetries. The main objective in this study is to propose such a tool in order to assist in detecting and evaluating the asymmetries that cause structural inequalities in the world economy [Stig02]. It is concerned with the design, implementation and evaluation of an intelligent systems model, namely the Growth and Trade Country Analyser [GTCA], which contributes to testing the effectiveness of alternative policies in sustaining long term economic development. Moreover, the GTCA allows experts to capture the important interactions among the policies pursued under different economic circumstances.

The importance of such a tool is substantial since it employs intelligent systems to extract previously unknown information to assist in the analysis of crucial country-specific structural issues and the modelling of policies to deal with these issues. It achieves this by analysing country growth structure and predicting international trade actions. It analyses provinces according to their growth type, determining the regions of a country that should adopt more rapid or gradualist policies for rural-agricultural to urban-industrial transitions and explores over or under-traded partner behaviour relative to bilateral trade flow knowledge. The Growth and Trade

Country Analyser aims to help experts in fields including development and information economics to design policies that will help countries achieve sustainable growth and also ensure that it is shared more equitably.

1.2 Research Goals

The research goal of this thesis is to investigate and develop tools for automating the application of intelligent systems in macroeconomic structural issues. The techniques introduced aim to provide a general purpose methodology for developing intelligent country modelling tools for understanding a country's economic prospective. The main research objective is introduced below together with a summary of the key aims.

1.2.1 Main Research Objective

The main objective of this research is:

> ***The design and evaluation of an intelligent decision making tool to analyse country growth type and determine trade behaviour from bilateral trade flows.***

The selection of successful country-specific growth and trade policies between high income developed countries and emerging developing economies has become a central component in the future of economic centres dynamics. There seems to be a knowledge gap between the ability to understand structurally different economies and trying to integrate them into the world economy while preserving their distinct economic identity. The main objective of this thesis is to reduce this gap through the design, analysis and experimentation of an intelligent systems model that automates knowledge discovery and translates implicit into explicit knowledge.

Choosing the correct intelligent computational techniques to explain this knowledge gap in both growth and trade policies is not trivial. Adding to this decisive selection of the main components of the system is the fact emphasised by the NFL theorems [WolpMa97] that there is no universal algorithm for all possible problems. There is no single intelligent system to both provide a direct visualisation of different provincial growth types and discover the equation determining a country's trade environment. Consequently, we employ two different intelligent techniques, Self-Organising Maps (SOMs) to extract information about the effect of differing provincial growth policies and Genetic Programming (GP) to discover the mathematical relationship in country-specific trade actions.

Conventional statistical tools rank growth according to indicator volume which does not consider the numerous differing factors that affect and characterise each province's growth pattern. The SOM solves this problem since with its projection and clustering capabilities it is able to extract information from multiple criteria in parallel providing feature classification and representation in a graphic and intuitive form. We employ the SOM to synthesise and visualise multivariate provincial growth data sets, detecting and illustrating structures within the data and classifying the provinces into categories with different growth identities.

The second intelligent technique we use is GP due to its unique ability to identify and explain relationships in continuously evolving dynamic environments. Conventional models usually focus on finding the coefficients determining trade for a particular country rather than identifying the functional form that best specifies the country's bilateral trade actions. GP is employed to genetically breed populations of trading rules in order to find a symbolic relationship that best expresses China's bilateral trade environment through Darwinian natural selection processes.

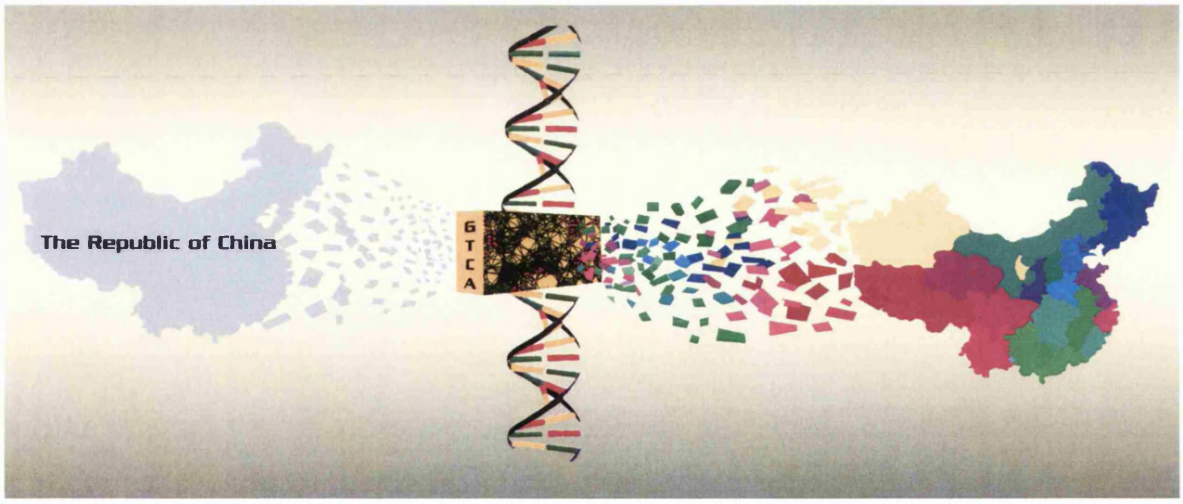


Figure 1.2 – Component Sequence of the GTCA model

Figure 1.2 graphically depicts the different components and processes taking place in the GTCA. We start with an initially grey geographical map of China, with little or no knowledge about the data sets we have collected depicted by the grey pieces floating towards the GTCA model. Once inserted into the model two different processes take place; provincial growth type classification and identification of trade environment created by partner countries trade actions. The main GTCA box with neurone connections represents the SOM and the DNA helix features the GP structure. Once processed, the data emerges from the GTCA tool with different colours illustrating the new knowledge acquired from the tool. Each coloured data input now explains a specific feature of the Chinese environment on a provincial and national level.

1.2.2 Thesis Aims

The key thesis aims that complement the overall objective described in the previous section are listed in this section with a brief explanation. A detailed discussion of each aim is found in the respective chapters throughout the rest of the thesis.

The proposal of intelligent modelling of economic policies

Erroneous decision-making on growth and trade policy issues due to asymmetries in information has proved extremely costly for countries and their governments. Although this is a problem affecting all countries in the world economy, there is no tool available to assist experts and world organisations to design better policies based on the knowledge acquired from policy interactions under different economic circumstance. The demand for an intelligent tool that reduces the knowledge gap between structurally different economies and the policies needed to regulate long-term economic development is the reason for proposing intelligent modelling of economic policies and introducing the Growth and Trade Country Analyser (GTCA). The GTCA employs two crucial intelligent components, Self-Organising Maps and Genetic Programming.

The investigation of SOM provincial growth analysis

In section 1.3 we provide a survey of many different applications of the SOM. The World Poverty Map application provides initial indications of the SOMs ability to portray complex correlations of unstructured data that form a socio-economic 2D mapping. It is this feature of the SOM that we use to investigate China's distinct and differing provincial growth types (Chapters 2 and 4).

Employing GP to determine bilateral trade actions

This objective is related to the use of GP to find the equation that portrays China's trade environment and the use of this equation to identify future partner country's actions. GP is modelled to determine the relationship between China and trade actions of developed and developing partner countries. Unlike other economic applications in GP – surveyed in section 1.3 – there is no formal equation expressing a country's trading environment, thus we do not know what the final relationship should be and assess the results according to our reference data (Chapters 3 and 5).

The identification of crucial components, capabilities and limitations of this intelligent analyser

The validation and assessment tests on growth policy analysis at a provincial level and bilateral trade behaviour at a national level that are performed should identify benefits and boundaries of the two individual sub-systems and the system as a whole (Chapters 6 and 7).

1.3 Brief Overview of Intelligent Economic Modelling

In this section we provide a brief introduction to intelligent techniques and their application to economic modelling. Most of the early research in intelligent systems was biologically inspired and did not directly relate to devising new methods of computation. However, in the 1940's mathematical modelling of neurones was introduced by the McCulloch-Pitts model [McCulP43] and Hebbian [Hebb49] learning, which researched adaptivity and learning in simple computing devices for modelling neurological activity in the brain. Computational biology techniques were established later in the 1960's by Rossenblatt's [Rosen62] study of perceptrons and in 1970's Kohonen's [Koh89] modelling of activity in the visual cortex concentrated on finding biological credible models of self-organisation.

The same period launched the field of evolutionary-inspired computation algorithms for optimisation and machine learning. Rechenberg [Rech73] introduced evolution strategies, designed to optimise real-valued parameters for devices such as airfoils, a concept that was further developed by Schwefel in the 1970's. Meanwhile, Fogel, Owens and Walsh [FoOW66] developed evolutionary programming, where candidate solutions (finite-state machines) to given tasks were evolved by randomly mutating their state-transition diagrams and selecting the fittest. In contrast to the above techniques, [Holl75] studied adaptation as it occurs in nature and focused on developing mechanisms to import natural adaptation in computer systems through the development of genetic algorithms (GAs). Inspired by Holland's technique Koza [Koza89] developed genetic programming (GP) where computers try to program themselves by evolving programs. The past 40 years have spawned many important intelligent techniques. This study will focus on the Kohonen Self-Organising Map and Genetic Programming which will provide the background for much of the work presented in the later chapters.

1.2.1 Established Techniques

The focus of biologically inspired techniques on the mathematical modelling of neurological activity in the brain generated artificial neural networks, intelligent techniques that imitate specific procedures of the nerve cells in the brain. Neural networks are able to learn directly from data sets, they adapt to changing knowledge environments and make flexible decisions on imprecise and incomplete data. They are classified as employing supervised or unsupervised learning. Supervised networks use external criteria to match the network's output, whereas unsupervised apply input patterns but supply no target outputs. In unsupervised learning there is more of an interaction between neurones, typically with interlayer connections between neurones promoting self-organisation [KinJ97].

The Kohonen Self-Organising Map (SOM) [Koh89] is such an unsupervised neural network. Kohonen showed that SOM feature maps can be developed in artificial neural systems as a consequence of simple learning rules. The layered structure of self-organisation of the visual cortex in the cerebral hemispheres of the brain with no external teacher to guide the structures development is the process that inspired Kohonen to create the SOM. In the visual part of the cerebral cortex, electrical stimulation of the cells produces the sensation of light, causing specific layers of neurones to be sensitive to particular orientations of input stimuli, responding either to horizontal or vertical lines. The orientation-specific layout of the SOM mimics these workings in the cortex and is a unique method that combines projection and clustering algorithms [Kaski97], [Ves97], [Ves00].

SOMs have been successful in many different application fields including industry [SiVe99], financial forecasting [MarSer95], textual data mining (WEBSOM) [KoKHL00], brain organisation [KohHa99], and world economic profiles [KaKo96]. From these successful applications important features of its capabilities emerged, including its ability to represent graded relationships and its innovative manner in finding unexpected structures in historical and high-dimensional data. This performance of the SOM is influenced by the choice of cost function, variables, neighbourhood architecture and learning duration.

The other important technique employed in this thesis is obtained from the field of evolutionary-inspired computation algorithms developed by [Koza89] called genetic programming (GP). Genetic programming is a descendent of genetic algorithms which mimics the process of natural populations that evolve according to the principle of the "survival of the fittest" stated first by Charles Darwin [Dar59]. It automatically generates and evolves computer programs towards increasingly better regions of a search space through the use of randomised processes of selection, crossover and mutation. It favours the best solutions to a specific problem without being explicitly programmed and maintains a certain level of diversity rather than focusing on one specific solution [Koza92a]. The GP technique allows the size and complexity of candidate solutions to increase over evolution by operating on nonlinear tree structured material with crossover operators defined so as to preserve the syntactic correctness of the program [Kinn94].

The GP technique has succeeded in evolving correct programs to solve problems in a number of varied application fields, including performing optical character recognition [Andre94], protein classification [Handl93], image processing [DaBRV96], electronic circuit design [Koza96] and car monitoring for pollution control [HaBM94], [Lan98]. GP uses fitness evaluation, a population of program solutions to search the space rather than a single solution, and also probabilistic transition rules. Research in all these applications fields concluded that the genetic programming approach should be employed in complex real-world applications due to its power of effect.

1.3.2 Recent Developments

One of the most important recent economic developments of the Kohonen Self-Organising Map (SOM) is the formation of a socio-economic mapping, the World Poverty Map [KasKo95], that visualises World Bank indicators that describe welfare and poverty structures of the world. Through this research application, an exceptional capability of the SOM to portray complex correlations in unstructured statistical data on a 2D map has been identified. General-purpose function estimators employed by some intelligent systems researchers give rough calculations that can fail to achieve desired accuracy and reliability relative to the eventual extensive costs associated with wrong decision-making [Kaski97]. SOM-based exploratory data analysis was employed for the creation of the World Poverty Map in order to solve the above implications with the only intervention being introducing new data sets presented in an easily understandable format, as the map compares standards of living in different countries.

Another important application includes the Spanish banking crisis of 1977-85 and the financial state of Spanish companies in 1990 and 1991, which demonstrates the SOMs capability, as an unsupervised neural network, to cluster input patterns according to their similarity when outputs are unobtainable [SerMa93]. In the first case study, without prior information, the self-organising map discovers similarities between the patterns and clusters the banks according to their solvent or bankrupt state. In the second case study, using financial data for 1990 and 1991 from 84 Spanish companies, the map finds features of large liquidity and small profitability or large debt where similar companies are clustered. [DebKo98] present a collection of SOM applications in economics and finance from various authors.

[Koza92a] employed genetic programming to rediscover basic physical laws including Kepler's 3rd law and Ohm's law from experimental data. [Levy92] adds that not only did GP rediscover Kepler's 3rd law but also discovered an earlier conjecture by the mathematician from one of its interim solutions with high fitness. [Koza92b] then applied the technique to eliciting the quantity theory of money or the exchange equation which relates the price level, gross national product, money supply, and velocity of money in an economy. Koza set the basis for GP applications in economics by evolving the entire equation and not just the parameters of the exchange equation, thus demonstrating its capability as a knowledge discovery tool. Finding the functional form of this econometric model is viewed as searching a space of possible computer programs for the particular computer program that produces the desired solution. This desired solution can be found by genetically breeding populations of computer programs in a Darwinian competition using genetic operators.

[Koza92b] motivated a series of economic applications of genetic programming in the mid-90s including [Neel97] and [AlKarj99] that adopted the GP approach to discover profitable technical trading rules for the foreign exchange market and stock market. Keber in [Keb99] and [Keb00] showed that genetically determined formulas outperformed calculations of analytical approximations of volatility based on the Black-Scholes model. [Chid00] also calculated option prices deriving approximations that depicted that GP-models outperformed various other models.

Both SOMs and GPs are well established search techniques in Artificial Intelligence. The task of selecting the one that represents and analyses the application problem in the most efficient manner is not trivial. [WolpMa97] argue in the No Free Lunch (NFL) theorems that there is no universal algorithm for all possible problems since if a technique fits a particular problem accurately, there exist other problems for which its performance is as limited as random search. Consequently, we employ each search technique to perform a particular task, which is outlined in the following section.

1.3.3 Intelligent Modelling Design Requirements

Requirements of economic data analysis and intelligent systems essentially have a similar objective; to design a model that can generalise future events based on past behaviour. These requirements for both scientific fields are listed below and figure 1.3 indicates how these validation criteria are equivalent and complement each other.

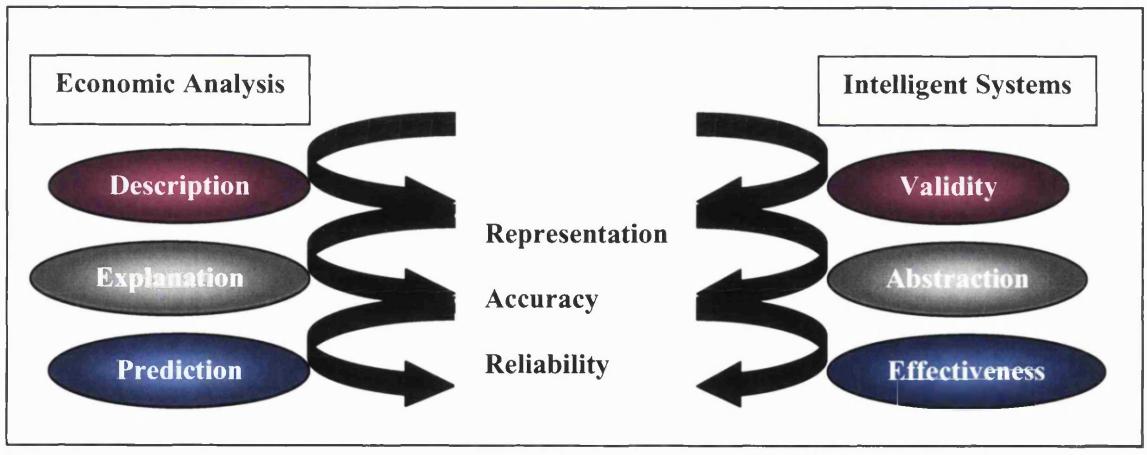


Figure 1.3 – Intelligent Country Modelling Design Criteria

The criteria for assessing economic forecasting include [Chat89]:

- *Description* → The search for dominant features that describe the main properties of the specific data set;
- *Explanation* → Detecting how the most important variables affect each other;
- *Prediction* → Finding future values.

The criteria for assessing the performance of an intelligent system [MiCM86] are:

- *Validity* → The representation accuracy when compared to the real-world problem;
- *Abstraction level* → The capacity and accuracy of exploring the data set for an accurate representation; and
- *Effectiveness* → The reliability of the representation.

Both validity and description criteria require representations of the specific data set. Abstraction and explanation are both focus on determining the accuracy of the representation through the detailed exploration of the data. Finally, effectiveness measures the reliability and accuracy of the prediction. It is these similarities shared between the two fields of economic analysis and intelligent techniques that make economic analysis an extremely convenient domain in which to apply intelligent system techniques.

1.4 Research Contributions

The main contributions of this thesis are:

1. The design and evaluation of an intelligent decision making tool to automate knowledge discovery of previously unknown relationships between economic growth and trade elements while solving technical implications including noisy asymmetric trade flows and multivariate provincial growth range data.

2. The design and development of the first successful intelligent decision-making tool that identifies some of the many crucial features of China's economic growth structure and trading environment through observation of its entire reform period.

3. Incorporating projection and vector quantisation methods in parallel, the SOM algorithm is used to characterise inputs it has never seen before and locate missing input data based on the trained map. It classifies underlying policies promoting province-specific growth and visualises important features of the different dimensions of growth inequality by modifying its internal state.

4. The use of GP to search for the symbolic relationship expressing and determining a dynamic evolving environment from the formation of over or under-trading rules relative to bilateral trade flows. The discovery of secondary symbolic forms highlighting the level each variable influences the main trade relationship and the acquisition of these secondary equations from the GP via high partial crediting.

5. The creation of a database (electronic version) of the Chinese provinces for the entire reform period including agriculture, industry, construction, population, energy and GDP figures.

1.5 Thesis Map

Chapter 2 gives an insight into China's internal state, observing the economic reform strategies, and the growth divergence of eastern, central and western provinces. Recent and pre-reform economic growth in the different regions is overviewed. The need for a system that lays the foundation for sector-specific analysis and investigates the strategic considerations that could influence policy directions is emphasised. The Kohonen algorithm is introduced together with its features, including the ability to encapsulate clustering and projection methods, its integration with exploratory data analysis and the criteria for selecting its main parameters.

Chapter 3 focuses on the structurally diverse Chinese trading policies and how they affect China's trading partner's behaviour. In order to determine the Chinese trade environment we employ genetic programming to identify the mathematical relationship measuring trade flows between China and its trading partners. The preliminary GP steps are introduced together with the selection of primary and secondary genetic operators. We discuss tree representation, effective program size, program efficiency, and genetic diversity and provide the mathematical proof of GP evolution using Price's Covariance and Selection Theorem to artificial evolution.

Chapter 4 focuses on the first part of the GTCA. It explores the strategic importance of SOM-based design in providing direct visualisation and location of the numerous Chinese provinces according to their distinct competitive growth strategies and geopolitical identities. The spectrum of SOM's abilities is widened by applying it to overall provincial growth self-organisation of the Chinese provinces. The extent to which each indicator influences the growth type of each province is also investigated by the SOM.

Chapter 5 focuses on the second part of the GTCA that tries to identify the functional form that best specifies China's bilateral trade behaviour through empirical trade flow data. Genetic programming is employed to portray the main relationship of China's trade pattern in the international market and forecast the future relationship of the state variables of the system. Discrepancies in export figures between China and its trading partners mainly due to re-exports via Hong-Kong are adjusted. The theorem of the gravity equation is employed to obtain trade flow knowledge from input data.

Chapter 6 described the combination of the algorithmic components explored in Chapters 4 and 5 providing the basic structure of an intelligent country analyser encapsulating direct visualisation

and location of complex unstructured correlations with the search for mathematical relationships determining evolving environments. The Growth and Trade Country Analyser (GTCA) is validated and assessed on its ability to investigate economic policy features in the European Union countries firstly in terms of the entire system and then with its separate components. The performance and applicability of the entire systems model as well as its individual components is validated. Finally, the model is assessed according to the design requirements and the research contributions are revisited.

Chapter 7 concludes this research leading to significant findings about the effectiveness and importance of the GTCA system created and applied, and for the future prospective of its application spectrum. The intelligent country modelling research area was proposed with the design of an analysis tool that comprises two separate intelligent systems that investigate different aspects of the Chinese economy. The final section proposes various new features for the GTCA system and some research areas that can be studied to extend and upgrade the system.

1.6 Summary

This chapter has introduced intelligent country modelling to assist experts in modelling alternative economic policies. It has also explained the motivation for the creation of a growth and trade country analyser based on intelligent country modelling. The analyser model outlines the ability to detect explicit knowledge, avoid extensive costs, achieve desired accuracy and present complex incomplete data sets in an easily understandable form. These are benefits that promote automation with minimal intervention, capture explicit features and solve technical implications and are introduced in the following chapters.

Chapter 2

SOM-based Provincial Growth Design

Chapter 2 gives an insight into China's internal state, observing the economic reform strategies, and the growth divergence of coastal, border and inland provinces. Recent and pre-reform economic growth in the different regions is overviewed. The need for a system that lays the foundation for sector-specific analysis and investigates the strategic considerations that could influence policy directions is emphasised. The Kohonen algorithm is introduced together with its features including the ability to encapsulate clustering and projection methods, its integration with exploratory data analysis and the criteria for selecting its main parameters.

2.1 China's Internal State

In order to understand China's national economy two distinct stages of development are described:

- *Adoption and implementation of a Soviet-type economy from 1952-1977*
- *Gradual economic reform toward a market-led economic system since 1978.*

Prior to 1978, China faced a hostile international environment with political isolation and economic embargoes. Political leaders adopted a heavy industry-oriented development strategy to catch up with developed western countries. By 1992 the solid foundations were established when China formally embraced the theory of building "socialism with Chinese characteristics" adopting the profile of a "socialist market economy". This "gradual" reform strategy was described as a dual-track approach - the coexistence of a market track and a plan track - and was launched simultaneously in the agricultural sector, in rural and urban manufacturing industry, as well as domestic and foreign trade [Good97], [FZZ02].

This transformation from an underdeveloped country with a high incidence of poverty to an emerging industrialised one and from a command to a market economic philosophy was conducted in ways that ensured that the underlying social philosophy and the social fabric of the society were preserved [StiHS00]. China's reform experience depicts that by retaining previously developed social and organisational capital and transforming it in ways that enhance future efficiency and productivity sets the stage for the next steps in a continuing dual-transformation reform process [Lar94].

2.1.1 Factors of Pre- and Post- Reform Growth Policies

The initial reforms were intended simply to make improvements to the performance of a command economy. Its successful evolutionary approach to decentralisation and rural reformation based on the three key processes in growth analysis - enhanced efficiency and productivity, accumulation of capital, and sectoral reallocation of factors - produced a market economy intertwined with a mixture of features of a command economy. For more than two decades China's real Gross National Product (GNP) grew around 9.6% per year compared to the average of 5.6% between 1953 and 1978 [StiHS00], [Nas99], [Chow94].

Policies that contributed to the qualitative differences of the pre- and post-1978 economic growth status include [StigHS00]:

- The opening of the Chinese economy to trade and direct investment,
- Extension of the range of physical and managerial techniques, and
- The reintroduction of economic incentives.

These policies can be explained with *two interrelated factors* [StigHS00]:

- The central strand of economic growth in a developing economy is the transfer of labour from farming to non-farming activities, and
- The rural economy has been the major locus of institutional transformations in the form of the decollectivisation of farming and an explosive growth of rural industry.

Results of these policies so far depict that governmental production-enhancing investments, such as agricultural research and development, irrigation, rural education, and infrastructure (including roads, electricity, and telecommunications) contributed not only to agricultural production growth, but also to reduction of rural poverty and regional inequality. Investment in the priority sectors of agriculture, physical infrastructure (telecommunications, transportation, and energy), and services (health, education, banking, and insurance) increased substantially. The rapid demographic transition to low-fertility slowed the annual population growth rate boosting economic growth since declining fertility led to higher rates of saving. The rise in real income in both urban and rural areas reduced the gap between these two incomes as well as the number of

people living below the poverty line. This growth in China was fuelled by a combination of two economy-wide reform policies [Chow94], [FZZ02], [LloZh00]:

- *Decentralisation* and
- *Restructuring of the rural economy.*

Decentralisation

Decentralisation before the reform era and promotion of self-sufficiency for regions provided a valuable platform for China's transition. One striking feature of China's economic transition is its strong dualistic nature which is heavily dominated by state-owned enterprises (SOEs) and tight controls while also influenced by high degree of liberalism and the township, village and private enterprises (TVEs). In the late 1970s rural poverty and unemployment drove the peasants to develop rural enterprises, when the central government decided to liberalise controls of the economy with the active support of the local governments. Entrepreneurs initiated the growth process by identifying and developing markets for products where each locality had a comparative advantage. As their comparative advantages and local needs differed, their development paths also varied. Although competing with SOEs for markets and resources, rural enterprises complemented them through backward and forward linkages thus pressuring them to perform more efficiently. The rural enterprises pushed forward towards reform by introducing institutional innovations such as flexible interest rates, share-holding enterprises and speciality markets that assisted in building the groundwork for a viable market system to which the SOEs must eventually adapt [KooYe99], [Hend99].

The introduction of decentralisation and market incentives in the reform era suggested that the centres of economic power were moving away from the centre to the localities; nonetheless, the centre still controlled the appointment of senior local leaders, and exercised control at a provincial level. The Chinese central government enhanced intergovernmental relations and enforced rural enterprises by substantially reducing its involvement in the direct management of the economy, turning instead to the principles of macro-economic control [Good97], [FZZ02]. China demonstrated enormous alertness, in first decentralising, providing significant scope and incentives for local economic activity, and then recentralising aspects such as control of the financial system [Hend99].

Restructuring the Rural Economy

The very large rural sector provided both the source of decentralisation and productivity growth in the very early stages of China's transition and the source of labour for industrial growth. In turn both decentralisation and rural organisational and institutional traditions provided the basis of the new largely collective, rural and industrial enterprises that drove China's growth from the mid-1980s. The gradual restructuring began with the central and local authorities designing several experiments that were evaluated initially to one province as a pilot study for new projects

followed by trials in several others. These trials were equally successful in all affected regions and depicted the gain from even partial reforms. The pace and degree of reform varied greatly across regions due to the gradualist and selective approach adopted, contributing to the widening of regional economic disparities [Stig02], [Hend99].

Rapid growth in agricultural labour productivity and rural income increased rural demand and generated tremendous labour surplus and initial investment. This led to explosive growth in the rural collective industry resulting and to a mixture of collective and private enterprises across rural localities. The loss of agricultural land in combination with a trend towards higher demand for agricultural products has resulted in discussions about the long-term capacity of the country to feed itself and its consequences. However, this might be solved by the high grain production which in China translates to food security and by the increasing possibility of China becoming a large vegetable exporter with a comparative advantage in producing vegetables for export, primarily due to its abundant rural labour resources. Investments in transportation and storage infrastructure as well as improvement of firms in grading and packaging standards are likely to make China a fierce competitor in world vegetable markets [FZZ02].

2.1.2 The Provincial Growth Debate

Most Chinese provinces are the size and scale of a European country in population, land area and social complexity. Figure 2.1 displays a geographical map of China with the provinces divided into three regions; eastern, central and western.

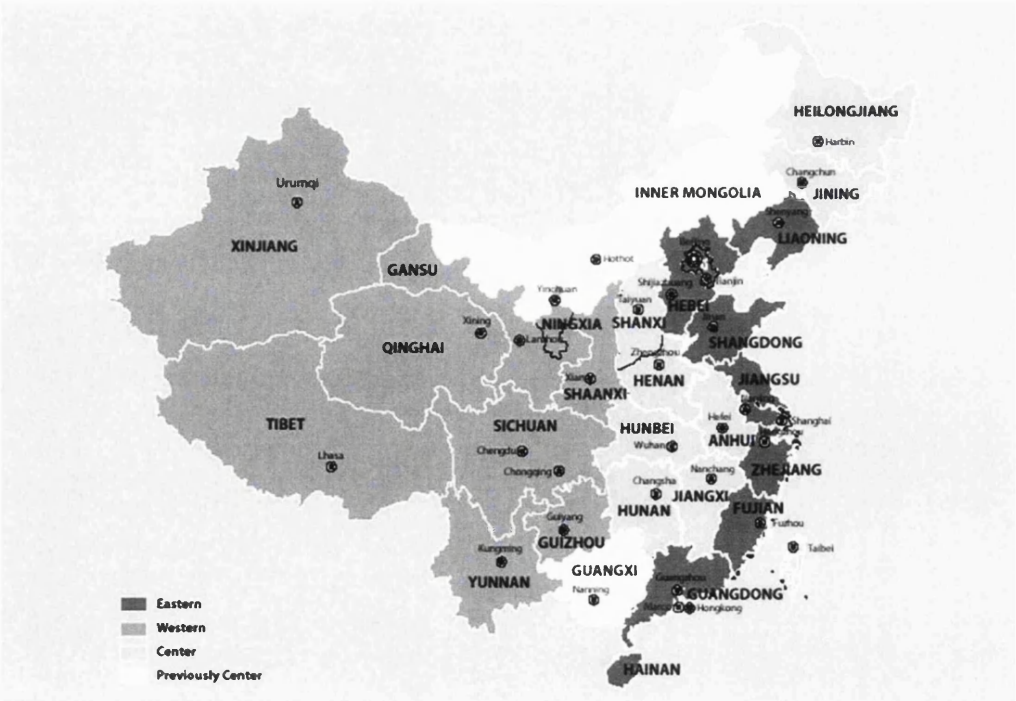


Figure 2.1 – Geographical Map of China (Source: ADB)

Table 2.1 complements figure 2.1 with information on the regional composition of percentile area occupation and population for each of the three regions.

Geographical Position	Area (%)	Population (%)
Eastern	18,18	43,94
Central	39,74	51,89
Western	42,07	4,17

Table 2.1 – Regional Composition of Land and Population (Source: ADB)

There are 3 factors determining the trajectories of economic development of China's provinces in the reform era [Chow97]:

- *Natural Endowment*, including geographical position, accessibility and natural resources;
- *Relations between province and the centre*, particularly as reflected in the latter's changing policies towards different provinces at different times; and
- *Economic Culture*, including a range of dominant beliefs and the behaviour of local government and the public in solving their economic problems.

Macroeconomic policies during the reform period benefited more the provinces that enjoyed the location advantage of being on the coast, and was less significant for the economic activities in the western provinces that account for the bulk of the rural poor [Nas99]. However, with east and south coastal provinces as their key customers, China's inland provinces have demonstrated that their location is not too remote to supply the rest of its interior.

The coastal region is the most developed with the best infrastructure and human capital as well as superior agroclimatic conditions. The western region is the least developed and has poor natural resources and social infrastructure. Major growth potential for agricultural production lies in the central region, where land is relatively infertile and agricultural production is still the main source of farmer's income. China is gradually eliminating the previous economic system with a crucial growth campaign for future growth performance and development of the west by undertaking large infrastructure projects to develop provinces linked by defective and inadequate infrastructure [Lar94], [Good97], [FZZ02].

By disaggregating the growth analysis into different regions it is revealed that for all types of government spending, returns to investments in poverty reduction were highest in the western region, while returns on agricultural production growth were the highest in the central region for most types of spending. Furthermore, for all types of government spending, investment in the western region led to the greatest reductions in regional inequality, while investments in either coastal or central regions worsened the existing large regional inequalities. In terms of regional

priorities, if the government aims to maximise poverty- and inequality- reduction effects, then investment should be targeted to the western region. In order to maximise returns from investment in agricultural production, investment should be targeted to the central region [FZZ02].

The patterns of provincial trade differ between three broad provincial groups. Coastal provinces have considerable advantages of global access putting them in the forefront of internationalisation and making their trading patterns similar to national trading patterns. Provinces with both border and coastal characteristics have the highest volume of trade and highest trade dependence. Coastal provinces have high volumes of trade but with low trade dependence rates. Border provinces have unique trade orientations and are variable in amount of trade and dependence [Good97], [Hend99].

Coastal and border areas have been encouraged to utilise the comparative advantage of their geo-political locations of economic involvement – trade, technology transfer and investment – to create external economic links abandoning the principle of self-reliance and adopting the open door policy in economic relations with the rest of the world. Hong Kong remains a significant export destination but beyond southeast China its share is exceeded by the other major trading partners that include Taiwan, South Korea and Japan that has a large presence throughout the Chinese economy. The possibility of entrepöt trade over inland borders is not insignificant, especially for Xinjiang through Kazakhstan and Yunnan and Guangxi through Vietnam. The coast has the advantages of money and the national structure of the economy. The inland areas have the advantages of means of basic production [GooSe94].

2.1.3 Prospects and Challenges

China's future growth prospects and challenges are considered below [FZZ02], [StigHS00]:

- *Sectoral transfer*: the fraction of the labour force in agriculture is still as high as 47% and the process of increasing productivity is still inferior to other activities.
- *Technological advance*: A more active role from the central government is required for the promotion of specific basic industries for rapid technological development.
- *Resource allocation*: China's World Trade Organisation (WTO) membership will help improve market development by further enhancing competition and efficiency as well as lead to greater integration of the economy as a whole.
- *Importance of social capital preservation and reinforcement:* Encouraging the dissociation of the social from the production functions and certifying limited disparities between the wealthiest and poorest provinces. In the process of maintaining and strengthening its national safety net is probably China's biggest challenge.

- *Utilisation of geopolitical potential:* Further encouragement and exploitation of the comparative geopolitical advantage of border provinces.

China's reform era is approached from one of three broad historical perspectives on the dynamics of social and political change [Good97]:

- The experience of Western Europe in the early part of the 19th century that gave birth to notions of capitalism and liberal–democracy;
- The implosion of communism and the subsequent political disintegration that characterized the former Soviet Union and Eastern Europe after 1989; and
- The transformation of authoritarianism that has been the hallmark of much of East Asia, but particularly Japan, South Korea and Taiwan since the 1970's.

None of the above three perspectives can be a precise predictor of China's interior future development. The variety of different economic, social and cultural environments, and different rates, sequences and processes of modernisation offer interesting possibilities about the interconnections of economic, social and political policy change. There have been few attempts to identify or conceptualise the dimensions of provincial variations in policy and performance, before the reform era let alone since 1978. Therefore, there is a necessity for a system to analyse the underlying policies of China's development through the reform period that promoted province-specific growth. The system should convey an immediate, clear and accurate illustration of the different dimensions of provincial growth levels.

2.2 Introduction to SOM-based Modelling

The intensifying competition for resources and markets has forced the Chinese provinces to differentiate themselves by developing distinct growth identities emerging as competitors for preferential policies. Although the necessity for differentiation has been realised, the selection of the underlying policies that precisely capture and promote province-specific growth in different sectors is not obvious. The Self-Organising Map (SOM) is employed to analyse fundamental growth determinants that affect the evolution of emerging growth policies and enable policy experts to gain an explicit view of province-specific growth types. The detection and accurate display of new knowledge from massive amounts of data depicts the path towards automation of knowledge discovery and the translation of implicit into explicit knowledge. Advanced intelligent analytical tools are applied for more effective knowledge and information management, concentrating on approaches for synthesising and visualising large multivariate data sets, detecting and illustrating structures within the data. Incorporating projection and vector quantisation methods in parallel, the Self-Organising Map is able to infer such relationships and classify them into different categories, modifying its internal state as it learns [DebKoh98].

2.2.1 Issues in SOM-based Exploratory Data Analysis

Exploratory data analysis acts as the route from knowledge extraction to non-trivial discovery and extraction of previously unidentified patterns. Multiple steps comprise exploratory data analysis from setting up the goals to evaluating the results and reformulating the goals based on the results through a feedback loop. Conventional graphical techniques including 2D graphical profile plots, scatter plots, Andrew's curves and Chernoff's faces represent high-dimensional data by each dimension governing some aspect of the sample factors and then integrating them into one. Though useful for illustrating summaries, their major drawback is that by not reducing the amount of data, the result is incomprehensible when dealing with large data sets, since all items are displayed. The SOM analysis solves these problems by reducing the original data set dimensionality while preserving properties of the input structure that are critical for the accurate translation and description of the input patterns. It allows direct visual relationships between elements in large, complex data sets and characterises inputs never encountered before and even forecasts values of missing inputs based on its resulting mapping analysis. This is achieved by incorporating:

- *Clustering Methods* and
- *Projection Methods*.

Clustering Methods

Clustering methods [Ander73], [JaiDu88], [JarSi71], [SneaSo73], are employed to automate the construction of categories or taxonomies of similar data items reducing the amount of inputs and minimising the effects of human biases or errors in the grouping process. Clustering techniques group information in the same manner that humans categorise knowledge. Clustering methods include, hierarchical and non-hierarchical clustering. Hierarchical clustering progresses successively by either merging or by splitting clusters with the end result being a dendrogram, which shows how the clusters are related. Non-hierarchical clustering emphasises on local structures, as it directly decomposes the data set into a set of disjoint clusters in order to minimise dissimilarity in the samples within each cluster, while maximising the dissimilarity of different clusters. An established non-hierarchical clustering method is the K-means [McQue67], where the goal is to find the average squared distance of the data items from their nearest cluster centroids. Most clustering algorithms prefer certain cluster shapes and sizes leaving other clusters empty when the centres lie far from the data distribution. This makes the interpretation of the clusters difficult and stresses the need for investigation of the clustering tendency [Kaski97], [DebKo98].

Projection Methods

Although clustering reduces the amount of information, there are methods that reduce dimensionality called *projection methods*, representing inputs in a lower-dimensional space but

preserving certain critical properties of the data structure. The projection methods can be divided in *linear* and *non-linear projection methods*. Principal Component Analysis (PCA) [Hotel33], is a standard linear projection method that displays data on a subspace of the original space that best preserves data variation. It does not take into account structures consisting of arbitrarily shaped clusters, since it describes linear subspaces. Multidimensional scaling (MDS) [KruWi78], is a non-linear method which creates a matrix that analyses subjective evaluations of pair-wise similarities. The method allows for visual inspection of the set with sufficient dimensionality reduction where entities are represented as vectors when only some evaluations of entity dissimilarities are available. Closely related to MDS is Sammon's mapping [Sam69], which tries to optimise a cost function describing the accuracy of pair-wise distances preservation in the data set. Principal Curves (PC) [HasSt89], are smooth curves projecting inputs onto a non-linear manifold where each point of the curve is the average of all points that project to it [MulCh95], [Ritt92].

2.2.2 The Kohonen Self-Organising Map (SOM)

Inspired by the self-organisation in the brain's visual cortex and created by Teuvo Kohonen [Koh89] the SOM incorporates the main qualities of both clustering and projection methods. Kohonen has shown that feature mapping can be developed in artificial neural systems as a consequence of simple learning rules. In the visual part of the cortex - the most complex layered structure in the largest parts of the brain the cerebral hemispheres - electrical stimulation of the cells produces the sensation of light. In addition, detailed analysis has shown that specific layers of neurones are sensitive to particular orientations of input stimuli, so that one layer responds maximally to horizontal lines while another to vertical.

The SOM Algorithm

Teuvo Kohonen modelled this effect of orientation-specific layout phenomenon of cells in the brain's cortex region using only locally interconnected networks and restricted adaptation of weight values to localised "neighbourhoods" [KohHa99]. The SOM network is able to infer relationships and learn more as more inputs are presented to it. It classifies inputs into different categories and modifies its internal state to model and characterise unknown or missing features found in the training data. The network adopts two assumptions: that cluster members are defined by input patterns that share common properties; and that the net will be able to identify common features across the range of these patterns. The cluster members are generated from the input patterns. Experimental work has proven that it is a valuable tool in numerous fields, including data mining with applications in full-text and financial data analysis [DebKo98], various engineering applications in pattern recognition [KoOSVK96], chemical structures [Nach00] and web-text information retrieval [LHKK00].

The main reasons for using the SOM algorithm include:

- *It is a numerical and non-parametric method;*
- *No a priori assumptions about the distribution of the data need to be made;*
- *It is a method that can detect unexpected structures or patterns by learning without supervision;*
- *It allows direct visual relationships between elements in large complex data sets.*

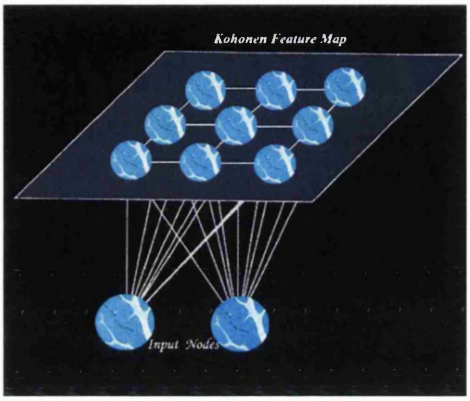


Figure 2.2.a The Kohonen feature map

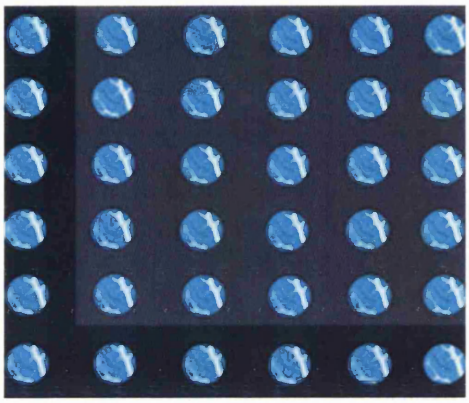


Figure 2.2.b Initial neighbourhood status

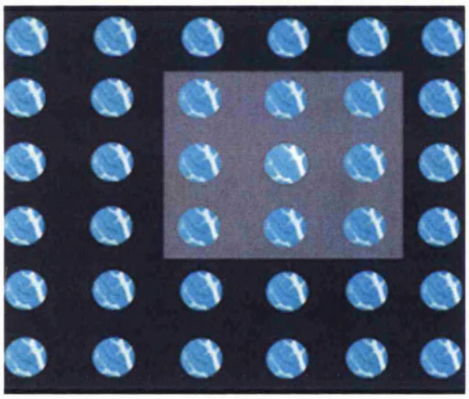


Figure 2.2.c Decreased neighbourhood size

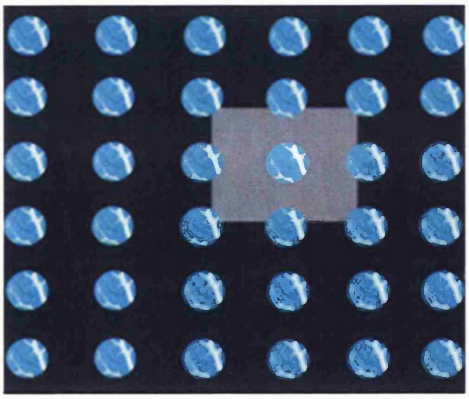


Figure 2.2.d Winning node

The SOM network depicted in figure 2.2.a is formed by two layers of processing neurones: the first is an input layer containing processing nodes for each element in the input vector; and the second is an output layer of processing nodes fully connected with the input layer. Each node has a unique weight vector with its dimensionality defined by the number of components in the input vector. The strength of their connections is inversely related to their distance. Starting from randomised weight values in figure 2.2.b the output units gradually align themselves such that when the input pattern is presented, a neighbourhood of units responds to this input pattern forming a map area that is selectively optimised to represent the average of the training data for that class. As the training proceeds in figure 2.2.c smaller numbers are updated until at the end of the training only the winner node shown in figure 2.2.d is adjusted. The winner is the output unit

whose incoming connection weights are the closest, in terms of Euclidean distance, to the input pattern given by minimising the following equation,

$$\mathbf{C(x) = arg\ min\{||x - m_i||^2\}} \qquad \text{(Eqn 2.1)}$$

where x is the input, m_i is the reference vector it has to match and $C(x)$ is the winning node for the input pattern x.

When an input pattern is presented to the SOM network it does not assume any initial functional form; rather it lets the map units in the output layer compete with each other to match the input pattern and be declared the winner. It should be noted that weight vectors have the same dimensionality as the input patterns. The winner node gets adjusted but also the weights of the adjacent output units in close proximity of the neighbourhood of the winner are adjusted, moving them closer to the input pattern and also allowing them to learn. The degree of adaptation is guided by the learning rate factor $a(t)$, which can be initially selected close to unity and decreases as training progresses. The learning factor can even decay with the distance from the winning output. The units are chosen by means of a neighbourhood function $h_{ci}(t)$ which is based on the distances from the winner node as measured in the 2D grid formed by the SOM. Combining these principles of SOM training, the reference vectors $m_i(t)$ are changed according to the following adaptation rule [Koh89],

$$\mathbf{m_i}(t+1) = \mathbf{m_i}(t) + \alpha(t)\mathbf{h_{ci}}(t)[\mathbf{x}(t) - \mathbf{m_i}(t)] \qquad \text{(Eqn 2.2)}$$

where $x(t)$ represents the current input pattern at iteration t, $a(t)$ is the leaning rate such that $0 < a(t) < 1$ decreasing with t, and $h_{ci}(t)$ is the neighbourhood kernel function of the winning unit $C(x)$, defined by the neighbourhood radius and the distance between the map units, also decreasing with t. Figure 2.3 presents a brief structure of the pseudo code for the Kohonen SOM.

The Kohonen Algorithm

1. Initialise weights randomly
2. Set topological neighbourhood function $h_{ci}(t)$
3. Set learning rate parameters $a(t)$
4. While stopping condition is false, do
   For each reference vector m_i do
   For each x, compute do
   $C(x) = arg\ min\{||x - m_i||^2\}$
5. Find index x that minimises $C(x)$
   For all units x within the neighbourhood:
   $m_i(t+1) = m_i(t) + a(t)h_{ci}(t)[x(t) - m_i(t)]$
6. Update learning rate
7. Reduce radius of topological neighbourhood
8. Go to step 4 until stopping condition satisfied

Figure 2.3 – Pseudo code for the Kohonen SOM

Figure 2.3 gives a brief structure of the pseudo code for the Kohonen Algorithm. The network organises the nodes in the grid into local neighbourhoods $h_{ci}(t)$ by a cyclic process of comparing inputs to reference vectors at each neurone. The amount they learn will be governed by these local neighbourhoods which are a decreasing function of the distance from the winning output on the map lattice. As training progresses the size of the localised neighbourhood around the winner becomes smaller with weights between input vectors and output neurones representing a typical or prototype pattern for the subset of units that forms a particular cluster. Consequently, through this process the SOM algorithm discovers the number of clusters formed and identifies their location in relation to the other clusters on the map. It can even classify vectors that it has never seen before as long as they are spatially close to the training nodes.

2.3 Overview of SOM Design Parameters

The previous section depicted how the Kohonen algorithm takes randomly organised set of nodes and produces a feature map that has local representation and is self-organised. In this section we overview the SOM design parameters from setting up the goals and deciding the design parameters to the criteria for selecting the best maps.

2.3.1 Main Features

The SOM learning process is a stochastic process that shows variations in its mappings, [DebKo98] advise to compute several mappings in order to ensure good quality. Maps are selected according to how well they describe and preserve the location of the input data and according to the distinct SOM features that best translate the implicit knowledge in the input dataset of the selected application. These distinct features that assist the SOM to visualise and classify the data inputs include the neighbourhood function and learning rate, neighbourhood preservation parameters and clustering quality measures.

Training the Map

There is a selection between training the map sequentially i.e. following the main process described above and using batch training. The batch training algorithm comprises two phases: the first phase employs a relatively large initial learning rate α_0 and neighbourhood radius σ_0 whereas in the second phase both learning rate and neighbourhood radius are small from the beginning. This research uses the batch training algorithm because it is nearly 10 times faster [**batch** (***0.4s***) vs **sequential** (***4s***) (100 map units)] and presents the whole data set to the map whereas the sequential training algorithm selects a single input vector at a time at random. In each training step the data set is partitioned according to the Voronoi regions of the map weight vectors, so each vector belongs to the data set of the map unit to which it is closest. The sum of the vectors in each Voronoi set is calculated as follows [Ves97]:

$$S_i(t) = \sum_{j=1}^{nvi} x_j \qquad \text{(Eqn 2.3)}$$

where nv_i is the number of samples in the Voromoi set for neurone i. The new values of the weight vectors can be calculated as:

$$M_i(t+1) = \sum_{j=1}^{m} \frac{h(t)_{ij}\, s(t)_i}{nv_j h(t)_{ij}} \qquad \text{(Eqn 2.4)}$$

where $h_{ij}(t) = e^{-d_{ij}^2 / 2\sigma_t^2}$ is the Gaussian neighbourhood function (σ_t : neighbourhood radius, d_{ij} : distance between map units i and j), $s_j(t)$ is the sum of vectors in the Voronoi set for neurone j, nv_j is the number of data vectors that neurone j is the best match for and m_i is the number of model vectors.

Large maps are more likely to have each input vector attracted or characterised by a different neurone. They make good lookup tables for initial data representation in order to obtain the general clustering dynamics, but their generalisation ability may suffer. Smaller maps provide more data compression, however if they are too small may provide only very coarse differentiation [Kaski97]. The optimal number of neurones or size of a SOM is therefore a question of selecting the optimal granularity or abstraction of the data for the specific application. Once the training is completed the best map is then selected on its meaningful representation and its quality properties.

2.3.2 Meaningful Representation

Preservation of neighbourhoods and neighbouring map unit relationship is a key property for acquiring a meaningful mapping. The SOM consists of neurones forming a topographic mapping from the data space onto a 2-D output space. The neurones are connected to adjacent neurones by the neighbourhood function, which dictates the map topology. The algorithm acquires both projection and clustering methods which means that it does not only find the winning weight vector but also finds and updates its topological neighbours, creating the incentive for neurones on the grid that have similar weight vectors to merge into the same category, cluster [Kivi96].

The accuracy of local neighbourhood is improved by increasing the flexibility of the mapping by gradually reducing the radius of the neighbourhood function. Longer distances are not preserved well which can lead to topographical distortion. There are many ways of measuring topographical distortion, [Ults93] lists a number of them, but for this application we employ the simplest measure of topographic preservation which calculates the percentage of distortion in SOM, and therefore ranks the adjacent neurone's relationship, is such that for an input i, if the winning unit is not a neighbouring neurone of the node that has second smallest distance from input then there is a distortion in the map. Therefore the topographic error (t_{err}) is defined as:

$$t_{err} = \frac{\sum_{i=0}^{N} | r_{bi} - r_{sbi} | > 1}{N} \qquad \text{(Eqn 2.5)}$$

where r_{bi} and r_{sbi} are the positions of the best matching unit (BMU) or winner and the second BMU of input i on the SOM respectively and N is the total number of inputs and their modulus is the shortest link distances between them.

Although preserving all data distances is impossible it is crucial to preserve these local distances indicating local density in order to preserve the cluster structures. The density of the data is reflected in the density of the model vectors which is reflected in the local distances between neighbouring model vectors. This numerical nature of the method enables it to represent graded relationships. High values for the distances between the neighbouring map units indicate a cluster border whereas uniform areas with low values indicate the body of the cluster [Kivi96].

a. hexagonal lattice b. rectangular lattice

Figure 2.4: (a) Hexagonal; and (b) Rectangular Lattice Structures

The topology of the map is determined by two factors: *local lattice structure* and *global map shape* (sheet, cylinder or toroid). Figure 2.4 shows both hexagonal and rectangular structures with the innermost polygon corresponding to 0-neighbourhood and 1-neighbourhood. The lattice structure type influences the topographic quality of the SOM. In rectangular form, each node has four adjacent neighbouring nodes and the data samples are usually distributed in oblong format. In hexagonal form it has six adjacent neighbouring nodes and is preferred for visual inspection of the map since it does not favour horizontal or vertical dimensions [Kaski97]. In addition, Kohonen used a hexagonal halftone raster for images to show that when the number of lattice surrounding a neurone increased, the resolution of data encoding also improved.

2.3.3 Quality Scope of Clustering

The map quality is measured in terms of the training data [DebKo98]:

- *Number of Clusters* → If the map size is large, the number of clusters will be equal to the number of observations. Maps that are very small suffer from insufficient clustering tendency and are usually selected for data reduction.

- *Cluster Quality* → Low clustering tendency indicates low cluster quality with a small number of clusters or small differences between them. High clustering tendency indicates high cluster quality with several distinct clusters or with significantly dissimilar features.
- *Clustering Stability* → Tested on the basis of several out-of-sample data sets as well as several random-selected historical samples, some with more noise, some with less noise, some with and some without outliers.

The measure that assesses the quality of the map resolution by measuring the average distance between each data vector and its BMU is the *mean quantization error* q_{err} *of a single unit* and is computed by

$$q_{err} = \frac{1}{N} \| m_0 - x \| \qquad \text{(Eqn 2.6)}$$

where N represents the number of input data x and m_0 is the n-dimensional weight vector initialized with random values [Ves97].

The *mean quantization error* Q_{err} *of a map* is computed after a fixed number of training iterations t as:

$$Q_{err} = \frac{1}{u} \sum_i q_{err} \qquad \text{(Eqn 2.7)}$$

where u refers to the number of units i contained in the m, q_{err} is computed as the average distance between weight vector m_i and the input patterns mapped onto unit i.

2.4 Summary

This chapter looked at China's pre- and post- reform economic growth strategies, debated the effects of diverged growth between central, eastern and western provinces, and considered future prospects and challenges. It then discussed the need for a system that can assist in understanding and evaluating provincial growth policy directions. The Kohonen algorithm seems to be such a system since it provides a direct visualisation, translation and classification of high-dimensional data. Its clustering and projection features were discussed as well as its integration with exploratory data analysis and the parameter selection criteria.

Chapter 3

GP Bilateral Trade Flow Design

Chapter 3 overviews the past performance and future challenges of China's foreign trade focusing on the transition from a centrally-planned to a market-oriented economy, geopolitical dynamics between trading partners, and re-export channel's significance in bilateral trade. Genetic programming (GP) is used to investigate these factors that determine China's international trade environment by breeding populations of bilateral trading rules. GP properties including tree structure, selection of parameters and genetic operators are introduced.

3.1 China and International Trade

As new centres of economic activity emerge the management of trade relations between the high income developed countries and the emerging high growth developing economies in transition to a market economy is likely to become an increasingly central element in the future of international trade. One of these economies is China, which on the 11th of December 2001 became the 143rd member of the World Trade Organisation (WTO)[2]. This is a major issue that raises the question of how to deal with structurally different economies while trying to integrate them into the international market system [HoeKo95], [FuLa03].

[2] The World Trade Organisation (WTO) builds upon the organisational structure of tariffs (low tariff rates to member countries) and trade. It has 4 main functions including implementing the Multilateral Trade Agreements; providing a negotiation forum on trade issues; administering the understanding of dispute settlement; and co-operating with the World Bank and the International Monetary Fund (IMF) to achieve reliable policy-making.

3.1.1 Centrally-planned to Market-oriented

China has two main tariffs: the Most Favourite Nation (MFN) rate, which is applied to goods imported from countries having agreements of reciprocal tariff preference; and the General Rate applied to countries without such agreements. Despite its dual-track economy, its trade structure has been converging towards that of the top market economies since the early 1980's, when economic reform started. China's export similarity with the top market economies has increased due to its capital-intensive, industry-oriented development strategy. One important measure of the degree of international economic integration of a country is the fraction of world trade, for which it is responsible. Chinese trade has expanded more than twice as rapidly as world trade thus dramatically increasing its share and ranking among the world's top exporters [Lar94].

The agreement of accession to the WTO signalled that its open market reforms will continue firmly yet gradually integrating China into the global trading system. These reforms have changed trading partner's behaviour because China's trade does not solely depend on international trade due to large inter-provincial investments, making its domestic market the main beneficiary of the WTO agreement. China has gradually opened up to international markets through trade tests in several coastal provinces integrating itself into the international trade markets in a secure and efficient manner. Its large exports and even greater future export potential have made its trading system subject to unusually high levels of scrutiny in discussions in the WTO [Lar94], [Zhao97].

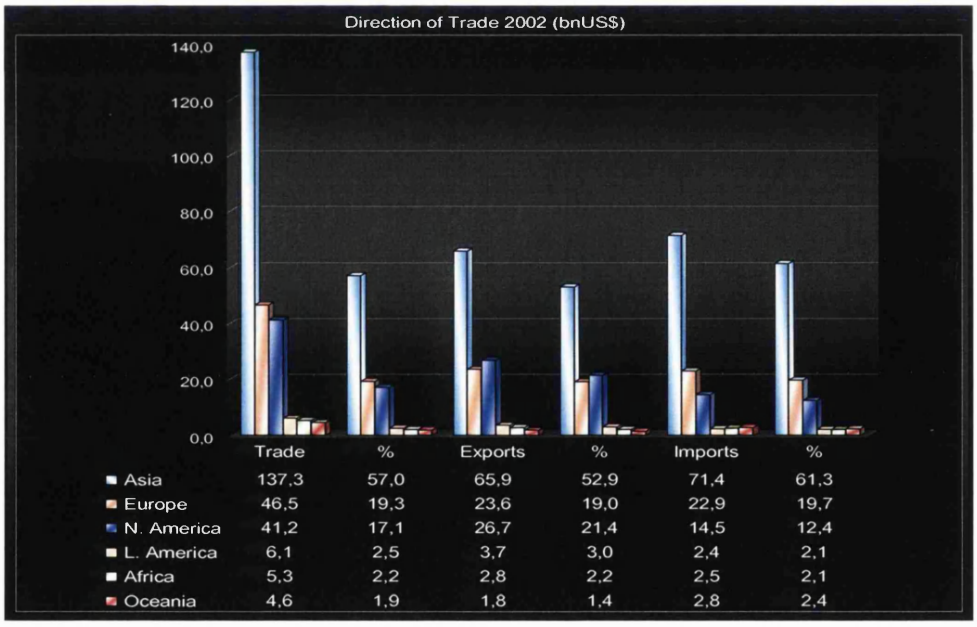
China's trade regime has evolved from the pre-reform era to the reform period though still retaining features from the past despite the points of resistance to the evolution of Chinese contemporary trade policies. The pre-reform Chinese trade regime was dominated by 16 Foreign Trade Corporations (FTCs) with effective monopolies in specified external trade product ranges with limited importance to conventional trade policy instruments such as tariffs, quotas and licences [IaMar01]. The continued reform in all areas in China has been the fundamental driving force of its economic success in the past and seems to be the most important factor of economic success in the intermediate future. Rapid growth of foreign trade provided the momentum for China's economic reform, making future trade performance of critical importance for the future of its economy as a whole [Lar94]. Factors that contributed to this unique trade performance include [Zhao97]:

- *The reform measures of the foreign trade sector including: i) reduction and ultimately removal of exchange rate distortion; ii) decentralisation and improved autonomy of foreign trade corporations; iii) gradual reduction of trade planning.*
- *The abundant supply of cheap labour making China one of the most important suppliers of labour-intensive products in the world market.*

- *The favourable policy toward foreign investment, starting from a mere 0.05% of total Chinese exports in 1980 to generating an impressive 41.4% in 2000.*
- *The necessary conditions for the rapid expansion of its exports including a welcoming environment for international trade.*
- *The development of indirect trade policy instruments (tariffs, licences, quotas and duty exemption schemes) that were absent under the planning system.*

3.1.2 Bilateral Trade Dynamics

Until the 1960's China's economy was highly dependent on increased exports of coal, petroleum, and petroleum products and had firm trade links with the Soviet Union; since then it fundamentally reoriented its pattern of trade. Since the reform period China has traded with countries all across the continents with Asia, Europe and North America accumulating the largest part of total foreign trade volume, 137.26bnUS$, 46.55bnUS$ and 41.1bnUS$ of total trade respectively as shown in figure 3.1. Data information for both figures 3.1 and 3.2 is provided by the Ministry of Foreign Trade and Economic Cooperation (MOFTEC) in China.



	Trade	%	Exports	%	Imports	%
Asia	137,3	57,0	65,9	52,9	71,4	61,3
Europe	46,5	19,3	23,6	19,0	22,9	19,7
N. America	41,2	17,1	26,7	21,4	14,5	12,4
L. America	6,1	2,5	3,7	3,0	2,4	2,1
Africa	5,3	2,2	2,8	2,2	2,5	2,1
Oceania	4,6	1,9	1,8	1,4	2,8	2,4

Figure 3.1 – China's Trade Dynamics across the Continents (Source: MOFTEC)

Figure 3.2 depicts China's top trading partners which include Japan, the United States (US), the European Union (EU), Hong Kong, the Association of Southeast Asian Nations (ASEAN), Taiwan, the Republic of Korea (ROK), Russia, Australia and Canada. China's country partners from the EU and ASEAN groups are included in the top trading ranks by measuring an aggregate of their trade volume for all countries in each of the two groups.

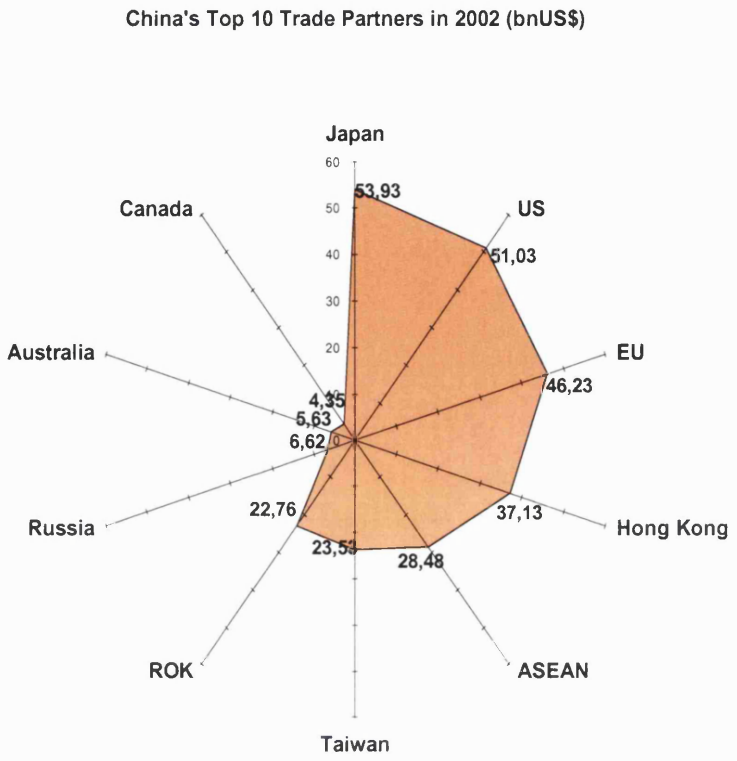


Figure 3.2 – China's trade volume with main trading partners (Source: MOFTEC)

Sino-Japanese Trade: Japan has been China's biggest trade partner for the past 7 years with Sino-Japanese economic and trade relations playing an essential role in making Japan the main source nation from which China absorbs foreign investment and introduces technologies. For over two decades, Japan has also been the biggest provider of governmental loan and gratuitous assistance to China accounting for 112bn¥ covering education, medical treatment, agriculture, and other areas. Figure 3.2 shows how Sino-Japanese trade sustains an increasing trend at the highest scale [MOFTEC].

Sino-US Trade: Trade with the US has caused bilateral trade frictions. The different structures of their economies together with the large trade deficit with China have been an obvious target of US trade policy. Bilateral trade imbalances often lie at the centre of trade disputes with the US being the only developed country that does not automatically offer Most Favourite Nation (MFN) treatment to Chinese products. If MFN status were withdrawn, the country's exports to the US would be subject to "3rd column" or "general rates", which are 5 to 10 times higher than US MFN tariffs. Imposing such high general rates would reduce exports to the US leading to a trade war that would be harmful for parties involved, making this possibility of MFN withdrawal

negligible. China heavily depends on the US market for its exports, whereas the US plays a less important role in supplying imports to China [Zhao97], [MOFTEC].

Sino-EU Trade: The EU countries group is third largest export market for China (after the US and Japan) and the second source of imports (after Japan). From the reform period to present trade has increased more than 30- fold and reached around 47 bn US$ in 2002. China is the third most important trading partner for the EU (after the US and Japan). Whereas the EU enjoyed a trade surplus with China at the beginning of the 1980s, their trade relations are now marked by a sizeable and widening EU deficit (around €47 bn in 2002). It has supported China's entry to the WTO and with a budget totalling € 37m for projects focusing on capacity-building in China's government and administration it ranks first among the contributors. Within the EU, the largest trade partner since 1975 is Germany with their bilateral trade accounting for 33.6% of total China-EU trade and Sino-German bilateral economic cooperation with German enterprises' actual investment hitting 9.17 bn US$ in 2002, lower only to the UK [MOFTEC].

Sino-ASEAN Trade: China is expected to continue being a powerful drive of growth, especially in East Asia since imports from ASEAN countries have grown over the past 10 years by 389% whereas total imports increased from under 6% in 1990 to 9% in 2000. The ASEAN countries that have benefited the most from China's growth in the 1990s include Singapore, Thailand, Malaysia and Indonesia. Nonetheless, according to World Bank experts Inachovichina and Martin [IaMar01], China's WTO accession will create the most significant implications for these countries, while providing the greatest mercantilist benefits to exporters in Taiwan and Japan followed by the industrialised economies and the newly industrialised countries in East Asia.

Sino-Taiwan Trade: Taiwan had prohibited contacts with China since their separation in 1949 following a civil war, nonetheless indirect bilateral trade flourished after Taipei allowed civilian exchanges with the mainland. Direct trade across the Taiwan Strait is currently banned and only 67.9% of a total of 8,282 industrial products and 23.1% of 2,092 agricultural items are allowed to be imported from China mainly via Hong Kong and in the last decade also by Guangdong and Fujian. Bilateral trade between Chinese coastal province Fujian and Taiwan rises by 23.36% on a yearly basis [MOFTEC].

Sino-ROK Trade: China-ROK bilateral trade in 2002 has increased by up to 8 times of the 1992 level when the two countries set up diplomatic ties that were mutually beneficial and reinforcing. Increased bilateral trade growth has left a trade deficit with Korea, since imports account for US$18.3bn of goods from the ROK, whereas exports accounted for US$9.2bn to Seoul in 2002. China is the second-largest export market for South Korea (after the US), and with the inclusion

of Hong Kong it becomes the largest export market for South Korea absorbing 20.8% of total exports. Furthermore, Korea's dependence on the Chinese exports market has escalated from 3% (1992) to 14.6% (2002) [MOFTEC].

Sino-Russian Trade: Russia is China's eighth biggest trading partner with bilateral trade volume growing for five consecutive years reflecting the rapid rise in economic co-operation. China is Russia's fourth largest trading partner. The two countries are starting to pursue joint projects in fields like technology and services as well as energy resources and chemicals. Trade along the border plays an important role in economic co-operation aiming to boost bilateral trade volume to US$20 bn by 2010 [MOFTEC].

Sino-Australian Trade: Since the establishment of diplomatic relations in 1972, Sino-Australian trade has impressively increased from A$158m in 1972 to A$17.9bn in 2001. Australia is the ninth largest trade partner and the largest supplier of major commodities, such as iron-ore, wool and aluminium, important for the manufacturing of high-grade steel and garments as well as electrical machinery and appliances, and telecommunications equipment. Australia's imports of goods were valued at A$10.3bn in 2001, an increase of 13.7% year-on-year. China is the third largest source of imports such as textiles, clothing, toys, games, sporting goods, footwear and computers [MOFTEC].

Sino-Canadian Trade: The Canada-China partnership encompasses trade and economic exchanges, defence relations, sustainable development and legal cooperation. Canada is the tenth largest trading partner whereas China is the third largest partner for Canada. Canadian firms are establishing vital commercial and technological linkages with Chinese companies, including fashion and clothing, telecommunications, and life insurance [MOFTEC].

3.1.3 Re-export channel significance – Hong Kong

As Chinese export specialisation changed to more differentiated, labour-intensive manufactured goods sold on diverse consumer markets, the crucial role of Hong Kong as an entrepôt and the main exporting channel for Chinese goods increased exponentially. Chinese products were exported to an indirect channel of distribution, Hong Kong serving as the intermediary import distributor. Between 1979 and 1997 Hong Kong grew faster than any single country and by 1995 Chinese exports accounted for more than 70% of its trade [HuBr98], [SuYW91].

China is the most important source of goods re-exported through Hong Kong (57% of total re-exports). The criterion used to determine the nationality of a product is determined by rules of origin necessary when there is a desire to discriminate between sources of supply. Thus, Chinese

re-exports are shipped first to Hong Kong with large proportion of these re-exports being products of outward processing commissioned by Hong Kong companies in China. This large volume of re-exports via Hong Kong and the different treatment of these re-exports by China and its trading partners statistical agencies accounts for the large discrepancies between their respective trade figures. After 1997 a separate Hong Kong was integrated with China according to Sino-British agreements, however keeping the "one country, two systems" formula, Hong Kong remains a separate customs territory and a separate member of the WTO consequently leaving the complicated trade data discrepancies unresolved [FuLau98], [FuLau03].

China's trade structure has been converging towards that of the top market economies with trade regime reforms tailored to facilitate a new centre of economic activity emerging from a high growth developing economy. Due to its structurally diverse economy its integration into the international trade system while attaining geopolitical equilibrium with its partners seems challenging, since these diverse actions also affect their trading behaviour. Trading partners under- or over- traded behaviour related to these actions is investigated in this study by trying to find a mathematical relationship that expresses China's continuously evolving bilateral trade environment. Instead of finding the coefficients determining a particular economic model, Genetic Programming (GP) is employed to genetically breed populations of candidate trade rules that best fit the data from an infinite information search space while highlighting the extent to which each variable influences the form of the equation.

3.2 From Natural Adaptation to Computer Program Evolution

Genetic programming (GP) [Koza89] is an evolutionary algorithm technique that automatically generates and evolves computer programs which undergo adaptation without being explicitly programmed. *Evolutionary Algorithms*, based upon Darwinian evolution, is a *stochastic search* technique which tests millions of species in parallel in order to maintain the extraordinary diversity and complexity seen in the biosphere. *Stochastic search* techniques guide their probabilistic choices using information from their search. *Simulated annealing* which searches for minimum energy states mimicking the physical annealing process, where specific metals are heated and cooled down to form large soft low energy crystals is another stochastic search technique. Other established search techniques include *enumerative* and *calculus-based* techniques. *Enumerative* techniques are simple to implement but search every single possible point. *Calculus-based or hill-climbing* techniques treat the search space as a continuous multi-dimensional function, estimating new maximums until they reach a hill top. They are divided in *direct* - using the function values to estimate the location of nearby extremes - and *indirect* - using the fact that at the extreme the function's derivative is zero [Lan98].

3.2.1 Fundamentals of Evolutionary Computation

Evolutionary Algorithms form a family of differing styles of computing approaches based on evolutionary adaptation that search for "solutions" given all the possibilities. They evolve towards increasingly better regions of the search space through the use of randomised processes of selection, crossover and mutation. The environment delivers a fitness value for the new search points and the selection process favours the reproduction of the fittest individuals. Three decades of research [KinJ97], have shown that mimicking the search process of natural evolution can yield very robust computer algorithms, although these imitations are just simplifications of biological reality. There are four types of evolutionary algorithms distinguished by the different types of structures which comprise the individuals in the population, the allowable genetic variation and the selection procedures of the genetic operators used to create offspring. These four types are:

- *Evolution strategies* [Rech73], is a method introduced by Rechenberg which is frequently associated with optimising real-valued parameters for hydrodynamic problems. Task specific routines are executed using fitness-determining parameters and recombination is employed for both objective and strategy variables.
- *Evolutionary programming* [FoOW66], is a technique created by Fogel, Owens and Walsh in which finite-state machines are evolved by randomly mutating - mutation is the sole genetic operator employed - their state-transition diagrams and selecting the fittest solution.
- *Genetic algorithms* (GAs) [Holl75], [Gold89], are adaptive methods based on the genetic process of biological organisms used to solve search and optimisation problems. Fitness is determined by executing task specific algorithms with sexual recombination being the principal genetic operator and mutation included as a secondary importance operator.
- *Genetic programming* (GP) [Koza89], is a technique developed by John Koza and inspired by genetic algorithms where computers try to program themselves by evolving computer programs. GP is discussed analytically in the following section.

3.2.2 Overview of Genetic Programming

Genetic programming (GP) [Koza89] automatically generates and evolves computer programs which undergo adaptation without being explicitly programmed. It is a descendent of genetic algorithms (GAs) invented by John Holland [Hol75] in order to study the phenomenon of natural adaptation and develop techniques that imported important features of natural evolution and inheritance into computer systems. Genetic algorithms are mathematical algorithms that mimic this procedure of natural selection by randomly generating individual bit strings analogous to the four nucleotide bases (adenine, cytosine, guanine, or thymine) found in nature which is stored on strands of DNA [BoHSS94], [Smith98]. Natural systems reproduce either by just creating the

child from a copy of the parent's DNA with some random mutations (***asexual***) or by inheriting the DNA from both parents and then copying half of each parent's DNA and joining the two sections to create the new individual (***sexual***). The fitter individuals survive to reproduce and pass on their DNA to subsequent generations, increase their proportion in the population and consequently evolve the nature of the species as a whole. In GP the individuals are tree-structured computer programs with new programs produced by removing branches from one tree and inserting them into another creating a new syntactically valid tree. This feature of the GP technique is particularly interesting since it allows the size and complexity of candidate solutions to increase over evolution, rather than keeping it fixed as in the standard GA [LaPol02]. Figure 3.3 shows an example of such a tree-structured computer program.

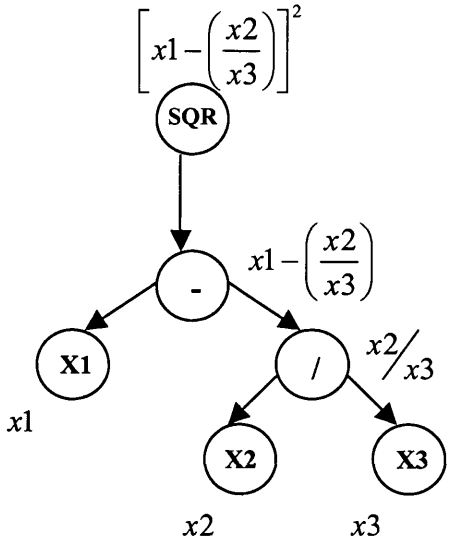


Figure 3.3 – $\left[x1-\left(\frac{x2}{x3}\right)\right]^2$ depicted as rooted point-labeled with ordered branches.

The key features of the GP structure are summed as follows:

- *Nonlinear Tree Structured Genetic Material*: Conventional GAs operate on linear genetic material whereas GPs generally operate on nonlinear tree structured genetic material.
- *Variable Length Genetic Material:* GP genetic material varies in size with considerable growth allowance from the original randomly generated generation but also with size limitations on its growth for practical reasons.
- *Executable Genetic Material:* GP is the direct evolution of computer programs so its evolved genetic material is generally executable. The genetic material is usually executed by some sort of interpreter in order to perform the desired function to obtain the fitness.
- *Syntax Preserving Crossover:* Crossover operators in GP are defined so as to preserve the syntactic correctness of the program, i.e. genetic material [Kinn94].

3.2.3 GP Preliminary Steps

[Koza94] describes genetic programming in terms of six components that assist GP to evolve programs that are difficult to write, these are:

- *Selection of Terminals and Functions*
- *Designing the Fitness Function*
- *Selection of Control Parameters*
- *Selection of the Termination Criterion*
- *Determining the Program Architecture*

Selection of Terminals and Functions

Terminals and functions are the components that ensure computer programs are syntactically correct to sufficiently express the problem solution and are able to evolve to a final solution. The more effective and informative functions and terminals are, the better defined the problem search space will be and the easier it will be for the GP system to find a credible solution. The search space boundaries should employ a wide set of functions and terminals though not too many since a change in the number can differentiate the difficulty factor for the GP system. Syntactically incorrect or semantically invalid individuals that are extracted from the fitness test evaluation can reduce the population diversity [Kinn94]. These two components employ a design property, *closure* that influences the choice of program components, problem representation and complexity. In closure, each of the functions accepts as its arguments any data type returned by any function or terminal, producing a GP result that is syntactically correct and executable. Functions $F = \{f1, f2, \ldots, fN_{fc}\}$ are formed in the junctions or *internal nodes* of the tree and terminals $T = \{l1, l2, \ldots, lN_{tm}\}$ at the end or *leaf nodes*, with the tree pointers connecting them indicating the order of evaluation and the brackets depicting the tree structure. The function set may include mathematical and arithmetic functions (+, -, *, /, sin, cos, exp, log); Boolean operations (AND, OR, NOT); and any other domain-specific functions. The terminal set is either comprised of variable or constant atoms [Koza89].

Designing the Fitness Function

The fitness measure causes the creation of structure via natural selection and the creative effects of genetic operations [Koza92a]. The fitness function guides the evolution of the GP population by crediting fitter individuals with high scores and also partially crediting improved solutions throughout the GP run, from the creation of the initial population to the discovery of the final solution [Lan98]. Favourably crediting an intermediate partial solution over another defines the direction of the evolutionary process encouraging individuals in the population to move in that direction. In partial crediting, the granularity of the fitness function is important since a fitness function with smaller amounts of information may not lead towards the solution. Furthermore,

careful consideration must be given to which sorts of programs would produce each partial solution and certifying that the correct program gets the better score. A set of fitness cases is run for which the correct output is known [Koza92a], [Lan98], [Kinn94]. The four fitness measures widely used are; raw fitness; standardised fitness; adjusted fitness; and normalised fitness.

Raw fitness: Raw fitness is the real-valued measurement of a problem evaluated over a set of fitness cases typically a finite sample of the entire domain space. The representative sample must be sufficiently large to include a number of different situations sufficiently so that a range of different numerical values can be obtained that forms the basis for generalising the results. In general, the raw fitness is the sum of the distances over all cases between the point in the range space returned by a tree-structured program for the set of arguments associated with the particular fitness case and the correct point in the range space it associates with. The raw fitness *r(i,t)* of an individual expression *i* in the population of size *N* at any generational time step *t* is [Koza92a]:

$$r(i,t) = \sum_{j=1}^{N_e} | S(i,j) - C(j) | \qquad \text{(Eqn 3.1)}$$

where *S(i,j)* is the value returned by expression *i* for fitness case *j* (of N_e cases) and where *C(j)* is the correct value for fitness case *j*.

Standardised Fitness: The standardised fitness *s(i,t)* restates the raw fitness so that a lower numerical value is always a better value. Standardised fitness equals raw fitness (*s(i,t)* = *r(i,t)*) if a smaller value of raw fitness is better, whereas for a greater value of raw fitness, standardised fitness is obtained by subtracting the observed raw fitness from the maximum possible value of raw fitness r_{max} as follows [Koza92a],

$$s(i,t) = r_{\max} - r(i,t) \qquad \text{(Eqn 3.2)}$$

Adjusted Fitness: The adjusted fitness measure *a(i,t)* is computed from the standardised fitness *s(i,t)* as follows:

$$a(i,t) = \frac{1}{1+s(i,t)} \quad \alpha \in [0,1] \qquad \text{(Eqn 3.3)}$$

where *s(i,t)* is the standardised fitness for individual i at time t.

Since the adjusted fitness is bigger for better individuals it exaggerates the importance of small differences in the value of the standardised fitness as it approaches zero and especially when it reaches zero and thus has found a perfect solution to the problem. Thus, as the population improves, greater emphasis is placed on the small differences that distinguish a fit from a fitter individual [Koza92a].

Normalised Fitness: The normalised fitness *n(i,t)* is computed from the adjusted fitness value *a(i,t)* as follows:

$$n(i,t) = \frac{a(i,t)}{\sum_{k=1}^{M} a(k,t)} \qquad \text{(Eqn 3.4)}$$

n(i,t) ranges between 0 and 1 and is larger for better individuals in the population [Koza92a].

Selection of Control Parameters

Genetic programming has 19 control parameters, 2 major and 11 minor numerical parameters, and 6 qualitative variables of executing a run. The two major numerical parameters are population size and maximum generations in a GP run. The population size must be larger than a critical minimum size in order to generate a reliable solution. Most GP populations are smaller than their optimum, constrained by the available machine resources [Koza92a] but do operate with large population sizes compared to other evolutionary computation techniques. If the population size determined is preventing a rapid solution either the population size is increased or the number of runs on a given population size is increased aiming for at least one successful run finding the solution to the problem. Multiple runs are another effective manner to utilise much larger populations though not the same as running a larger population. Nonetheless, increasing the population size beyond optimum size may slow down the GP process from producing a solution [Kinn94].

The 11 minor numerical parameters used to control the process are; crossover probability; reproduction probability; probability distribution of crossover points; maximum depth size for expressions created by genetic operations in a given run; maximum depth size for the random individuals generated for the initial population; mutation probability specifying the frequency of performing mutation; permutation probability; editing frequency; encapsulation probability; decimation condition. The 6 qualitative variables that select different ways of executing the runs are; initial random population ramped half-and-half; fitness-proportionate reproduction method; same method for first and second parent crossover selection; optional adjusted fitness measure; over-selection for populations of 1,000 and above; and not using elitist strategy [Koza92a].

Selection of the Termination Criterion

The termination criterion stops the evolution of genetic programs when either an exact or approximate solution is found or when a specific number of generations is reached. From observation the GP seems to run out before generation 50 and continuing the run only marginally increases the chance of finding a solution. [Koza92a] argues that sometimes it is more effective to run a GP several times than increase the number of generations used in a run. The runs in this thesis terminate either when an individual passes the whole of the fitness test case or the maximum number of individuals have been created, so usually around 50 generations.

Determining the Program Architecture

Computer programs select and employ skillful techniques to decompose a problem in separate modules and then combine these modules to construct a program which solves the original problem. Despite the fact that GP has no such skill, it has been successful in incorporating a degree of modularity into its systems with two distinct approaches:

- *Automatically Defined Functions* are evolvable subroutines within a genetic program consisting of terminals and the same functions, preserving the overall format of the program by ensuring crossover and other genetic operations acts only within each ADF.
- Encapsulation is an asexual process operating on one parent tree expression and automatically identifying potentially useful subtrees for future reference. The selected subtree is an inseparable single point no longer subject to potentially disruptive effects of genetic operations and is allowed to reproduce in future generations [Koza92a].

3.2.4 Genetic Operators for Modifying Structures

Evolutionary algorithms employ different techniques to decide which individuals will reproduce, the number of offspring as well as which individuals will be removed from the population, these are divided in primary and secondary operations. The two primary operations are reproduction and crossover.

Reproduction

Evolutionary algorithms decide which individuals will reproduce by rewarding better solutions with more offspring. It is important to determine the amount good individuals are rewarded since this allows it to reproduce several times decreasing the genetic diversity of the population while equally rewarding every individual lessens the selection pressure on the population to evolve in the desired direction. The reproduction operation is asexual in that it operates on only one parental expression and produces only one offspring expression on each occasion and consists of two steps; a single expression selected from the population according to fitness and the selected individual copied, without alteration, from the current population into the new population. If

$f(s_i(t))$ is the fitness of individual s_i in the population s_i at generation t, then, under fitness-proportionate selection, the probability that individual will be copied into the next generation of the population as a result of any one reproduction operation is

$$\frac{f(s_i(t))}{\sum_{j=1}^{N} f(s_j(t))} \qquad \text{(Eqn 3.5)}$$

Typically, $f(s_i(t))$ is the normalised fitness $n(s_i(t))$ so that the probability that individual s_i will be copied into the next generation of the population as a result of any one reproduction operation is simply its normalised fitness $n(s_i(t))$. The reproduction operation performed by fitness-proportionate selection is called fitness-proportionate reproduction [Koza92a], [Kinn94].

Tournament Selection: Various schemes have been used to rescale fitness values so that the number of expected offspring is within reasonable range, including doubling the rescaled fitness of the best member of the population compared to the worst one; or ranking candidate parents and using this fitness ranking to determine the number of new offspring. These schemes use the whole population and perform well in small concentrated populations but become onerous when maintaining global fitness data for selection in dynamic environments. Tournament selection is a stochastic reproduction method that does not use the whole population and performs rank selection only on local population statistics with an element of noise due to random selection. This method randomly selects a number of individuals from the breeding population, comparing them with each other and selecting the best one [Koza92a].

Steady State Populations: Traditional GAs [Holl92] evolve a sequence of discrete non-overlapping generations mimicking species of plants and animals that live no longer than a year, germinating from seeds in spring and growing during the summer and producing their own seeds in the autumn which survive during winter while their parent dies. In contrast there are species that live many years with no distinct boundary between generations. In steady state populations new children are continually added to the population that can be selected as parents for new offspring and can also be removed from the population at any point ensuring a constant size for the population. GPs mimic this process of the creation of dynamic (steady state) populations [Koza92a].

Crossover

The crossover operator consists of choosing and exchanging the subtrees beneath the randomly selected points of two tree-structured computer programs (typically of unequal size) to produce

two offspring allowing the program size to increase or decrease. Initially a random point in each parent is selected as the crossover point for that particular parent which is a rooted subtree consisting of the entire subtree below the crossover point [Koza92a]. Figure 3.4 shows how offspring 1 and 2 are produced by deleting the respective crossover fragments and inserting the crossover fragment (red line) of parent 2 at the crossover point (blue line) of parent 1 and vice versa.

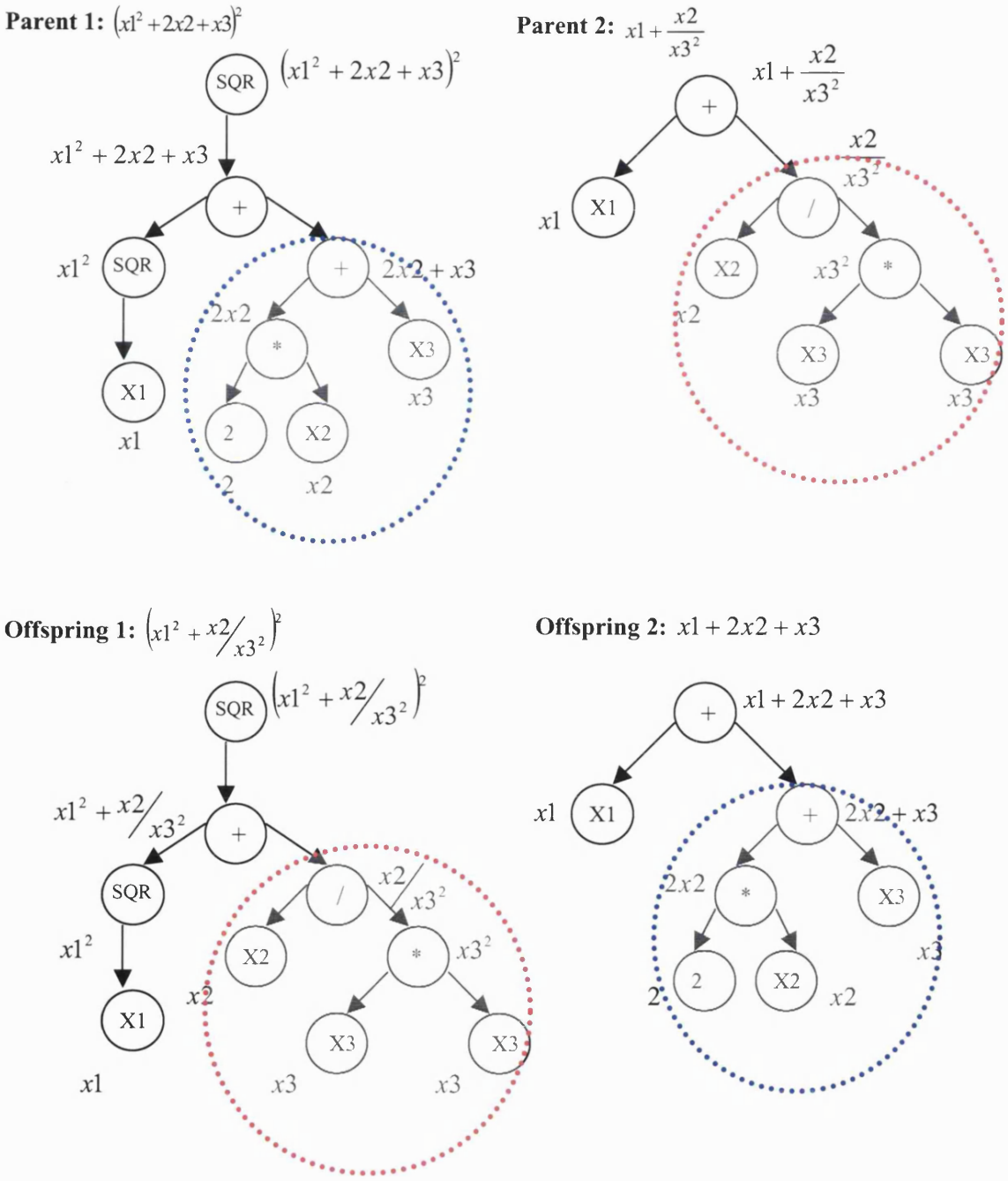


Figure 3.4: Offspring 1 and 2 produced by crossover of fragments selected from Parent 1 and 2.

When the first parent's crossover point is a terminal then the subtree from the second parent is inserted there and the terminal from the first parent is inserted at the location of the subtree in the second parent, whereas when terminals are located on both crossover points they are just swapped from one tree to another. If the crossover point for one parental expression is a root the entire first parent will become a subtree within the second parent making the crossover fragment of the second parent the second offspring. In traditional GAs crossover produces many copies of a particular individual with exceptional fitness relative to others increasing the tendency towards premature converge achieved by convergence to a globally suboptimal result. In contrast, in GP when two individuals mate the two resulting offspring differ due to different crossover points. The reproduction operation does create a tendency toward convergence however it also exerts a counterbalancing pressure away from convergence. A permissible size limit measured via tree depth is established for offspring created by the crossover operation that prevents the expenditure of computer time on a few extremely large individual expressions [Koza92a].

There are other crossover operation methods proposed, these include; context preserving crossover and one-point crossover. *Context Preserving Crossover* means that the subtree taken from one parent and inserted into the offspring is inserted in a similar position to the one it occupies in the parent, thus preserving the program syntax. Context is defined in terms of tree geometry. It is divided in; *Strong Context Preserving Crossover (SCPC)* and *Weak Context Preserving Crossover (WCPC). One-point Crossover* [PolLan97] is a crossover operator, acquiring the same crossover points from both parental sections, which when combined with node replacement mutation acts as a linear GP.

Secondary Operations

In addition to reproduction and crossover there are five optional secondary operations, these are:

Mutation: This operation is performed on a single node of a tree-structured individual in an effort to make the candidate solution fitter. Figure 4.5 illustrates this operation. Koza omits the mutation operator from the GP runs since he argues that large initial populations contain sufficient diversity so that only the crossover operator is used. [O'Reil95] argues that mutation combined with simulated annealing or stochastic iterated hill climbing can perform as well as crossover. Recent comparisons of crossover and mutation have suggested an advantage of the mutation operator over crossover especially in modelling applications with Koza now allowing for low levels of mutation [Koza92a].

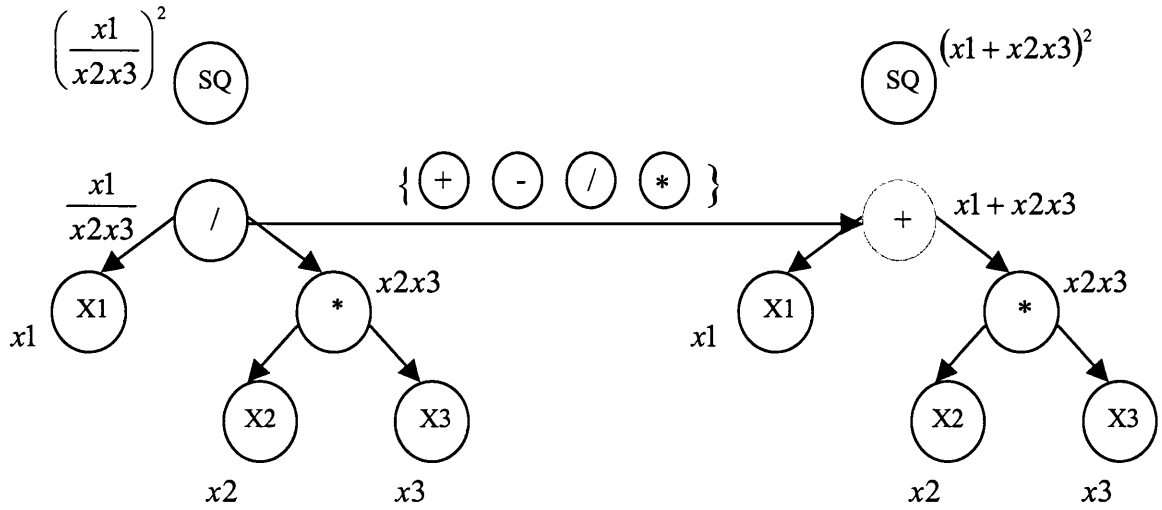


Figure 3.5 – Effect of the Mutation Operator on a single tree node from (/) → (+).

Point or *node replacement mutation* is similar to bit string mutation randomly selecting and changing a node in the tree, while ensuring that it has the same number of arguments as the replaced node. *Permutation* is another special mutation case where a mutation operator swaps the order of arguments of binary non-commutative, functions, rather than acting on any function. *Editing* is an asexual operation that operates on only one parental expression and is used to simplify expressions as genetic programming is running. The editing operation recursively applies a pre-established set of domain-independent and domain-specific editing rules to each expression in the population. *Decimation* deals with highly skewed distribution of fitness values resulting low variability and many unfit individuals by probabilistically selecting individuals on the basis of their fitness and disallowing reselection in order to maximise diversity in the remaining population [Koza92a], [Lan98].

3.3 Issues in GP Structures

The tree structure and fitness function largely determine the search space and consequently complexity and success of the GP search. Randomly generated tree-programs can be of different sizes with numerous numbers of nodes and levels. This section surveys these important issues.

3.3.1 Tree-Structure Representation

In order for the tree to be machine-readable we need to employ a method that will linearise the tree structure or put all nodes on one line. A tree traversal is employed for this process that visits each node in the tree exactly one time. GP work generally represents tree structures in pre-order depth-first traversal which proceeds as far as possible to the left (or right), then back up until the

first crossroad (root node), goes one step to the right (or left), and again as far as possible to the left (or right). Nonetheless, other methods have been used which are briefly outlined here. *Directed acrylic graphs (DAGs)* have been used for tree representation, since individuals descended from similar parents have common code thus a single DAG can hold the whole population by storing the common code once resulting in considerable memory reduction. *Linear representations* have also being applied [Banz93] but are converted to tree shaped programs when executed. *Parallel Distributed Genetic Programming (PDGP)* replaces program trees with directed graphs and although constrained by the predefined grid of nodes it is considered a powerful and flexible technique since the tree nodes have multiple outputs making the intermediate results readily usable in other parts of the program [LaPol02].

3.3.2 Effective Program Size

GP work typically uses a single fitness criterion ignoring size and run time issues. Program size can affect fitness since many modelling problems require solutions which fit the data and are short. GP solutions tend to grow in length, thus a parsimony pressure is included in the fitness function to encourage the evolution of shorter solutions [LaPol02]. Program growth is often dealt with by placing either a maximum tree depth or by constraining the GP to small programs in order to increase its performance. [Kinn94] adds a term inversely proportional to the program's length to its fitness generating shorter more general programs. [Bl96] shows that application of parsimony pressure methods on symbolic regression problems can be accomplished with the adaptive parsimony fitness component giving the best results while [GatRo97] use program size to resolve ties when programs have equal fitness.

3.3.3 Program Efficiency

GP determines an individual's fitness by running it. This makes it susceptible to syntactically invalid individuals that can effectively halt the whole GP system. Consequently recursive or iterative features are restricted in GP since they can promote very long or even infinite loops. [Cr85], [TelVel95] address the problem by implementing ad-hoc limits which give poor fitness scores to invalid individuals. [Koza92a] applies a limit on both the total number of iterations and each loop primitive. [MaxIII94] applies an external time limit that interrupts programs if they are still running giving a partial fitness while allowing other population members to be run. Programs which remain in the population are allowed to run for another time interval whereas looping programs are initially given low partial fitness and eventually removed from the population.

3.3.4 Genetic Diversity

The number of individuals contained in a GP population is usually small compared to the problem search space. Holland showed that by operating on fixed-length character strings the GA

processes information about an enormous number of unseen hyper-planes or *schemata*. For each generation an estimated value of the average fitness for each schema is computed. Although reproduction and crossover operate on the number of individual present in the population, implicit computation operates on a considerable number of schemata [Poli01].

A GP schema is the set of all individual rooted, point-labelled tree expressions that share common features containing one or more specified subtrees. There is a size limitation for both of the initial random tree size and the tree growth size resulting from crossover and defined by the total number of points in the tree. The average fitness of the GP schema is the average of the fitness values of all individual trees belonging to it. Subtrees from relatively compact high-fitness individuals are used as building blocks for constructing new individuals. [Tack93] shows how GP building blocks (subtrees) are repeated frequently throughout the population. [Kinn94] adds that if building blocks have frequencies independent of position then their distributions will shift, and might result with entirely different sets of building block subtrees.

3.3.5 Mathematical Proof of GP Evolution

[Lan98] investigated the mathematical proof of how GP populations evolve depicting that Price's Covariance and Selection Theorem [Price70] can be applied to artificial evolution.

Price's Selection and Covariance Theorem: The population genetics theorem relates the change in gene frequency from one generation to the next, to the covariance of the gene's frequency in the original population with the number of offspring produced in that population. The frequency of the gene in the current population is given by

$$Q_{freqi} = \frac{\sum g_i}{N} = \frac{\sum q_i}{N} = \bar{q} \qquad \text{(Eqn 3.6)}$$

where $\mathbf{Q}_{freqi}$ is the gene frequency, g_i are the gene copies in an individual i, N is the size of the initial population P_i, q_i is the individual's gene frequency and $\bar{q}$ is the arithmetic mean of q_i in the population. The number of genes in the new population equals the number of successful crossover fragments produced by the former generation. The number of chromosomes in an individual equals the sum of the number in each of the fragments that formed it. Thus, the gene frequency in a population in the next generation is given by

$$Q_{freq_j} = \frac{\sum g_i'}{\sum z_i} = \frac{\sum z_i q_i'}{\sum z_i} = \frac{\sum z_i q_i'}{N\bar{z}} \qquad \Rightarrow$$

$$= \frac{\sum z_i q_i}{N\bar{z}} + \frac{\sum z_i \Delta q_i}{N\bar{z}} \Rightarrow$$

$$= \frac{\sum((z_i - \bar{z})(q_i - \bar{q}) + \bar{z}q_i + z_i\bar{q} - \bar{z}\bar{q})}{N\bar{z}} + \frac{\sum z_i \Delta q_i}{N\bar{z}} \Rightarrow$$

$$= \frac{\frac{1}{N}\sum(z_i - \bar{z})(q_i - \bar{q}) + \bar{z}\frac{1}{N}\sum q_i + \bar{q}\frac{1}{N}\sum z_i - \frac{1}{N}\sum \bar{z}\bar{q}}{\bar{z}} + \frac{\sum z_i \Delta q_i}{N\bar{z}} \Rightarrow$$

$$= \frac{\frac{1}{N}\sum(z_i - \bar{z})(q_i - \bar{q}) + \bar{z}\bar{q} + \bar{q}\bar{z} - \bar{z}\bar{q}}{\bar{z}} + \frac{\sum z_i \Delta q_i}{N\bar{z}} \Rightarrow$$

$$= \frac{\frac{1}{N}\sum(z_i - \bar{z})(q_i - \bar{q}) + \bar{q}\bar{z}}{\bar{z}} + \frac{\sum z_i \Delta q_i}{N\bar{z}} \Rightarrow$$

$$= \frac{Cov(z,q)}{\bar{z}} + \bar{q} + \frac{\sum z_i \Delta q_i}{N\bar{z}} \Rightarrow$$

where Q_{freq_j} is the gene frequency in the next generation, g_i' are gene copies in all successful fragments in the next generation produced by ***i***, z_i is the number of offspring, q_i' is the gene frequency in the offspring produced by ***i***, Δq_i is the difference between the gene frequency in the offspring produced by ***i*** and the gene frequency in ***i*** and $\bar{z}$ is the mean number of offspring produced. From the above equations we conclude that:

$$\Delta Q = \frac{Cov(z,q)}{\bar{z}} + \frac{\sum z_i \Delta q_i}{N\bar{z}} \qquad \text{(Eqn 3.7)}$$

Consequently, selection for reproduction is dependent upon fitness and on specific genes whereas determination of crossover points is random and independent of genes.

3.4 Summary

This chapter focuses on the structurally diverse Chinese trading policies and how they will affect its integration into the international trade system. This integration is investigated by employing genetic programming to identify the mathematical relationship measuring trade flows between China and its trading partners. The preliminary GP steps are introduced together with the selection of primary and secondary genetic operators. Issues in GP structures are discussed including tree representation, effective program size, program efficiency, and genetic diversity. The mathematical proof of GP evolution using Price's Covariance and Selection Theorem to artificial evolution is also given.

Chapter 4

GTCA Part I: Intelligent Growth Policy Mapping

The focus of this chapter is concerned with the first part of the GTCA which explores the strategic importance of SOM-based design in providing direct visualisation and location of the numerous Chinese provinces according to their distinct competitive growth identities. The SOM has been applied to provincial growth self-organisation of a country with the complexity of China. The effect of sector-specific influence for the growth type of each province is also investigated together with the validity and analysis of the cluster formations. The SOM's ability to interpret and deal with inaccurate and missing elements from the early reform years is investigated.

4.1 From Growth Information to Growth Type Policy Formulation

One of the most interesting features when investigating a country's interior is the formation of inter-provincial economic growth groups emerging as competitors for preferential policies. This intensifying competition for resources and markets has forced provinces to differentiate themselves by developing distinct growth policies and identities. Although the necessity for differentiation has been realised, the selection of the underlying policies that precisely capture and promote province-specific growth in different sectors is not obvious. We employ the Kohonen SOM to analyse fundamental and technical growth determinants that affect the evolution of established and emerging growth policies by aiming to directly identify, visualise and locate the distinct growth identities of the Chinese provinces.

4.1.1 Integrating SOM-based Data Analysis

The analysis of a country's long-term growth policies and structure is an important process in macroeconomic development. Conventional statistical tools apply ranking to interpret the long-term growth factors producing very accurate results. However, some provinces have very different growth type structures but the same growth indicator volume making it difficult to rank them only according to volume. Since this study is interested in exploring as many of the different numerous factors that determine the growth features in order to characterise each province, it needs a knowledge discovery tool that it able to translate information from multiple criteria in parallel.

The SOM solves this problem since its projection and clustering capabilities provide feature classifications in a meaningful data-driven approach based on multiple criteria, extracted and represented in a graphic and intuitive form. These capabilities are explored and further developed by applying them to spatial (provinces) and temporal (annual) Chinese growth data. The SOM illustrates some structures in the selected data set, and the features chosen to represent them ultimately determine the structures. If there is some prior knowledge about the input, this information is usually used for choosing the features. If no explicit clusters exist, this self-organising mapping method reveals open zones with irregular shapes and high clustering tendency. The data sets selected as well as their collection and pre-processing are discussed in the following sections together with the methodology employed, the results from actual processing of the data and finally multiple growth dimensions and profiles of the provinces from computation of the maps.

4.1.2 Data Collection

Choosing the data set is a very crucial step because from it depend the quality and suitability of the inputs. This study has collected long-term growth indicator information on 27 provinces over the reform period (1978-1998). Finding Chinese provincial data for the reform period, especially the first decade, was considerably difficult for the following reasons:

- Due to the fact that no single library or organisation had a collection of all the statistical yearbooks together, especially for the first decade. The School of Oriental and African Studies (SOAS) has the best collection and we had to collect them one-by-one.
- The data for most of the reform period was not available in electronic form so it had to be processed and inserted into worksheets.
- Provinces give different definitions for the same indicator thus making it even more difficult to attain correct meaningful data and limiting the indicators that could be used.
- Most western and some central provincial yearbooks had no values for most indicators for the first decade of reform, either due to lack of expertise, since these provinces did not

have the privilege of statistics experts that the coastal provinces had, or because the values were negligible.

- The Chinese Provincial Statistical Yearbooks are in Chinese and since there was no immediately available translator, the symbols for all relevant indicators had to be learnt in Chinese.

In Appendix B, we provide an annual data collection for all provinces and all indicators used in this research through the entire reform period.

4.1.3 Data Pre-processing

When collecting data it is important to choose a set of indicators that describe the phenomenon of interest alone so that the corresponding dimensions of the input space can be scaled according to these features. If the collection and pre-processing of the data features is satisfactory enough to reflect the requirements of the task, then it will reflect the evaluation and quality of the results, thus helping to avoid or even eliminate misleading conclusions.

The data collected over the reform period (1978 – 1998) includes information on 27 provinces these are; Anhui, Beijing, Fujian, Guangxi, Guizhou, Gansu, Guangdong, Hainan, Henan, Heilongjiang, Hunan, Hubei, Inner Mongolia, Jilin, Jiangxi, Jiangsu, Liaoning, Ningxia, Shanghai, Shandong, Shanxi, Shaanxi, Sichuan, Tianjin, Xinjiang, Yunnan and Zhejiang. The Statistical Yearbook of China [StatCHN] as well as the Provincial Statistical Yearbooks were consulted from 1978 to 1998 for country and provincial figures, the Facts and Figures Book [FaFi] various years and the China Energy Yearbook [StatEN] for years 1978-1989 and 1991-1998. The key long-term growth sectors selected for this study include aggregate values (total output) for population, construction, industry, agriculture, area (land occupation) and energy. In order to measure inequality and growth distribution levels, per capita values have been computed for the above 6 variables. Many more sectors that highlight the different determinants that affect long-term growth could be inserted into the SOM model, however problems experienced collecting, evaluating and organising the data – discussed above – limited the number of high-quality sectoral datasets that could be included. Since this seems to be the first application on modelling macroeconomic policies, many more factors could be inserted and analysed in future work. In addition to the problems experienced while collecting Chinese data, some adjustments had to be made to the available data sets, such as:

- Tibet, Qinghai and Hebei provinces had to be excluded from the sample; Tibet is seen more as a religious autonomous region than an emerging growth province; Qinghai had negligible values in almost every sector; and provincial data for Hebei was not available.
- Coal, oil and hydro-power production were analysed separately and combined into a single energy measure as an aggregate of energy production because they are the largest

components of energy production. The China Energy Yearbook (1978-1989, 1991-1998) was consulted.

- The Gross Domestic Product (GDP) was excluded because it would have been difficult to scale this central and strong indicator in relation to the others, since it is an aggregate measure of all the other indicators selected.
- The industry indicator had different definitions for the different types of industry (small-scale and large-scale) for each province, thus we took an aggregate industry measure.
- An aggregate measure was also determined for the agricultural sector.

After collecting and selecting the appropriate datasets, they must be pre-processed before being applied to the SOM algorithm. The data is scaled, taking growth rates of each province for each sector, and balanced using 3-year "moving" averages to smooth out the impact on the index of year-to-year fluctuations of growth rates in particular sectors. The variance of all indicators is scaled to unity, to reflect their relative importance to the long-term growth measure. By scaling data values that can produce an error during training of the net, outliers can be discarded in advance. However, the outliers in the SOM map displays affect only one map unit and its neighbourhood, while the rest of the data display is still used for inspection. The outliers can be easily detected based on the clustering display since the input space near the outliers is very sparsely populated. Furthermore, some data items initially detected as outliers were actually not erroneous but just had strikingly different features which we discuss in section 4.3.

4.2 The SOM Methodology

This section outlines the multiple steps that comprise SOM-based data analysis from setting up the goals and deciding the SOM parameter criteria for selecting the best maps to analysing the clusters formed. After selecting and pre-processing the data source, variables, scope, and quality that are meaningful in relation to the objectives and likely to influence the results, it is necessary to devise a methodology to select the most suitable map for identifying growth type structure in provincial China.

4.2.1 Map Selection Criteria Revisited

The SOM learning process is a stochastic process that shows variations in its mappings. Consequently, several mappings are computed in order to ensure good quality. Maps are selected according to how well they describe and preserve the location of the input data and according to the distinct SOM features that best translate the implicit knowledge in the input dataset of the selected application. We use different initialisations, neighbourhood functions and learning rates in order to select the best maps, while ensuring the topology is preserved since it is the feature that most affects the SOM's performance. Selection of the desired display size, neighbourhood and learning rate, as well as map tuning is required for optimal clustering and visualisation.

Training Procedure

The amount the map learns is initialised linearly favouring a law of type $a(t) = a_0(1-T/t)$ [Koh89], where $a(t)$ is the learning rate, a_0 is the initial learning rate and T is a parameter advised to have the value of 100 times the number of neurones or more. The batch training algorithm is employed because it only takes a few seconds and presents the whole data set to the map before any adjustments are made. The Gaussian neighbourhood function is selected where $h_{ij}(t) = e^{-d_{ij}^2 / 2\sigma_t^2}$ is the neighbourhood function, σ_t is neighbourhood radius and d_{ij} is the distance between map units i and j.

Both large and small mappings are selected in this work in order to fine-tune the map to stretch out to the data input units and give the best possible interpretation of the data set. We initially train a larger map with a learning rate of 0.1 and a starting neighbourhood radius of 3 to obtain an initial data representation and the general clustering dynamics. Then we complement this by fine-tuning with a smaller neighbourhood radius of 1 and learning rate of 0.02 which provides sharper differences between clusters and better stability.

Topology Preservation

Preservation of neighbourhoods and neighbouring map unit relationship was introduced in Chapter 2 (Section 2.3) as a key property of the SOM. Map topology is determined as the connections to adjacent neurones by the neighbourhood function. In this research we employ the hexagonal lattice structure for visual inspection of the map since it does not favour horizontal and vertical dimensions. The latter topology map factor selected is the global shape which is in sheet format. The mapping selected is used in section 3.3 as the foundation for displaying the cluster structure of the data, so it is very important that the inputs are represented in a meaningful manner. In order to increase mapping precision we gradually reduce the radius of the neighbourhood function thus preserving the accuracy of local neighbourhoods. The distortion of the topology is measured by employing the topographic error t_{err} described in section 2.3.1. Data density reflects vector density and consequently the local distances between neighbouring model vectors. It is impossible to preserve all data distances however it is crucial to preserve these local distances between neighbouring model vectors in order to obtain a meaningful representation of the cluster structures. The topographic error for the maps selected was zero or negligible, which is really important since that denotes that the relationship between the growth data input is nearly perfectly represented. This will be discussed together with the computations of the maps in the discussion sections for each map analysis in section 4.3.

Quality Scope of Clustering

The objective in this study is to better understand the clustering dynamics of the input data set and small maps give only very coarse differentiations, thus larger ones are employed to represent

the Chinese data sets in the best possible manner. After selecting the maps according to meaningful representation, we assess their quality according to the mean quantisation error Q_{err} introduced in Chapter 2 (Section 2.3). Due to the difficulties experienced with the Chinese data and explained in earlier sections, some clusters experience only minor differences, whereas others are significantly dissimilar indicating high quality. The clustering stability is also assessed by using different data samples, some with outliers and some without in various map generations described and discussed in section 4.3.

4.2.2 Cluster Analysis

In the previous section the methodology used to select the most suitable map for identifying growth type structure in provincial China was decided. This section is concerned with the clustering output of these selections and how it affected the SOM's ability to detect distinct features and preserve the location of the input data. The number of clusters formed is analysed together with their differences with neighbouring clusters and their dependencies on indicators.

Analysis of Main Cluster Characteristics

Several maps were computed before selecting the best one. Prior to listing the different types of experiments, as well as the selection and analysis of the best mapping, it is useful to explain features discussed in the following sections, such as:

- There were 27 data vectors each one representing an individual province. Each data point had a vector of dimension 5, explained as 5 long-term growth indicators for an average value of years 1978 to 1998.
- Each of these vectors was assigned a specific colour reflecting its distance from its neighbours. Data objects near each other were assigned similar colours, whereas ones further apart had dissimilar colour coding. The Euclidean distance was used to calculate differences between colours in neighbouring areas.
- The borders of each cluster were also determined by this colour coding which does not allow for inputs in different clusters to have the same or similar colour. In the case of data inputs spilling into neighbouring clusters, the SOM locates input vectors further apart according to their distance from the centre of the cluster there are classified in. The SOM weighs all 5 components of the input vector and forms a cluster according to similarities among all of them. If a vector is located outside the main body of the cluster it represents and near another cluster with different features, it does not mean that its has changed cluster features, rather that it is very far away from the other inputs in the main body cluster and from the cluster centre.

A summary of the most striking and interesting features of the experiments commenced in order to find the map that extracted the most significant knowledge is given below, these include:

- In some experiments that extreme outliers were allowed with significantly different magnitude to the rest of the data, the map located the input in one corner of the map and affected the other inputs by clustering them all together. The neurone with the outlier also had a strikingly different colour coding (very bright magenta) compared to all the others (shades of green).
- Different size maps were computed. A 8x3 mapping which did not show any clustering tendency, instead nearly each province had its own neurone, as well as a 3x8 mapping which clustered all inputs in two large main areas. In both types of experiments, the map seemed to stretch further away - either wider or longer - lowering the clustering tendency by locating approximately a neurone per province data vector and thus losing potential useful knowledge on distinct features of the inputs.
- Another set of experiments included, inserting new data values for some of the provincial input vectors. This had two different effects; Small differences between old and new data values would result in the SOM locally rearranging some provinces, thus affecting only the neurone or sometimes cluster where the changes took place; Large values would alter the structure of the particular cluster and also affect its neighbours, which was determined by the effect the larger data values had on the input vectors of its neighbouring clusters.

Figure 4.1 shows the clustering dynamics of the input data for the selected best map, divided in four main areas: yellow data inputs indicating cluster **A**; blue dots representing **B**; green inputs belonging to **C;** and purple dots forming cluster **D**. Neighbouring units are more densely packed in **A**, with the distances between inputs being relatively small. **A** also occupies most of the data inputs. In **B** inputs stretch to quite a significant radius with very low density, whereas **D** covers a much smaller area with few samples. Input data from **B** and **C** spill into all other clusters thus sharing features denoting low clustering tendency whereas **A** and **D** seem more distinct denoting a higher clustering tendency.

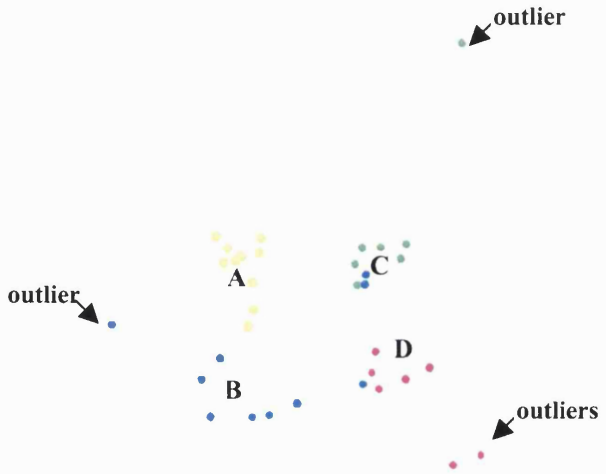


Figure 4.1 – Main Cluster Formations

Figures 4.2, 4.3, 4.4 and 4.5 depict close-ups on individual clusters A, B, C, D, respectively as well as their location and dependency on their neighbours. A description of each individual cluster follows with information including the percentage of the sample that each occupies and some quality information on their neighbourhood radius and density.

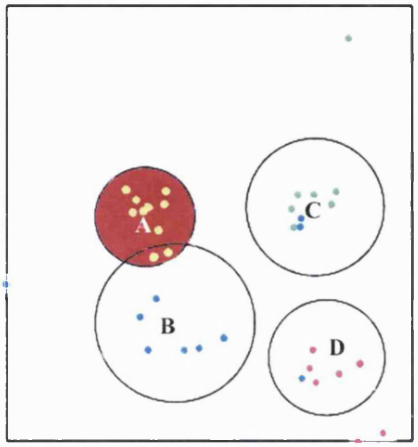


Figure 4.2 – Cluster A and neighbours

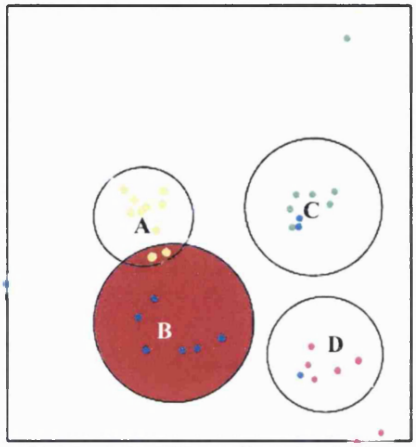


Figure 4.3 – Cluster B and neighbours

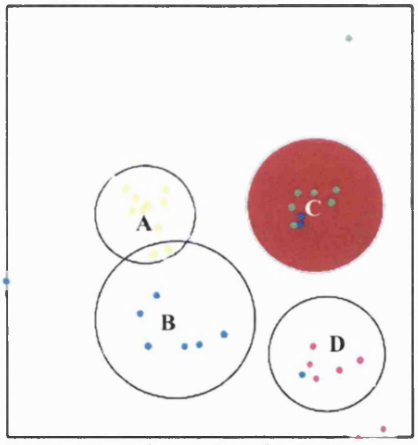


Figure 4.4 – Cluster C and neighbours

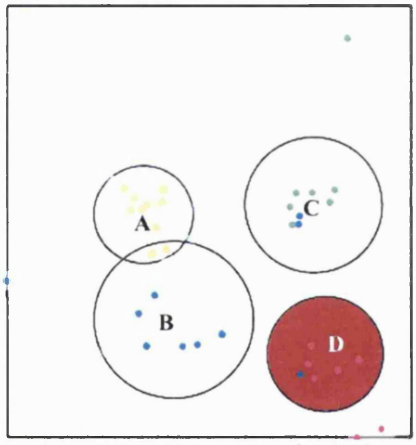


Figure 4.5 – Cluster D and neighbours

Cluster Info	**Samples**	**Density ratio**	**Radius ratio**
A	40%	0.83	1.28

Cluster **A** holds the largest share of the data samples (40%) and is densely clustered (0.83). These measures indicate a high clustering tendency with many data inputs all collected near each other thus denoting an area with dominant characteristics for the growth identity that it represents. Observing **A** and its local neighbours in figure 4.2 it is obvious that some of the samples have stretched out of the main cluster and spill over into neighbouring **B**, demonstrating a trend that moves to a cluster with different features explaining a high neighbourhood radius of 1.28. Nonetheless, the comparably high density ratio indicates good preservation of local distances and thus good structure preservation.

Cluster Info	**Samples**	**Density ratio**	**Radius ratio**
B	21%	0.57	0.76

B holds 21% of the data samples, making it the third largest cluster. An interesting feature when observing **B** in figure 4.3 is that although it does not own one of the larger samples, its data inputs spill into neighbouring **C** and **D** indicating a low clustering tendency and explaining the low density ratio of 0.57. The inputs stretch out to areas outside **B** indicating a transition from cluster type **B** to **C** and **D**, nonetheless only approaching the borders of both clusters. Although selective data inputs from **B** seem to be influenced by its neighbours, it has preserved some local distances between inputs denoted by a neighbourhood radius ratio of 0.76.

Cluster Info	**Samples**	**Density ratio**	**Radius ratio**
C	28%	0.53	1.28

C is the most interesting dynamic cluster depicting high inequalities in its samples with some merging with **B** and others stretching to other areas. It shows the lowest clustering tendency explaining the high neighbourhood radius ratio of 1.28 and the significantly low density ratio of 0.53 – the lowest of all clusters - relative to a second largest share of 28% of data inputs. An interesting observation from figure 4.4 is that despite their similarity of identities, the data points that represent **C** fuse only into the clustering borders of **B**, indicating differences that will be explained and discussed in section 4.3. The effect of the inclusion of outliers is most evident in **C** since some data points stretch to areas with no cluster formation trend. The density of the data in some areas is quite high but declines in areas where high number of local neighbourhood vector distances that are not preserved.

Cluster Info	**Samples**	**Density ratio**	**Radius ratio**
D	11%	1.23	1.64

D holds a mere 11% of the share of the samples, the lowest among all clusters. It experiences the highest density ratio of 1.23 thus preserving the data items distances and consequently its structure. In figure 4.5 we can distinguish **D** as the cluster with the most distinct features, with small local distances keeping away from any of its neighbours. Another important observation is that a few inputs have escaped the main body and even border of this otherwise densely packed cluster. This movement near the borders of the map explains a significant neighbourhood radius ratio of 1.64, which however is compared to a cluster of the smallest share of data inputs.

These features from all clusters are going to be used as the groundwork and investigated further when the individual provincial growth type are analysed according to their location on the Kohonen map in section 4.3.

Indicator Correlations

Characterisation of the clustering properties of individual growth indicators is another fundamental issue since interesting deviations or interactions between variables can be observed that can then be used to complement the main map that clusters provincial growth in latter sections of this chapter. The information used to determine the clustering of each indicator as well as their correlations employs sector-specific data samples averaged over the reform period (1978-1998) for all 27 provinces (Appendix B.1).

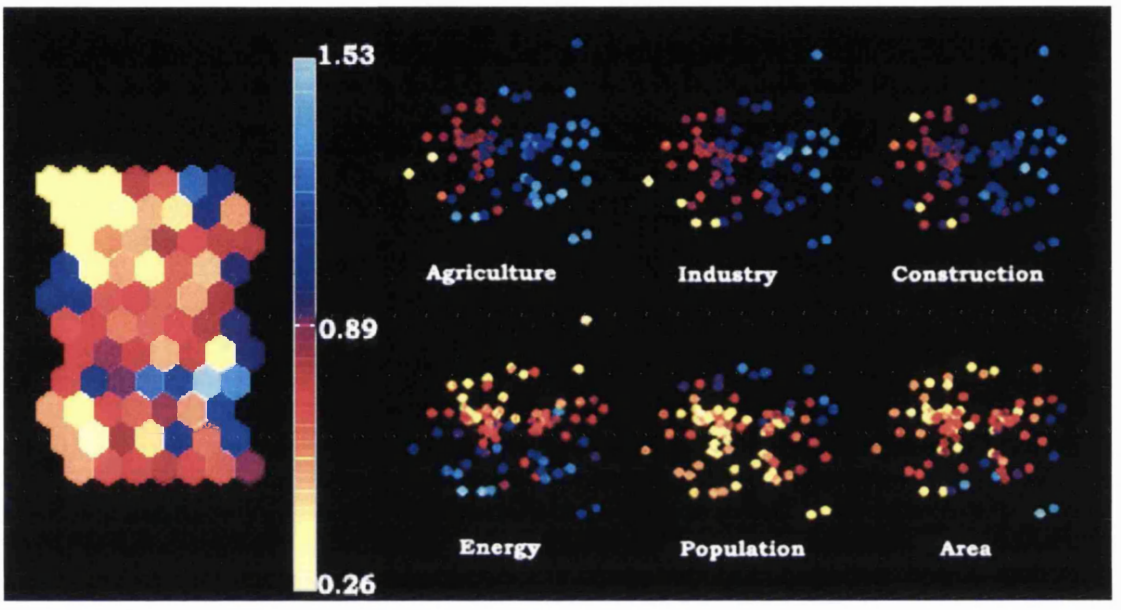


Figure 4.6: U-matrix colour coding and data sample analysis for each indicator:
a. Agriculture; b. Industry; c. Construction; d. Energy; e. Population; and f. Area.

Figure 4.6 shows graphs of the six indicators used in the main SOM together with the U-matrix measure. The U-matrix is the main distance measure that visualises distances between neighbouring map units either by calculating all or parts of the indicator values, thus depicting the clustering structure of each unit on the map [Kaski97]. The colour bar accompanying the U-matrix shows the scale of the variables which in turn indicates the values that the variables have in the map structure. Each figure corresponds to one indicator and each dot in each figure to one data sample. The colour coding reflects the distances between model vectors such that data objects close to each other are assigned similar colours, thus avoiding non-neighbouring areas to attain the same colour. An automatic method called the Comission Internationale de L'Eclairage (CIE) [KaVeKo00] is employed that almost mimics the human colour vision system. It employs a group of uniform spaces corresponding to the Euclidean distance which calculates differences between colours in neighbouring areas making it possible to visualise complex structures automatically.

Observing figure 4.6.a, it is clear from the domination of the blue and light blue dots that most of the data samples for agriculture are located in the range between 0.80 and 1.50, denoting high

values for this indicator. Similar scaling is indicated in figures 4.6.b. and 4.6.c, depicting the industry and construction sectors respectively. The colour domination for the next three indicators however dictates a decrease in scale with the yellow and red dots governing most of the map space. Analytically in figure 4.6.d, the energy sector divided in the yellow and red area and the blue and light blue area, with a few outliers on both ends (yellow – blue) of the U-matrix colour range. In figure 4.6.e the differences are significant, with the yellow dots ranging between 0.26 and 0.60 dominating the map units that represent the population indicator. The area indicator in figure 4.6.f forms a similar colour pattern as energy and population with yellow and red dominating the map space. However, the light blue dots – denoting the highest values – make interesting outliers corresponding to provinces with high land occupation and will be examined in section 4.3 together with the main SOM and the indicator profile analysis in order to interpret their position.

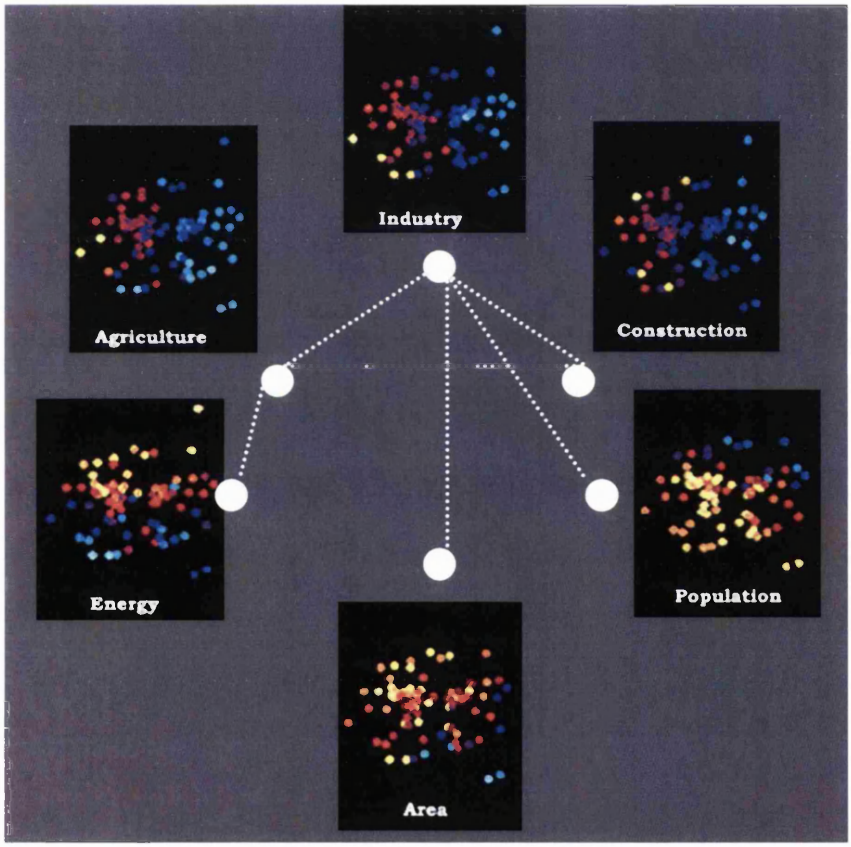


Figure 4.7 – Indicator Correlations

Figure 4.7 integrates the above information from the individual indicator figures to give their clustering patterns and relationship by linking them together by position using the same colour coding methodology as above. Industry, construction and agriculture all share the same features. Industry has the same distance from agriculture and construction whereas the latter two are further apart. Agriculture and energy are highly correlated linked by the smallest distance whereas area and population indicators though directly related to industry occupy the position furthest apart. Finally, the industry sector seems to have attained the central unit position since it has direct links to all the other sectors excluding energy.

4.3 Computation of the Growth Maps

In this section we aim to advance beyond classical growth analysis to the interpretation of mappings using sophisticated intelligent techniques that employ multi-dimensional input data sets. All the maps were produced from programming functions on the SOM Toolbox platform developed by Teuvo Kohonen and his team at Helsinki University of Technology (HUT). Information for the toolbox is introduced in Appendix A.1. This research aims to be the first to illustrate, analyse and discuss the following four aspects of province and sector-specific long-term growth data sets:

- The mappings of the provinces on the SOM demonstrating growth patterns in the reform era (1978-1998) and individual years (1978, 1988 and 1998). The clustering tendency in the data set is shown by displaying the distances between the growth types of neighbouring map locations with different colour levels and different intensity;
- The pair-wise comparison of individual growth indicators displayed with different colour levels on the map foundation;
- The comparison of the individual growth indicator in single pie charts for each map unit showing the relative proportion of each component to their sum in the specific map unit;
- The distribution of single and pairs of indicators on the map described by simple scatter plots and histograms.

We will use the SOM to cluster all indicators and provinces for the entire reform period, all indicators and provinces for individual years and all provinces for each individual indicator using incomplete and in some cases missing Chinese data sets to analyse these four aspects.

4.3.1 Emerging Growth Type Patterns in the Reform Era

The first aspect of the computations of long-term growth is given here with the creation of a single SOM from sector-specific per capita data for all provinces from 1978 to 1998 (analytical tables for each year are found in Appendix B.1). A smooth colour palette is superimposed on the neurones of the SOM providing a compact representation of the selected data thereby associating a colour code with each of the model vectors. The resulting global map summarises the similarities and dissimilarities among emerging and frontier provinces and their individual growth types and growth group identities. The data vectors are spread over a 6x4 map or 24 neurones with the ones most similar grouped together, while others depicting differences appearing further away from each other on the map. Figure 4.8a (***SOM map***) shows a 6 by 4 mapping of the provinces complemented by a relative importance figure and the relevant averaged data set over the entire reform period. We will now discuss how our findings of the clustering dynamics (section 4.2.2) are interpreted by the SOM to form the main growth zones.

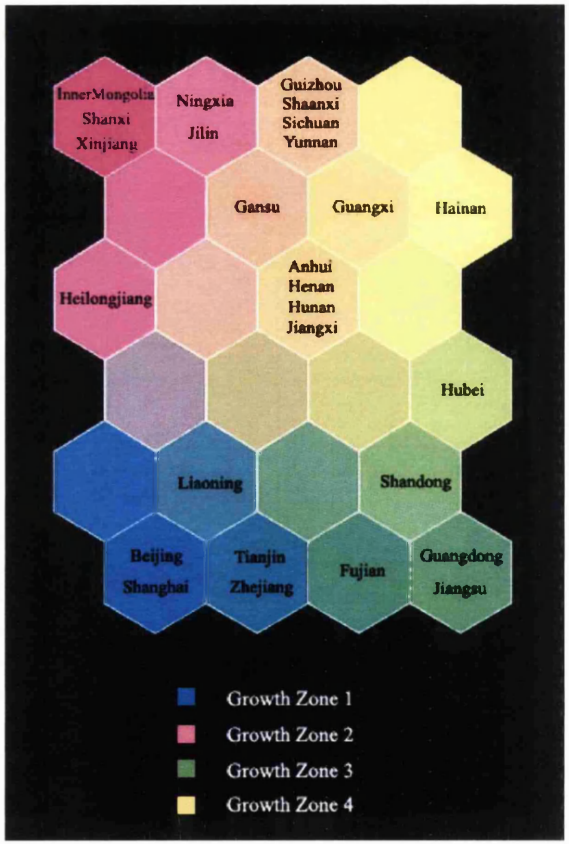


Figure 4.8a: SOM Provincial Growth Types

AVGPC7898	Area	Agriculture	Industry	Construction	Energy
I.Mongolia	5.0448	0.1019	0.2273	0.1905	2.1708
Shanxi	0.4918	0.0472	0.2359	0.2266	8.5281
Xinjiang	9.1586	0.1130	0.1872	0.2762	1.9155
Ningxia	0.1234	0.0396	0.0938	0.0252	0.2997
Jilin	0.7073	0.0909	0.2681	0.2368	1.1506
Guizhou	0.4647	0.0448	0.0772	0.0798	1.1638
Shaanxi	0.5701	0.0582	0.1622	0.1650	1.1038
Sichuan	0.4193	0.0617	0.1491	0.1547	0.6497
Yunnan	0.9508	0.0603	0.1188	0.1437	0.5885
Gansu	1.7864	0.0535	0.1600	0.1523	0.8070
Guangxi	0.5055	0.0759	0.1461	0.1002	0.2655
Hainan	0.4515	1.1505	0.4768	0.1576	0.0014
Heilongjiang	1.2430	0.0831	0.2834	0.2548	3.4262
Anhui	0.2248	0.0788	0.2141	0.1194	0.5922
Henan	0.1793	0.0710	0.2075	0.1006	1.0852
Hunan	0.3230	0.0788	0.1879	0.1291	0.6768
Jiangxi	0.0397	0.0076	0.0156	0.0089	0.0586
Hubei	0.3173	0.0837	0.3209	0.1770	0.2088
Liaoning	0.3505	0.0870	0.5391	0.4222	1.6211
Shandong	0.1731	0.0982	0.3914	0.1788	1.0395
Beijing	0.1348	0.0657	0.7857	1.4591	0.8694
Shanghai	0.0423	0.0608	1.6614	1.1056	0.0138
Tianjin	0.1181	0.0633	1.0396	0.6674	0.5265
Zhejiang	0.2285	0.0934	0.6935	0.4320	0.0412
Fujian	0.3637	0.1070	0.3610	0.2138	0.3168
Guangdong	0.2604	0.0885	0.5577	0.3771	0.2125
Jiangsu	0.1429	0.1055	0.6484	0.3576	0.3485

Table 4.1: Per Capita Indicator Data (averages)

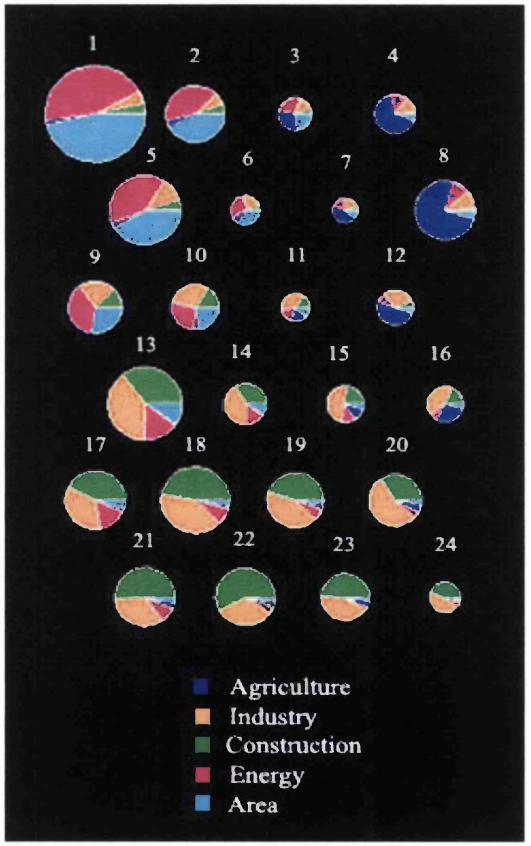


Figure 4.8b: Relative Importance Pie Charts

Figure 4.8a supplies information for the main SOM map with the four corner neurones of the map depicting the most dominant representative provinces of each of the four growth zones indicated with magenta, bright yellow, dark blue and dark green. Table 4.1 supplies a sample of the datasets used averaged over the entire reform period. Figure 4.8b measures the relative influence (percentage) of each indicator in each hexagonal neurone affecting the location of each province on the map and shows the sectoral growth rate of each province in each neurone affected by quicker (smaller pie chart) or slower (larger pie chart) growth. Pie charts are created indicating the relative proportion of each indicator to the sum of all indicators in a specific map unit and its magnitude as a whole. The size of the pie charts has been scaled according to the Euclidean distance between the indicators. Both figures are used to discuss our findings and interpret the formation of four main cluster areas dictating the main growth zones among the Chinese provinces discussed in this section.

Growth Zone 1: Coastal provinces with high living standards, high infrastructure and the highest industrial production.

Located at the bottom and middle left side of the main SOM map in figure 4.8a represented by shades of blue colour coding (also indicated by the colour bar adjacent to the growth zone number) this is the zone examined in section 4.2.2 as cluster **B**. Cluster **B** with 21% of the data samples indicated low clustering tendency and a low density ratio of 0.57 with its inputs stretching out to cluster types **C** and **D** nonetheless preserving some local relationships denoted by a neighbourhood radius ratio of 0.76. The provinces comprising this cluster growth zone are Beijing, Shanghai, Liaoning, Tianjin and Zhejiang.

This information about cluster B and the location of the provinces on the main map is combined with the relative importance of the indicators in neurones 18, 21 and 22 occupied by these provinces shown in figure 4.8b. In all 3 neurones the pie charts are mainly occupied by green and orange colours that represent the infrastructure and industry indicators respectively. This is the zone characterised by the best infrastructure and the highest industrialised production. Nonetheless, these provinces lack energy resources (purple colour), have relatively low agriculture-based industries (blue colour) and occupy small geographic area (light blue). In order to be able to retain this growth level in the future it needs to acquire energy and food resources from provinces in other growth groups that have such resources.

Growth Zone 2: Border provinces rich in natural resources with limited infrastructure.

Located at the very top left side corner on the SOM and represented by shades of magenta colour coding this is the zone examined in section 4.2.2 as cluster **D**. The provinces comprising this

cluster growth zone are Xinjiang, Inner Mongolia, Shanxi, Ningxia, Jilin and Heilongjiang. **D** forms the most distinct structure holding the lowest share (11%) of the samples with the highest density ratio (1.23) preserving the data distances, keeping away from any of its neighbours. When analysing **D** in section 4.2.2 we discovered an interesting feature where inputs escaped the main border of this otherwise densely packed cluster to approach the area near the borders of the map.

This distinct finding can be explained here by combining the SOM information and the relative importance pie charts. The provinces occupy neurones 1, 2 and 9 of the relative importance figure. Mainly influenced by the energy and the area indicators, these are the neurones where indicators have influenced their position the most out of all others in the relative importance figure. This explains the highest density ratio and distinct cluster structure which separates these neurones from the rest. The map has accurately placed Jilin and Ningxia in the middle top of the map which occupies neurone 2 where the influence is not as significant since the other provinces in this zone occupy the top 5 seats for energy production whereas they are ranked 9^{th} and 10^{th} respectively and also occupy a much larger area than these two provinces.

Another interesting feature is detected in neurone 9, occupied by Heilongjiang province which has a limited but consistent influence in most indicators representing the unit in cluster **D** that is moving towards cluster **B**, which may be denoting a transition to a different growth zone (growth zone 1) with higher living standards. The rest of the provinces in this group have relatively low living standards indicated by the limited pie chart proportion of the construction and industry indicators, which does not allow them to take advantage of their enormous area (Xinjiang occupies 1/6 of the total area of China). From figure 4.8b and the relevant pie chart it is observed that these provinces have considerable amounts of natural resources that they can exploit - Xinjiang's share of China's production is 10.9% and 12.3% for oil and natural gas respectively and has an advantageous bordering location with Kazakhstan (Caspian Sea) - to both China's domestic market and to oil multinationals if they focus on developing their infrastructure.

Growth Zone 3: Cluster of the highest growth rates with low energy resources.

Located at bottom and middle right side of the map represented by shades of green colour coding this is the zone formed by cluster **C**. The provinces comprising this growth group are Guangdong, Jiangsu, Shandong, Hubei and Hainan. In section 4.2.2 we characterised **C** as the most interesting cluster with some of its units merging with **B** and others stretching to areas where there is no cluster formation trend explaining the high neighbourhood radius ratio of 1.28 and the lowest density ratio (0.53). We also detected that data points in **C** fuse only into the borders of **B** which is growth zone 1, depicting the similarity of the growth identities of the two clusters since they both have middle to high living standards.

Occupying neurones 16, 20, 23 and 24 these provinces have good infrastructure and industry production with middle living standards and an increase in agricultural production in neurone 16 denoted by Hubei. Hainan's growth is very distinct since it is mainly influenced by agriculture and it is this feature that has positioned it to occupy one neurone (neurone 8) on its own explaining the units that spread to areas where there is no cluster formation outside the borders of **C**. Shandong and Fujian have low energy reserves but good exploitation of their advantageous agro climatic conditions of the region. An interesting feature is depicted in the neurone occupied by Guangdong and Jiangsu with green and orange dominating the pie chart nonetheless depicting low per capita growth which may be indicating a difficulty in equally distributing the zone's wealth across the different parts of the province thus creating social inequalities.

Growth Zone 4: Cluster of increased poverty and high inequalities.

Located at top, centre and middle right side of the map represented by shades of yellow and peach colour coding this is the zone formed by cluster **A**. The provinces comprising this zone are split in two main sub-cluster areas, the first comprising of Hunan, Henan, Jiangxi, Anhui and Gansu and the second of Yunnan, Guangxi, Guizhou, Shaanxi and Sichuan. **A** holds the largest share of the data samples (40%) and is densely clustered (0.83) with most inputs collected near each other indicating a high clustering tendency and some others spilling over into neighbouring **B** demonstrating a trend that moves to a cluster with a focus on infrastructural development and industry.

Guizhou, Shaanxi, Sichuan and Yunnan occupy neurone 3, where the sector significance drops and all the sectors are equally distributed though excluding construction. Neighbouring neurones usually share many common features, however that is not the case with Gansu (neurone 6) and Guangxi (neurone 7) provinces show unexpected contrasting characteristics with Gansu being mainly affected by energy and area whereas Guangxi by agriculture, also explaining why they are occupying a neurone each. These provinces comprise the first sub-cluster of growth zone 4, with low levels in every indicator. Anhui, Henan, Hunan and Jiangxi that form the second sub-cluster show the same features as the other provinces in this zone however, an increase in the pie chart of the industry sector moves them towards the more developed industrialised provinces explaining the trend of some units towards **B** of high industry and infrastructure levels. Finally, although high clustering tendency may be very important for feature preservation in this cluster it is translated to widespread poverty and the worst living standards shared by most provinces in this group in all sectors including infrastructure, industry and even agriculture.

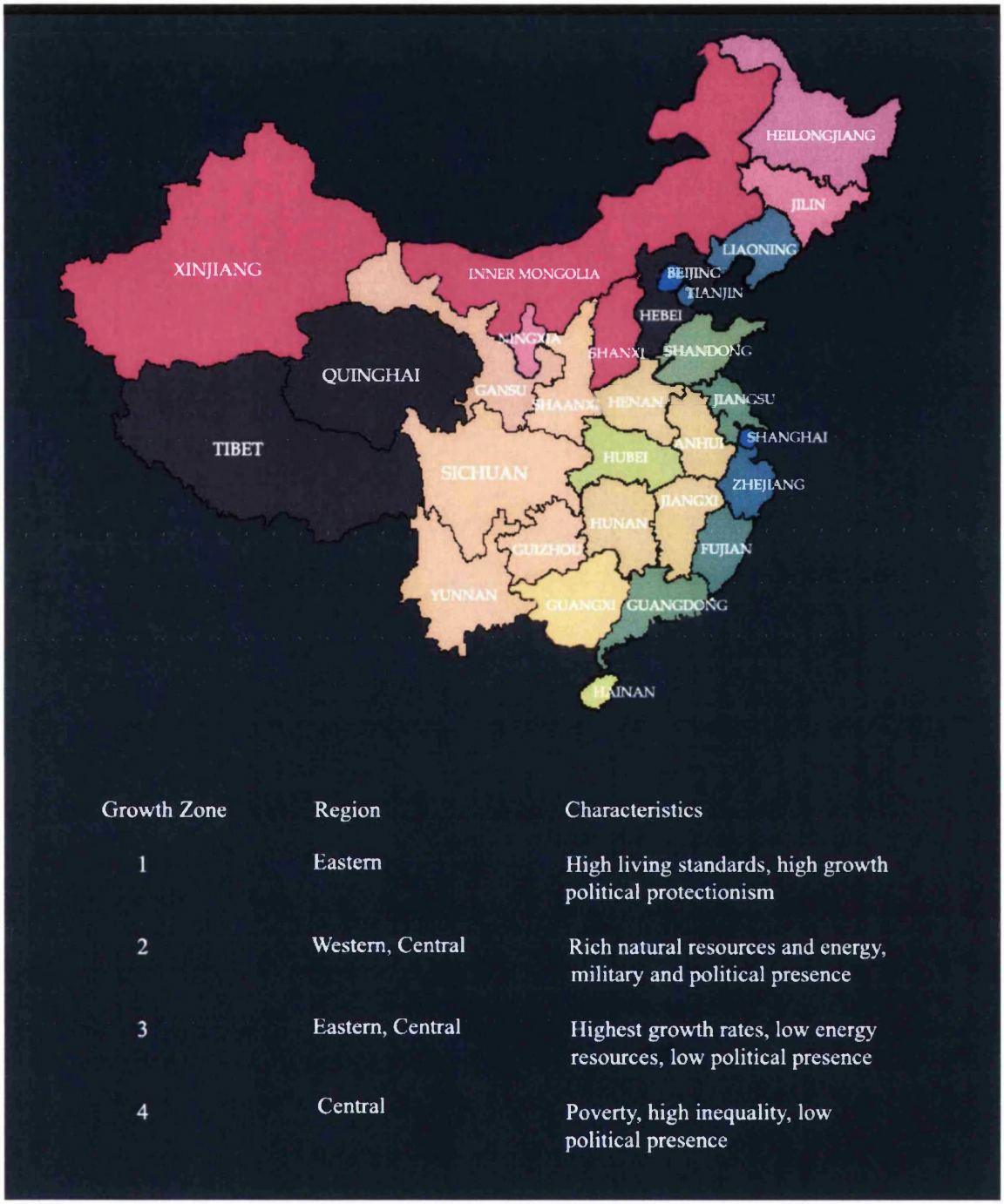


Figure 4.9 – China map representation of geographical growth dynamics

Figure 4.9 shows a map of China with the geographical positioning of each of the provinces with their individual growth type and colour coding determined by the main SOM map in figure 4.8a (section 4.3.1.). An interesting point observing the map of China is that although we did not supply the SOM with any information about the geographical location of the provinces, the four clusters formed also represent China's main geographical regions. Tibet, Quinghai and Hebei are not included in the sample due to adjustments explained in section 4.1.1.

4.3.2 Growth Patterns for Individual Years

In the previous section a single map was produced that clustered provinces based on growth data sets averaged over the entire reform period. The information collected for the maps in this section includes data for each indicator and each province for years 1978, 1988 and 1998 found in Appendix B.1. The schematic representation in Figure 4.10 shows 3 maps for years 1978, 1988 and 1998 discussed in this section which investigates the SOM's ability to cluster all indicators and provinces for individual years. We have selected these years because the SOM is aiming to interpret the provincial growth transitions and these years comprise the start (1978), intermediate stage (1988) and recent years (1998) of the reform period.

In the first map, which represents 1978, provinces group more densely in one neurone forming one large group at the top right side of the map and a few smaller ones on the lower part of the map. In 1988 the SOM starts to show a different cluster structure with provinces stretching out to neurones in every area of the map, with the exception of provinces in the poorest growth zone (growth zone 4) occupying only two neurones between them at the top right side of the map. As the provinces evolve from one decade to the next they assume different patterns that are formed in different ways affected by the growth policies they selected and jump from cluster to cluster to end up forming the four main growth groups in the 1998 mapping.

Discussion

By adding the knowledge attained from the previous sections and the results of the growth trend in individual years we made a very significant discovery of the attempt of the SOM to give an interpretation in both the vertical and horizontal dimensions of the map of the main policies that influence and locate growth type on the map:

The horizontal dimension of the map seems to reflect different levels of transition in the reform period, starting from Beijing on the left side, with increased political presence and old planned economic policies, and on the right side with clusters of provinces including Guangdong indicating an explosion in foreign trade and investments. The vertical dimension seems to indicate a poverty and inequality scaling with the poorest provinces occupying the top section of the map moving to better distribution of wealth as we approach the lower regions of the map where income equality and middle to high living standards are detected. The central part of the map seems to reflect provinces on the move, some experiencing low growth rates after the first decade and others converging towards wealthier areas of the map.

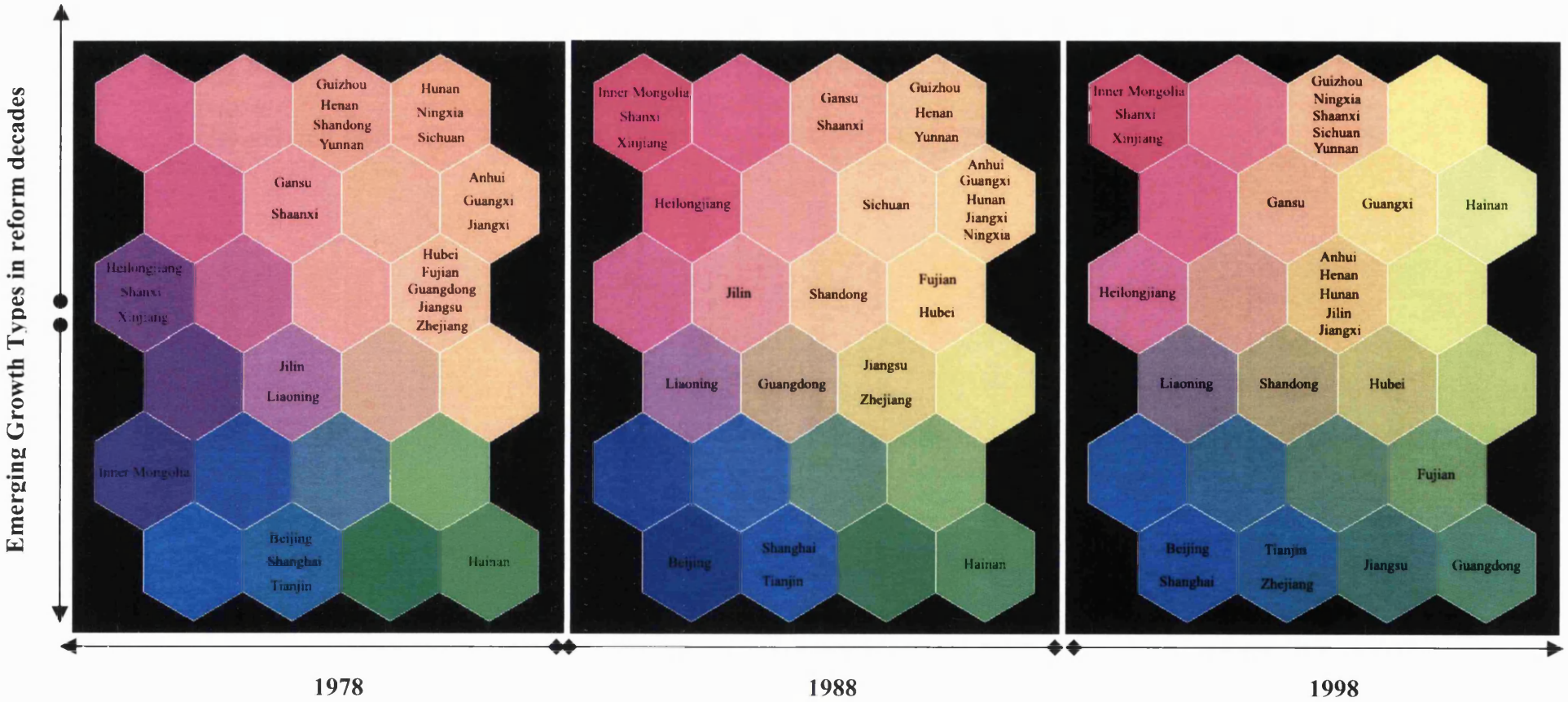


Figure 4.10 – Schematic representation of evolving growth patterns

This understanding contributes to better explaining the location for each province. The horizontal dimension for provinces on the bottom left neurones for all 3 mappings shows the Chinese capital Beijing, Shanghai, Tianjin and Liaoning which are provincial-level units of strategic importance to the central government [Hend99] also depicting high levels of equally distributed wealth which we measure by taking per capita values for all indicators. As detected in the main map, these provinces maintain high growth status through the entire reform period. Maintaining high growth levels can be explained since political presence is especially high - does not decrease in any of the 3 maps. Sustained political and military attention is also obvious in the top left neurones including Xinjiang and Inner Mongolia with the constant concern of invaders emerging from border provinces [Good97], [FZZ02], [Hend99]. On the right side of the map, Guangdong's rapid rise has resulted mainly from its bordering and integration with Hong Kong. Shandong together with Fujian seem to have developed rather slowly in the first decade but clearly accelerated during the second decade probably due to their key economic relationship with South Korea and Taiwan respectively, through decreasing surveillance from the central government [FZZ02].

On the vertical dimension the provinces of zone 4 have low growth levels through the entire reform period with only Anhui, Henan, Jiangxi and Hubei moving towards higher living standards. Guangdong, Jiangsu, Zhejiang and Heilongjiang make the transition to neurones at higher levels approaching the zone formed by Beijing, Shanghai, Tianjin and Liaoning which maintain their high growth status since 1978. Hainan is the most interesting exception in all three mappings since it keeps on occupying a neurone on this own moving from higher growth in the first decade to lower in the second. The initial high levels replicated the whole country's key growth transitions in the first decade and Hainan's experience of reforms as a testing ground for central government to try bolder economic and political policies [Good97]. Its use as a testing ground seems to be the reason why Hainan's development decreased during the second decade.

4.3.3 Indicator Profile Analysis

In this final section of the map computations, we discuss two figures:

1. A comparative figure (Figure 4.11) with scatter plots and histograms for all the variables distributed in an individual and pair-wise manner; and
2. Explanatory figures 4.12 and 4.13 for selected pair-wise indicators.

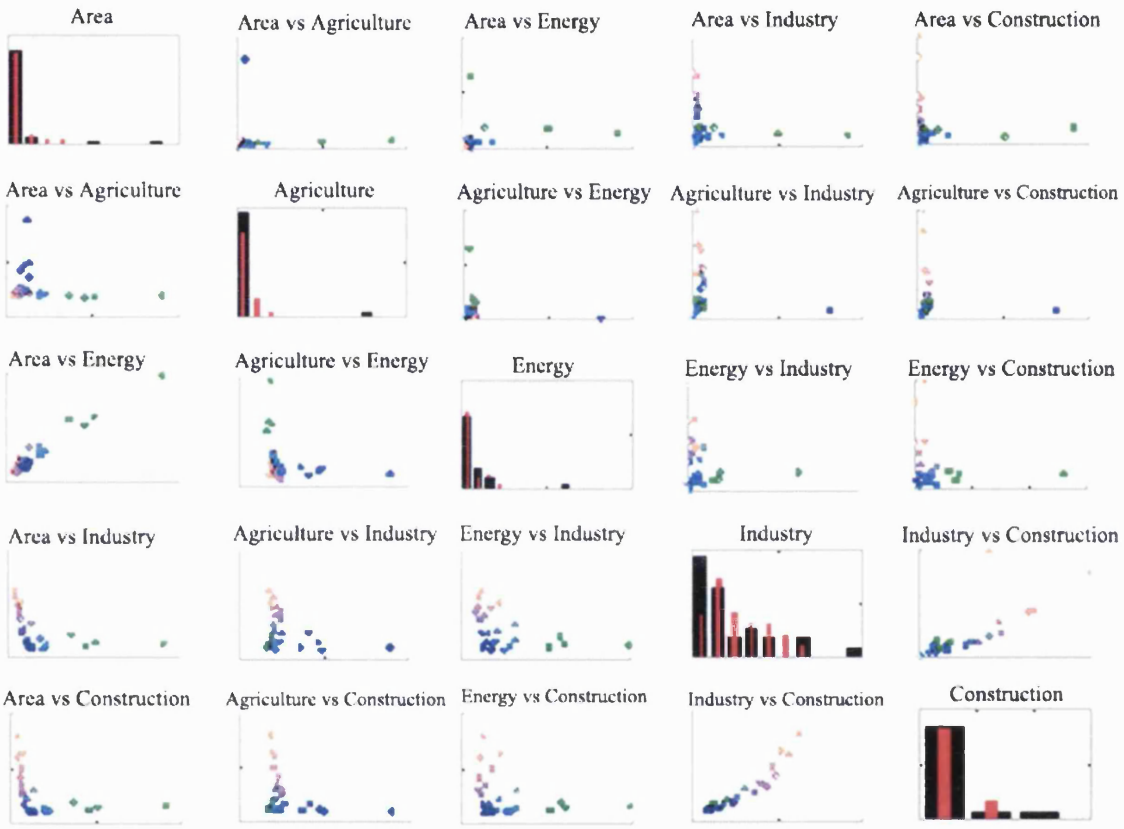


Figure 4.11 – Indicator individual histograms with pair-wise scatter plots on either side of diagonal

In figure 4.11 the histograms occupying the diagonal section of the graph provide analysis of each indicator, with black colour representing the original data set and red colour for the map prototype values. The scatter plots are located symmetrically on either side of the diagonal section with the colours that each data sample employs in figure 4.8. The right side of the diagonal section shows scatter plot formations of the original data points for comparisons of pair-wise indicators, whereas the left side shows the clustering tendency of the map prototypes. The SOM is used to measure the similarities or dissimilarities between indicators and to indicate the pair-wise subplots that should be further analysed. The indicators representing area and energy have a highly linear correlation shown in the map prototypes in subplot 3.1 of figure 4.11, which is not true though for the original data clustering. Industry and construction have a highly linear correlation in subplot 5.4 which was also identified in the indicator dependencies in section 4.2.2. Agriculture (subplot 2.1) and Industry (subplot 4.1) do not seem to share many features and show dissimilarities with the area sector.

Given these observations we compute maps for pair-wise indicators using a data set for all provinces for 1998 with per capita indicator information for agriculture and energy (figure 4.12) in the first map and industry and construction per capita in the second (figure 4.13). The SOM is employed to compute two more maps of pair-wise indicators for all provinces in a single year depicting its ability to extract different types of information. The specific pairs of indicators depicted in figures 4.12 and 4.13 were selected, the agriculture vs energy due to the striking clustering differences between original data and map prototypes and as for industry vs construction their relationship has been observed in most mappings and depicts the highest correlation in figure 4.11.

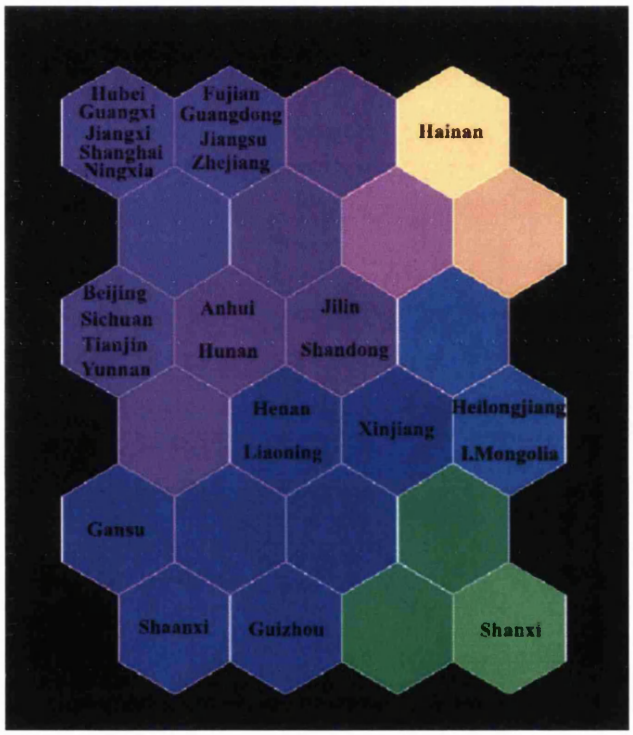


Figure 4.12 – Agriculture vs Energy (6x4) SOM

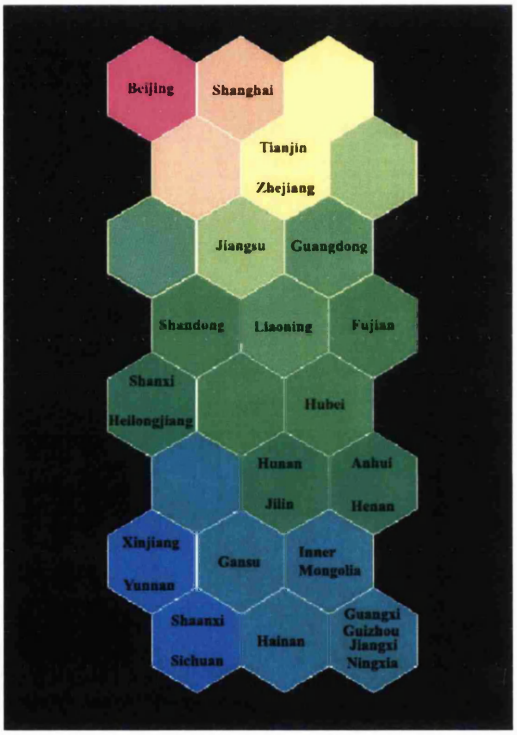


Figure 4.13 – Industry vs Construction (8x3) SOM

While observing the two maps the first interesting feature is that they have different shapes. Figure 4.12 shows a 6x4 map whereas figure 4.13 shows a 8x3 map. The reason for this is that the map stretches in different directions to match the entire range of original data points. Consequently, the second mapping is longer and narrower than the first, due to the fact that the original data has a longer range. Figure 4.12 shows the agriculture vs energy map with the two dimensions once more showing the overall trend of the inputs on the map, with the vertical axis indicating different levels of energy and the horizontal axis depicting levels of agricultural growth. Analytically, Guizhou, Shaanxi and Gansu have relatively substantial energy but few agricultural resources together with a poor infrastructural and industrial basis observed in the industry/construction mapping. Anhui and Hunan have substantial agricultural outputs but occupy the middle of the map since their energy levels are also satisfactory. Shanxi's single neurone location though close to the energy zone cluster with the richest natural resources (Xinjiang,

Heilongjiang and Inner Mongolia) indicates substantially high natural resources (lowest right side neurone) with good infrastructure and industry (figure 4.13 – middle left side neurone).

In the industry/construction map in figure 4.13 all the provinces in growth zones 1 and 3, including Beijing, Shanghai, Tianjin, Zhejiang, Jiangsu, Guangdong, Shandong, Liaoning and Fujian (reform era SOM – Figure 4.8a) occupy the top section of the map (highest values) with Beijing denoting a remarkable exploitation of per capita resources in these two indicators and obtaining the highest living standards followed by Shanghai, Zhejiang and Jiangsu. Sichuan, Xinjiang and Inner Mongolia's positions in the lowest cluster level indicate very low construction and industry levels. Their location on the map reflects infrastructural constraints including poor physical condition of the road and rail transport links – often indirect - to major cities [FZZ02]. In order for these provinces to reach their full growth potential and increase their cross-border and inter-provincial trade links, the infrastructure of key regional arteries has to be improved. Hainan's position on the map occupying the lowest level of industry and construction and highest level of agriculture confirms its uniqueness discussed in earlier sections.

4.4 Summary

In this chapter we investigated the strategic importance of SOM-based design in providing direct visualisation and location of the numerous Chinese provinces according to their distinct sector-specific growth identities. The SOM dealt with inaccurate and missing elements for the early reform years and was successful in classifying and visualising the provincial growth patterns using panel data for indicators and provinces over the entire reform period. We continued to show that the SOM can used to examine the development of specific indicators after computing the main mapping. In addition, we illustrated how the main map can be used in combination with the relative importance pie charts to get a better understanding of how each indicator affects the location of each province in specific growth zones. Another important discovery through observation of individual year mappings 1978, 1988 and 1998 was the horizontal dimension reflecting different levels of transition in the reform period and the vertical dimension of the map indicating levels of poverty and inequality. Consequently the SOM was able to investigate the impact macroeconomic policies had on each of the Chinese provinces over the reform period.

Chapter 5

GTCA Part II: Evolving Bilateral Trade Flows

Chapter 5 focuses on the second part of the GTCA that tries to identify the symbolic expression that best specifies China's bilateral trade behaviour through rules formed from empirical trade flow data. Genetic programming is employed to portray the main relationship of China's trade pattern in the international market and identify the future relationship of the descriptive variables of the system. Discrepancies in export figures between China and its trading partner's mainly due to re-exports via Hong-Kong, are adjusted. The theoretical concept of the gravity equation is used to obtain trade flow knowledge from input data.

5.1 Evolution of China's Trade Environment

The focus of the second part of the GTCA is to evolve bilateral trading rules in order to identify under- or over- trading actions between China and its main partners. Genetic programming is employed to evolve rules that express actions in China's trade environment and try to identify the geopolitical dynamics in their bilateral trading actions. We use trade flow data to identify the main symbolic relationship formed as well as secondary relationships selected - due to high partial fitness crediting - that affect the Chinese trading environment.

5.1.1 – Using Evolution to Identify Trade Actions

China has gradually opened up integrating itself into the international trade markets in a secure and efficient manner (exports reached 25% of total production in 2002 compared to 5% in 1982). Due to its structurally diverse economy its integration into the international trade system seems challenging. These diverse features also affect its partners trading behaviour. This work aims to determine under- or over- traded behaviour related to these actions by finding an expression that highlights the effect a set of selected independent variables (inputs) has on the associated dependent variables (outputs) for China's real-world trading environment. Once this model has been determined it can be employed to predict an increase or decrease in the factors affecting the Chinese bilateral trade system. Instead of finding the coefficients determining a particular economic model, genetic programming is employed to genetically breed populations of candidate trade rules that best fit the data, thus identifying the symbolic form of trade actions.

Initially, genetic algorithms were employed for this application for their speed of operation and power of effect. This speed and power has made them highly suitable and applicable in many fields, including economics, business-related and financial applications. The main interest in this part of the GTCA is the modelling of a dynamic trade environment where rules constantly change size and complexity, increasing the understanding of the final solution. The standard GA is more suitable for fixed length representations and tends to produce many copies of a particular individual with exceptional relative fitness, thus getting trapped at local maxima. In addition, it has difficulty handling a changing environment after convergence has occurred. In contrast, GP employs non-linear tree-structured genetic material which alters in size, shape and complexity and performs operations in a hierarchical manner. Crossover operators are defined to preserve the syntactic correctness of the program and although in the GP reproduction process there is a trend towards convergence, there is also a counterbalancing pressure away from it.

5.1.2 Fundamentals of the Gravity Equation

The aim of part II of the GTCA is to find the main relationship that portrays China's trade pattern in the international market and predict the convergence or divergence of partner's trade flows to it. Applied international economics provide us with a model widely used for determining trade flows, the gravity equation. The gravity equation relates trade flows between trading partners with their income, income per capita and the geographical distance between them. The gravity equation is specified as follows:

$$\ln(X_{ijt}) = \beta_0 + \beta_1 \ln(Y_i Y_j)_t + \beta_2 \ln(YPC_i YPC_j)_t + \beta_3 \ln DST_{ij} + \beta_4 \ln M_{ijt} + \varepsilon_{ijt} \quad \text{(Eqn 5.3)}$$

where $\ln(X_{ijt})$ is the exports trade flow between trading partners i and j at time t with their size approximated by income (product GDP for i and j) $\ln(Y_i Y_j)_t$ income per capita (product GDP per capita for i and j) $\ln(YPC_i YPC_j)_t$ and transaction costs measured by geographical distance $\ln DST_{ij}$ between trading partners. β_N is a vector of nuisance coefficients and ε_{ijt} is a log-normally distributed error term. We add a fourth variable to this equation which describes the imports trade flow, $\ln(M_{ijt})$.

The gravity model [Linn66], [Ander79] states that larger and richer countries trade more with each other - positive effects for income and income per capita - than with smaller and poorer countries and trade less –negative effect for distance - when a significant geographical distance separates them. We selected an additional variable to investigate the interaction between export and import trade flows. The balance of trade is measured by the difference between exports and imports. Significantly large bilateral trade imbalance is key to the US-China trade relationship increasing to 11.7bn US$ in August 2003 and breaking the 11.3bn US$ high recorded the previous month. The National Association of Manufacturers in the US warned that if this trend continuous, with Chinese imports six times higher than US exports, the trade gap will triple to 300bn US$ by 2008 [Hugh03]. Economists state that global trade imbalances reflect domestic macroeconomic factors. In contrast, policy makers and trade negotiators argue that such imbalances represent measures of gains (trade surplus) and losses (trade deficit), affecting real trade policies and even creating an environment for trade wars [FuLau03].

The gravity equation is an empirical model that has been successfully applied to a wide range of countries and periods including estimating trade flows for European Union countries [GlRos01], for Iran and its trading partners [Kalb01], and for the OECD countries [Berg85], [Berg89], [Rose00]. Nonetheless, the gravity model tends to weigh equally all data inputs, thus producing coefficient values dominated by the smaller trade flow values due to the different magnitude in values between developed and developing countries. This negative effect was first realised when regressions were performed using the econometric package E-views to obtain trade flow predictions from the gravity equation in order to compare them with our results. The deviation between the gravity equation's prediction for the trade flows and the actual trade flow values was significant. Tables 5.1 and 5.2 depict calculations for two years, 1990 and 1997, certifying that this negative effect is not a special case that occurred in one annual observation. The ten Chinese trading partners were selected in order to include both developing and developed country values. The statistical resources for the actual trade flow observations are referenced in section 5.1.2.

1990 (US$m)	Actual Flow	Predicted Flow	Deviation
HK	41,728	1,983	39,744
US	11,905	4,942	6,963
Korea	669	5,400	-4,731
Japan	16,866	13,593	3,273
Germany	5,042	2,221	2,822
Canada	1,927	1,042	887
Argentina	326	133	193
India	270	532	-262
Venezuela	42	186	-144
Vietnam	3	136	-133

Table 5.1 – Actual, predicted trade flows and their deviations for 1990

1997 (US$m)	Actual Flow	Predicted Flow	Deviation
HK	50,795	7,770	43,025
US	49,034	20,132	28,901
Korea	24,021	35,288	-11,266
Germany	12,677	7,889	4,788
Japan	60,810	57,208	3,602
India	1,835	3,967	-2,132
Canada	3,908	3,175	733
Argentina	1,186	595	590
Venezuela	150	615	-464
Vietnam	1,436	1,070	366

Table 5.2 – Actual, predicted trade flows and their deviations for 1997

The countries are listed according to the absolute value of the deviation, not taking into consideration positive or negative signs. Observing both 1990 and 1997 tables, Hong Kong (HK), the United States (US), Korea, Germany and Japan – developed economies with large trade flows - have the highest deviations, while India, Venezuela and Vietnam – developing economies with relatively small trade flows - have the lowest deviations between the actual and predicted trade flows. These results confirm the belief that by including developed and developing trade flows with very different magnitudes and weighing them equally, the gravity equation coefficients are dominated by the small trade flows. The significance of the gravity model is unquestionable as an important and widely used tool in economic analysis nonetheless this is an important problem that affects the trade flow predictions in a negative way. Our aim is to improve these features by finding methods to reduce the noise effects caused by asymmetries and equal weighing of the data values and obtain an accurate relationship between the variables.

5.1.3 Data Treatment

Although collection of Chinese trade statistics was relatively easier than provincial statistics, during pre-processing we realised that there were wide discrepancies between Chinese and trading partner's export data sets. This section details all the data processing and is divided in three parts:

- *Collection of the Data.*
- *Reconciliation of Data Inconsistencies.*
- *Data Pre-processing*

Collection of the Data

The selection of a satisfactory data collection strategy is extremely important. This is stressed even more here since the resulting symbolic equation of bilateral trade actions is solely dependent on the data inputs provided and thus misleading inputs could confuse the GP process in finding a credible solution.

Initially the data collected was for 217 countries from 1978 to 2001. The data collected from 1978-1996 is just referenced. The data employed for the GP includes 193 observations for all variables for 49 countries from 1997 to 2001. Appendix B.3 gives analytical tables of the data sets used. Trade data between 1978 and 1997 was obtained from the "Direction of Trade" (DoTS) data set developed by the International Monetary Fund (IMF), as well as the Organisation for Economic Co-operation and Development (OECD) and the Ministry of Foreign Trade and Economic Co-operation (MOFTEC) in China. The IMF data set covers bilateral trade between 217 country codes from 1948 to 1997. There are though many gaps in the data and not all areas are countries in the conventional sense including colonies, countries which recently gained their independence, territories and so forth. Appendix B.3 lists the 50 trading partner countries selected for this study. Countries were selected according to the volume of trade they had with China. Maddison [Madd01] collected and analysed a large data set for the OECD Development Centre Studies from which the GDP and GDPPC data for years 1978-1998 was obtained. Trade data (imports and exports) for years between 1997 and 2001 was obtained from the Global Trade Atlas supplied by the Global Trade Information Services and balance of trade was computed from these figures. GDP and population data from 1997 to 2001 was collected from the World Development Indicators supplied by the World Bank [WDI9701]. The size of a trade flow is measured in this study at the point of export. The distance measure between two countries is computed in nautical miles by the shortest navigable distance between the main ports of the respective countries.

Reconciliation of Data Inconsistencies

Due to wide discrepancies in bilateral trade data compiled by China and by its trading partners, it was necessary to make some adjustments to the data sets. These sharply different public perceptions of China's trade relationship with the rest of the world based on misleading figures have created serious policy conflicts [Rusk03]. There is persistent doubt about the quality of Chinese data, but few proposed solutions to this problem have been suggested. The Sung-Lardy [Sung91] method has tried to solve this problem, although there are some limitations in the methodology which include:

- The exclusion of indirect trade through Hong Kong (HK), assuming that Chinese exports to HK are equivalent to HK's total imports from China, which include retained imports and re-exports;
- Respective import figures include both direct and indirect trade requiring that Chinese or partners' imports from HK only match HK's domestic exports.
- Disregarding the difference between an export flow and the corresponding import flow arising from transit lag and cif (cost insurance freight).

Huang and Broadbent [HuBr98] develop a methodology to provide more accurate estimates for bilateral trade flows, extending the Sung-Lardy [Lar94], [Sung91] method and achieving reconciliation of the two data sets by China and by its major partners. This methodology investigates the main reasons behind these discrepancies, taking into consideration the likely distortion on both Chinese and partner data sets and including indirect trade via Hong Kong. The determination of Hong Kong's role as an exporting channel for Chinese goods is achieved using HK figures. It also estimates a new re-export margin (value-added trade) on Chinese exports and takes into account valuation and transit lag when comparing export and import series. This model has been used in this thesis to reconcile the inconsistencies in the data sets and to provide the proposed system with the revised observations in order to get rules that accurately express the Chinese trade environment.

The two formulas used to adjust the export figures are given below and the results obtained from them using our data sets are analysed in section 5.2.2. The first equation (Eqn 5.1) is employed to estimate total Chinese exports from trading country import data and the relevant Hong Kong (HK) data. Chinese exports to its partner are given by the partner's imports from China plus a proportion of the difference between the partners' imports from HK and HK's domestic exports to that trading partner. This difference is part of HK's total re-exports to the trading partner with the proportion obtained by the Chinese total re-exports share.

$$\hat{X}_{C,A} = M_{A,C} + \left(M_{A,H} - DX_{H,A}\right)\left(\frac{RX_{H(C),A}}{RX_{H,A}}\right) \qquad \text{(Eqn 5.1)}$$

where A is the partner country; C is China; H is Hong Kong; $\hat{X}_{C,A}$, is C's total exports to A; $M_{A,C}$, is A's imports from C; $M_{A,H}$ is A's imports from H; $DX_{H,A}$, is H's domestic exports to A; $RX_{H(C),A}$, is H's re-exports from C to A; and $RX_{H,A}$, is H's total re-exports to A.

The result of the second equation should give a similar estimate to the first equation for our results to make sense. We compute the second formula from Chinese exports data plus a proportion of the difference between China's exports to HK and HK's retained imports from China. The difference between the above two is part of China's re-exports to the world and the proportion is given by the partner country's share in total Chinese re-exports to the world.

$$\hat{X}_{C,A} = X_{C,A} + \left(X_{C,H} - RM_{H,C}\right)\left(\frac{RX_{H(C),A}}{RX_{H(C),A\&B}}\right) \qquad \text{(Eqn 5.2)}$$

where $X_{C,A}$, is C's exports A; $X_{C,H}$ is C's exports to H; $RX_{H(C),A}$, is H's re-exports from C to A; $RM_{H,C}$, is H's retained imports from C; and $RX_{H(C),A\&B}$, is H's re-exports from China. In addition the data is refined; firstly the inclusion of the re-export margins on the goods that pass through HK and the differences between recorded imports and exports due to the transit lag and differing valuation basis. These two equations are used here to reconcile the inconsistencies in the data sets used by the GP system. Tables 5.1 and 5.2 provide results computed for the period 1997-2001 from the Huang-Broadbent model for two of China's major trading partners: Japan and US.

JP trade (US$m)	$M_{A,C}$ Raw Data	$X_{C,A}$ Raw Data	$X_{C,A}$ Formula 1	$X_{C,A}$ Formula 2	$X_{C,A}$ Mean
1998	44,231	29,692	38,071	32,546	35,308
1999	44,632	32,399	38,622	35,044	36,833
2000	54,391	41,654	48,268	45,029	46,648
2001	64,326	45,078	57,213	48,966	53,089

Table 5.3 - Export Data Reconciliation for Japan (1998-2001)

US trade (US$m)	$M_{A,C}$ Raw Data	$X_{C,A}$ Raw Data	$X_{C,A}$ Formula 1	$X_{C,A}$ Formula 2	$X_{C,A}$ Mean
1997	62,600	32,702	50,155	45,890	48,022
1998	71,200	37,975	57,976	50,182	54,079
1999	70,000	41,945	55,229	53,227	54,228
2000	100,100	52,104	85,298	65,262	75,280
2001	102,300	54,318	84,027	67,645	75,836

Table 5.4 – Export Data Reconciliation for US (1997-2001)

This process was commenced for 50 countries, and from it more consistent values for total exports $X_{C,A}$ from China (C) to a specific trading partner (A) were obtained. Following these calculations, the entire data sets for GDP, GDP per capita, distance, import and the reconciled export figures for the 50 trading partners from 1997 to 2001 were pre-processed. The procedure employed to pre-process the data sets is discussed below.

Data Interval Coding

Once the data has been reconciled it is necessary to use a method to pre-process it effectively. The data consists of values of significantly differing magnitudes, including some of the poorest developing countries and highly advanced and industrialised countries. It is necessary to sort the data in such a manner that no useful information for the GP is lost and to ensure that the asymmetry of the data values that creates considerable noise does not control the final form of the solution thus giving a misleading symbolic relationship. The data values of the selected variables largely determine the success or failure of the system. Class intervals or bins are used to divide the data into discrete cut-off values for each variable starting with the range minimum and successively adding the class interval to it [Guj03]. The reason for using class intervals is due to the different magnitude of China's trading partners, i.e. in 2002 China's exports to Japan were 48,483billion US$ and to Uzbekistan 104million US$. The negative effect of this different magnitude was emphasised when regressions were performed using the econometric package E-views. A sample of the results was provided in section 5.1.2.

Each class interval is divided as follows; code 1 contains the lowest 5% of the data values for GDP increasing to code 25 that includes the highest 5% of the specific parameter. A percentile data analysis tool was used to rank the data and assist in determining the number of class

intervals, which was decided for observation of the percentile ranking as 25. The same procedure is employed for GDP per capita, imports and exports trade flows. All variables are coded in increments of 1/25, in order to maintain consistency. The distance parameter is also computed in the same manner however the class intervals differ since distance between countries does not change over the 5-year period selected. Consequently, for the distance measure, code 1 is near the lowest 2% of the historical range of the data values and code 25 includes the highest 2%, i.e. Tajikistan in 1997 with 2 million US$ exports from China is in code 1 and the United States in 2001 with 54,319billion US$ is in code 25.

5.1.4 Logic Combination of Trading Rules

In this part of the method we develop a set of rules that allow for the possibility of extracting complex logical relationships between the four independent variables expressing over or under – traded behaviour of export flows. The task is to encode China's trade over the 5-year period using the codes obtained from the pre-processing of the data in the previous section. We combine the four conditions defined below with logic gates AND, OR, NOT to form sets of rules according to the formation of the four variables in our data set.

The following 4 conditions are defined that have to be true for an increase in partner export trade $\ln(X_{ijt})$:

- ***(M1)*** *Model1*: Trading partners have high incomes $\ln(Y_i Y_j)_t$.
- ***(M2)*** *Model2*: Trading partners have high incomes per capita $\ln(YP_i YP_j)_t$.
- ***(M3)*** *Model3*: Trading partners are closer by $\ln DST_{ij}$ or are neighbours.
- ***(M4)*** *Model4*: Trading partners export high $\ln(M_{ijt})$ (imports to China).

Figure 5.1 shows a tree-structured program of an example trading rule. The 3 internal points are labelled with logic operators {AND, OR, NOT} and the 4 external points are the four model conditions graphically depicted as follows:

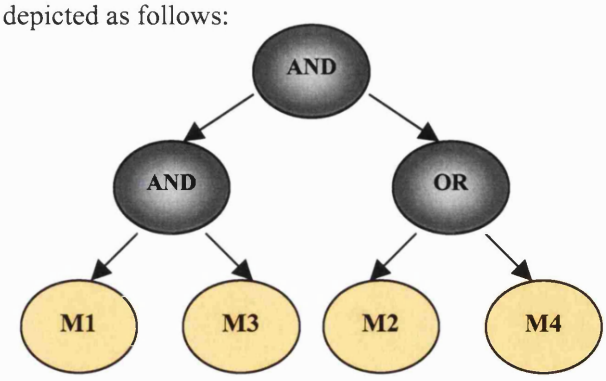


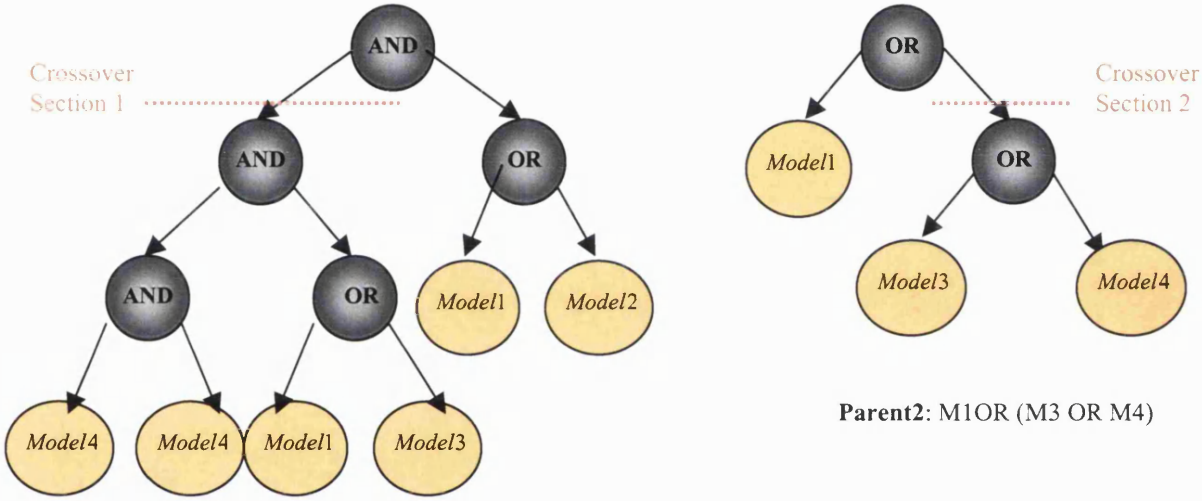
Figure 5.1 – Tree structured representation of example trading rule

This can be expressed in the following pre-order traversal format (discussed in tree representation in chapter 4):

(AND (AND M1 M3) (OR M2 M4))

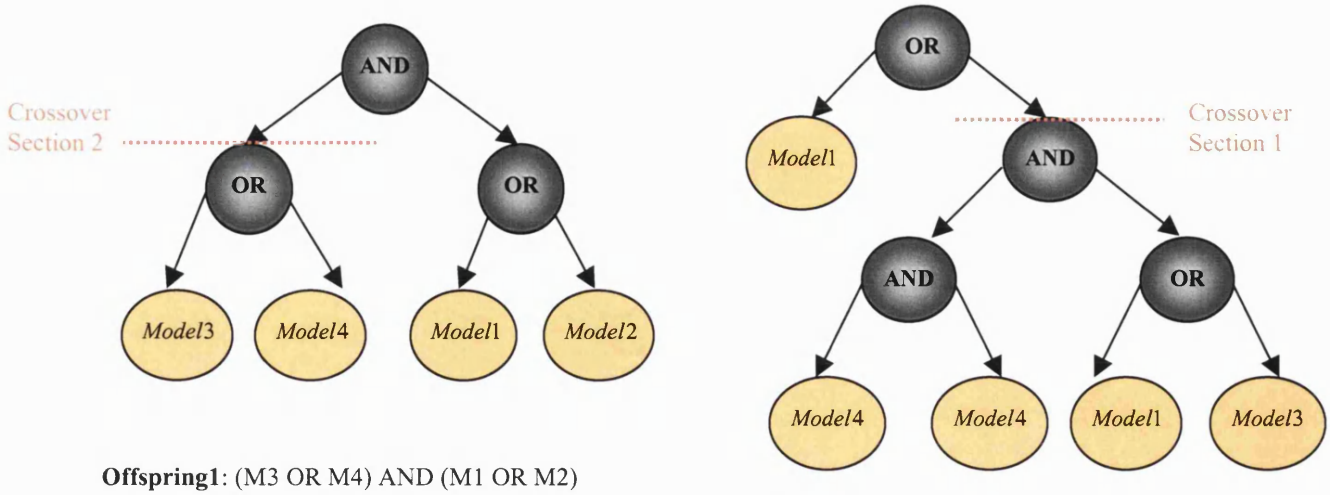
From the above logic combination of the 4 models we decode the following trading rule:
Trading partners that have high incomes and are neighbours and have high income per capita or high imports will trade high else they will trade low.

During the GP process, operations including crossover and mutation take place. These are graphically depicted in figures 5.2 and 5.3 respectively.



Parent1: [(M4 AND M4) AND (M1 OR M3)] AND (M1 OR M2)

Parent2: M1OR (M3 OR M4)



Offspring1: (M3 OR M4) AND (M1 OR M2)

Offspring2: M1 OR (M4 AND M4) AND (M1 OR M3)

Figure 5.2 Crossover operation for trading rules resulting in Offspring 1 and Offspring 2.

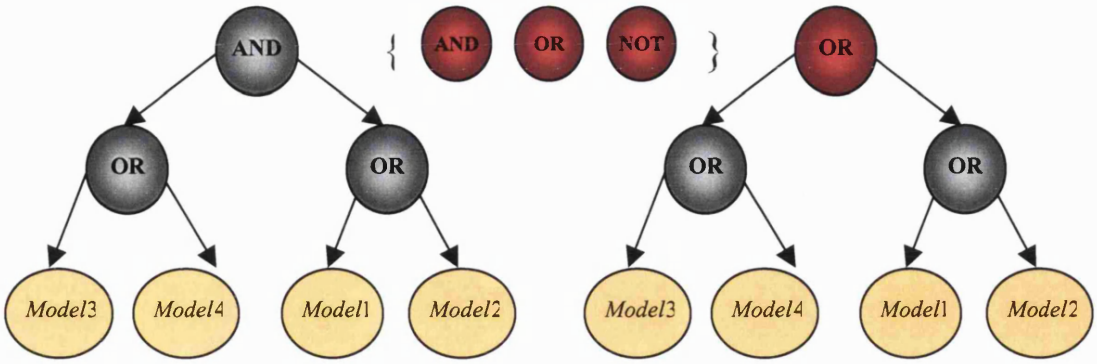


Figure 5.3 Mutation of trading rule from AND->OR logic gate

In Figure 5.2 an example of a possible crossover between two trading rules was demonstrated. The two parental trading rules are combined to produce two new rules. Depending on the fitness crediting each new rule is assigned, it will be selected again or be dismissed from the population if it has a very low fitness. The mutation operator is used less frequently in the GP run, figure 5.3 shows the power its effect has to either assist the GP to produce a better or worse rule. It changes the rule expression from considering two sets of conditions i.e. M3 or M4 and M1 or M2 to considering only one of the four conditions (M3 OR M4 OR M1 OR M2) for the specific trading rule as being true for higher trading actions.

5.2 The GP Environment

This section introduces and discusses the modelling of the GP environment. Initially the software used in order to implement the GP is introduced together with additional features coded for the specific parameters necessary for the specified application. The selection of the parameters including the terminals and functions, the fitness function, population size, and genetic operators is also discussed.

5.2.1 GP Implementation

For the GP implementation a GP library written by Bill Langdon at University College London was used as the basis. Features specific to our application were added. This software was used for the following reasons:

- It is a public domain software
- Since it is written in C++ code, negligible changes were made to functions created in earlier programs.
- It is stable.
- It has a flexible structure so new features were easily added to it.

Several features were added to the software, listed below:

- Extended selection of functions set to include logic gates, AND, OR and NOT.
- Changed the function that calculates fitness to count the number of individuals and then give a resulting fitness.
- Initially only one type of terminal set was available, it was extended to include combinations of letters and numbers especially useful when GP deals with breeding populations of trading rules.

5.2.2 Selection of Parameters

Following the coding of the variables describing trading actions we are now faced with the problem of finding the desired solution in the space of possible tree-structured solutions. In this section the selection of parameters that assist the GP in this process are outlined since it controls which structures from the search space will be selected, modified or improved to assist in finding the best program for our application.

Terminals and Functions

The GP environment is provided with a set of cases which forms the basis for evaluating particular tree-expressions. The terminal set is: T = {M1, M2, M3, M4}. These terminals correspond to the independent variables of the model and are constructed according to the values in our data set. The unknown functional relationship could involve a combination of economic functions combining the four independent variables and the dependent variable. The GP is not informed of the nature of the relationship between the dependent and independent variables, being able to guide the process by itself. In this work we include only logic gates to make the process easier, so the function set is: F = {AND, OR, NOT} which when merged with T, gives the following: $C = F \cup T$ = {M1, M2, M3, M4, AND, OR, NOT}.

Fitness Function

Each individual in a population is assigned a fitness value as a result of its interaction with the environment. The raw fitness of a computer program is measured by the number of trading rules (out of a possible 193) that it satisfied at the end of each generation. The program that gets the most trading rules is the fittest individual. The standardised fitness, acts as the error, and is calculated as the total number of fitness cases minus the raw fitness.

Parameters

The following parameters have been used in the GP environment: 500 individuals in the population; 90% of the new individuals are created by crossover; of the other 10%, 99% are direct copies from the previous generation and the remaining 1% are mutated copies. The GP is run for 51 generations (an initial random generation called generation 0 plus 50 subsequent generations). A maximum tree depth of 25 was established, preventing large amounts of computer time being expended on a few extremely large and unfit solutions. Table 5.5 lists the selected parameters.

Objective	Find the over or under traded relationships for the Chinese trade environment.
Terminal Set	M1, M2, M3 and M4
Function Set	AND, OR, NOT
Fitness Cases	193 trading rules (fitness cases)
Raw Fitness	Number of trading rules satisfied at end of generation.
Standardised Fitness	Total number of trading rules minus raw fitness.
Hits	Equals raw fitness
Parameters	P = 500, G = 51, Max initial tree size = 25, crossover 90%
Success predicate	A tree expression solution that scores 193 hits

Table 5.5 – List of Selected Parameters

Different initial conditions were used, such as:

- Reducing crossover from 90% to 60%, the remaining 40% was 50% direct copies and 50% mutation, which improved the GP's performance since it took less time to find the tree expression solution that scored 193 hits.
- Reducing tree depth from 25 to 15 was very useful since the number of trading rules that had numerous branches with no new information decreased significantly although the GP did not reach a solution quicker with this alteration.

5.3 Determining China's trade environment

Koza [Koza92a] set the basis for GP applications in economics by evolving the entire exchange equation to demonstrate its capabilities as a knowledge discovery tool. In addition, to the rediscovery of Kepler's 3rd law, GP also discovered earlier conjectures that the mathematician considered, which were decoded from interim solutions with high fitness. GP contributes significantly to economic applications since it is an intelligent technique that evolves hierarchies of subroutines (building blocks) from an infinite information search space to continuously improve its selections and results. One of the most valuable uses of GP is not the explicit form of the rules produced but rather the variables that the system highlights by occurring in many fit rules even when starting from a variety of different initial conditions.

Finding a mathematical relationship for a scientific process has always been an interesting yet difficult problem to solve. This thesis employs observed data and the theory of the gravity equation to demonstrate how to discover the relationship that identifies China's trade environment. The gravity equation states the relationship between GDP, GDP per capita and geographical distance. The imports trade flow is also added to these three variables due to the direct effect a bilateral trade imbalance has on trade behaviour. The relationship between the trade flows and the various explanatory variables is usually estimated by Ordinary Least Squares (OLS) regression methods. Similar to most conventional economic techniques, OLS tries to find the values of the coefficients and constants required that best fit the gravity equation. However, it is equally important to use the observed data to find the functional form that best expresses the interactions between these variables.

Although there are statistical techniques that can generate similar results concerning the variables, GP identifies the main symbolic relationship that expresses China's and trading partner's actions from an infinite bilateral trade flow data search space. In addition to finding the main relationship, the system has also discovered secondary solutions for these actions which the GP has partially credited due to their high fitness trying to demonstrate some the many components in China's trade. These secondary solutions are thought as partial relationships that influence the actions in this environment but to a lesser extent or even try to improve the main relationship by indicating the necessity for additional variables are needed or different combinations of the existing variables.

In this section we discuss our results for the main relationship portraying trade actions as well as partial equations according to GP computations. We also discuss the performance of the GP measured by fitness for different generations.

5.3.1 The Symbolic Expression of China's Bilateral Trade Actions

The main relationship was computed using the models and codes designed and explained in section 5.2.3. The significance of the gravity equation is unquestionable as an important and widely used tool in economic analysis, however the equal weighing of all data inputs results in coefficient values that are mainly dominated by minor trade flows. Our aim is to eliminate the noise effects and asymmetries caused by this equal weighing of the data which tends to control the final form of the solution thus giving misleading relationships for the variables.

The main relationship from the best-of-run individual is given by the following expression:

(OR (AND Model2 Model3)(AND Model1 Model4))

which has a standardised fitness of 0 and is true for all 193 cases. This individual is equivalent to the following trade relationship:

Model2 AND Model3 OR Model1 AND Model4

which is translated as follows:

Trading partners will trade high either if they have high income per capita and are neighbours or if they high imports and high income, else they will trade low.

Figure 5.5 graphically depicts the above rules as a rooted, point-labeled tree with ordered branches.

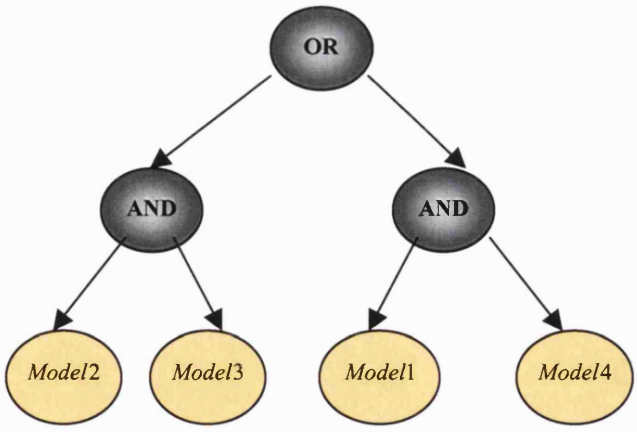


Figure 5.5 –Graphical representation of main trade actions relationship

This equation gives the main relationship between partners in China's trade environment and confirms the gravity equation theory which states that trading partners trade more when they have higher incomes, income per capita or are neighbours. In addition the GP gives a positive

relationship between export and import trade flows, i.e. high exports lead to high import trade behaviour from trading partners and vice versa. The GP resulted in this relationship in many different initial conditions and tests introduced in section 5.2.2 which confirmed its dominance in the first tests. It is the best of run individual tree expression with standardised fitness = 0 for 193 cases. The GP credited some solutions with high partial fitness throughout the run. We selected the two with the most interesting features describing the variables and discuss them below.

5.3.2 Secondary equations

We decided on computing secondary equations for China's trade expression for two reasons:

1. *The discovery of interim partial solutions with high fitness by the GP for Kepler's 3rd law.*
2. *The multiple dimensions of international trade actions and even more so when measuring China's trade environment.*

The first partial expression with standardised fitness = 9 which is equivalent to being true for 184 cases is given by the following expression:

(OR(AND Model2 Model3)(AND Model1 Model4)(OR (AND(NOT Model4 Model3))))

This individual is equivalent to the following trade relationship:

[(Model2 AND Model3 OR Model1 AND Model4) OR (Model4 AND (NOT Model3))]

which is translated as follows:

Trading partners will trade high if they have high income per capita and are neighbours or if they export high. Trading partner will trade high if they export more and if they are not neighbours.

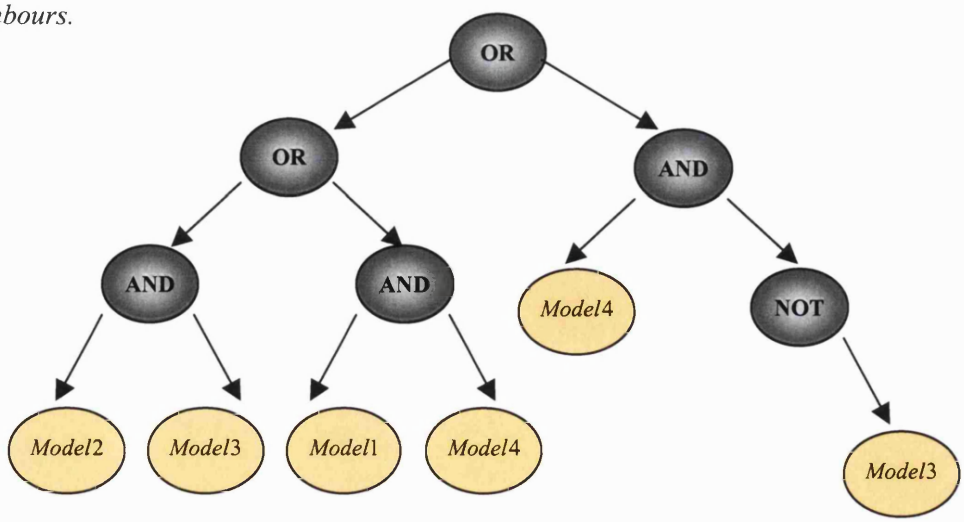


Figure 5.6 Graphical Representation of secondary equation no.1

This equation includes the basic structure of the main relationship and confirms it. In addition, it gives a very interesting result which stresses that China's trading partners will be affected more by their higher exports to China even if they are far away, thus describing different aspects of the gravity equation. The inclusion of M4, the increase in imports in the right as well as the left subtrees of the trading rules indicates the dominant features of this variable on the trade expression affecting the relationship for 184 situations more than any other single variable.

The second partial expression with standardised fitness = 132 for 61 cases is given by,
(OR (AND Model4 Model4) (AND Model2 (AND Model3 (NOT Model1 Model1))))

This individual is equivalent to the following trade relationship:

Model4 AND Model4 OR Model2 and Model3

which is translated as follows:

Trading partners that export high will trade even higher or trading partners that have high income per capita and are neighbours will trade more, else they will trade less.

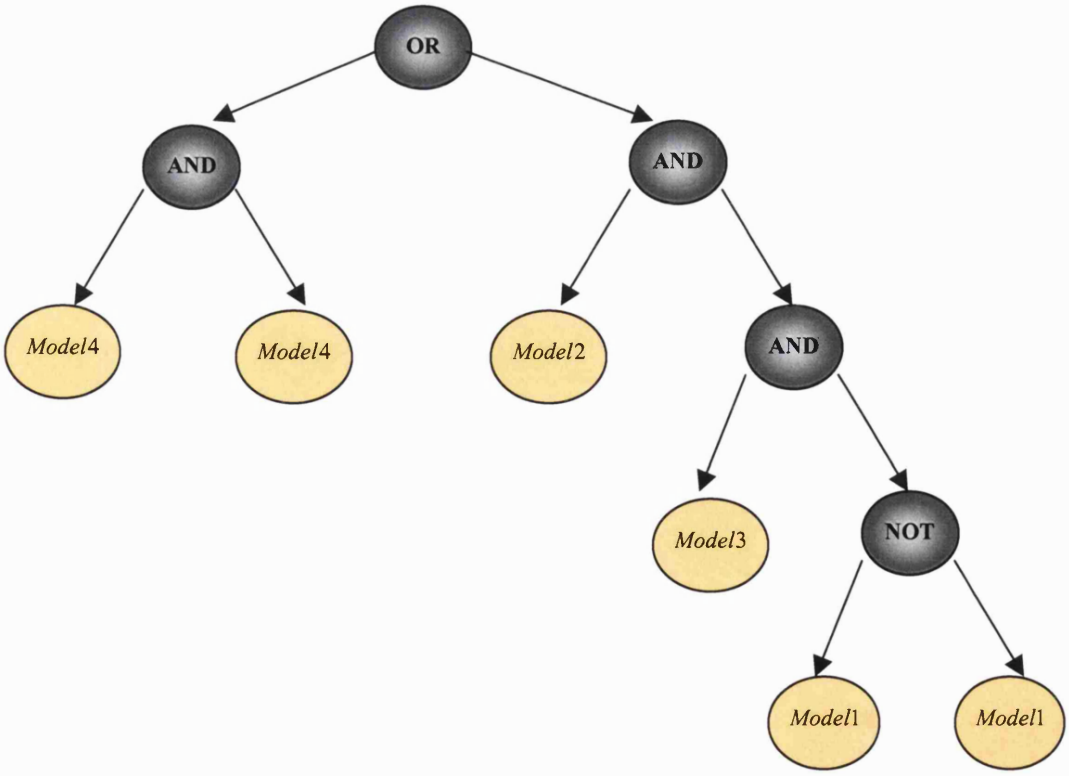


Figure 5.7 Graphical Representation of secondary equation no.2

This equation fits only 61 cases however it was included six times in the specific run. This is the only equation from the three selected that does not include one of the main components of the basic structure of the main relationship. M1, indicating high income, is cancelled out, resulting in a rule that is once more mainly affected by balance of trade and then by the main components of income per capita and geographical distance.

Koza [Koza89] has researched this effect of partial crediting. This thesis goes a step further in finding these equations that get partial crediting. The results obtained from the GP run show that there is a significant potential of exploiting the gravity equation. Both the main symbolic relationship and the secondary ones include the basic structure of the gravity equation for China and confirm it. In addition, the secondary equations indicate variations to the equation's main format reflecting the different trading behaviour among groups of China's partner countries.

5.3.3 GP Performance

The fitness of each computer program was measured by the number of trading rules it satisfied. Each trading rule represented an individual trading pattern for China's trading partners for each year. If a rule was repeated more than once for a country it was eliminated since that would guide the GP to indicate a relationship for China's environment based on numerous identical cases rather than equally assess the whole range of cases for all countries. The standardised fitness was used to calculate the error, i.e. the share of total rules that the program got wrong. The lower the standardised fitness the higher the program performed.

We performed quite a few runs with different initial conditions in order to select the run with the most interesting and distinct features in its generations. The initial population was selected randomly with no fitness function to guide the process by selecting the better performing individuals and as expected was quite unfit.

- *Generation 1*: the best-of-generation individual has a raw fitness of 10 (i.e. 10 hits from a possible 193) and a standardised fitness of 183.
- *Generations 2 and 3*: this fitness remained the same with no actual improvement and the only variation being that more individuals scored 10 hits than in the first generation.
- *Generation 5*: By the end of this generation half of the individuals (total of 500 individuals per generation, 50 generations per run) scored 12 hits with a standardised fitness of 181, indicating a very high error and unfit individuals.
- *Generation 10*: the individuals were scoring between 57 and 68 hits with standardised fitness of 136 and 125 respectively. This is the generation where the second of our selected secondary equations with 61 hits scored appeared six times.

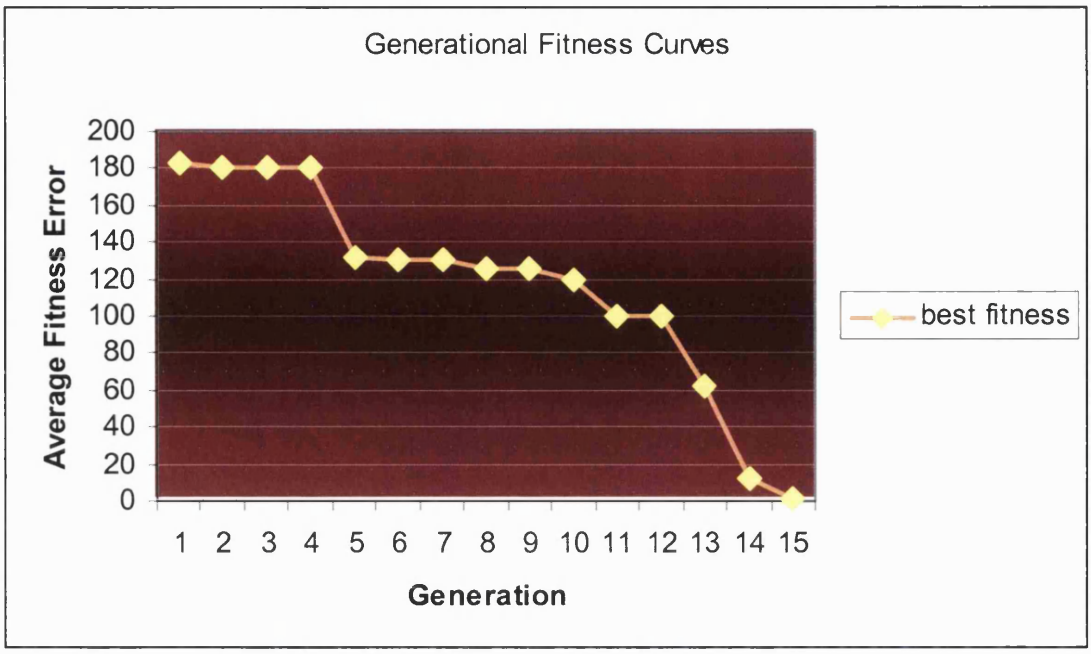


Figure 5.8 – Performance Graph for consecutive generation of selected run

- *Generation 15:* the best-of-run individual was found scoring 193 hits with zero standardised fitness. This individual was detected four times in the last generation. It satisfied all conditions for the trading rules and gave the GP's explanation of China's bilateral trade environment. The first secondary equation scoring 184 hits was found in this generation. Figure 5.8 gives a graphical representation of the evolution of these generations.

5.4 Summary

This chapter introduced the second part of the GTCA employing genetic programming to find the relationship expressing an increase or decrease in the factors affecting China's bilateral trade system. Instead of finding the coefficients determining a particular economic model, populations of candidate trade rules that best fit the data were genetically bred to identify the symbolic form of these trade actions. Bilateral export data inconsistencies were reconciled and class intervals were used to eliminate the noise and asymmetries created by trade flows of different magnitudes from developing and developed countries. The theoretical basis of the gravity equation was confirmed for China's bilateral trade with the selected countries and a positive relationship between import and export trade flows was identified. Initial results were confirmed further by the secondary equations selected due to high partial fitness crediting, which depicted a high dependence on the balance of trade among countries and a lesser affect of distance and for some cases an elimination of the importance of high income.

Chapter 6

The GTCA Model: A Case Study

Chapter 6 combines the algorithmic components explored in Chapters 3 and 5 that provide the basic structure of an intelligent country analyser encapsulating direct visualisation and location of complex unstructured correlations with the search for mathematical relationships determining evolving environments. The Growth and Trade Country Analyser (GTCA) is tested on its ability to investigate income convergence in the European Union countries firstly in terms of the entire systems model and then with its separate components. The performance and applicability of the model are tested followed by a critical assessment of the analyser and its components according to a set of criteria. Finally, the research contributions presented in section 1.4 are revisited.

6.1 Experimental Design

In Chapter 3 and 5 we introduced intelligent growth policy mapping and the evolution of bilateral trade flows respectively, designed and built as the two separate components of the Growth and Trade Country Analyser (GTCA). In this chapter these components are used to analyse and predict a specific macroeconomic problem. The aim in this chapter is to:

- *Conduct another experiment of a similar type with an independent data set in order to assess the GTCA systems model applicability and capabilities and refine its components.*
- *Compare results for independent data with results in Chapters 3 and 5.*
- *Assess the analyser and its components on a set of criteria.*

6.1.1 The Case Study

The separate parts of the GTCA are tested on the income convergence types and trade dependency of European Union countries. The first part of the GTCA using SOMs is employed to determine the trend and relative importance of evolution of income disparities over time across EU countries. The second part of the GTCA, employing genetic programming is tested using data sets for domestic demand growth and exports as percentage of GDP to identify the relationship between the features affecting trade dependency.

An article analysing convergence by Barro and Sala-Martin [BarSiM92] examining forces that lead to convergence between countries over time in the levels of per capita income and product sparked most of the literature in this field. Barrell and Dury [BarrDu01] argue that it is not possible to have one European main group and divides the European economies into a core group consisting of Germany, Netherlands, Austria and France and another periphery group consisting of Italy, Spain and Portugal. De la Fuente [deFue00] and Luginbuhl and Koopman [LugKo03] confirm the group division of EU countries by observing different strategies of the EU countries in the euro zone economy and testing for convergence in GDP series of five European countries, respectively.

Trade dependency theory states that rich and powerful countries collectively form the core of international trade with the poor countries forming the periphery since their well-being depends from trade with these countries. This creates an unbalanced and unfair trading environment. The core produces more luxury goods, while the periphery specializes in basic and industrial goods. Cardoso and Faletto [CarFa69] were the first to state the definition for dependency theory arguing that economic development frequently depends on favourable conditions for exports. Trade dependency is affected by exports as a percentage of GDP and domestic demand growth. As the exports percentage of GDP increases the dependency of trade increases since a large amount of GDP is contributed by this factor. In contrast, if a country's domestic demand grows the trade dependency decreases since the country depends more on its domestic market.

6.1.2 Data Collection

The data collected is for the last five years (1997 to 2001). It includes data sets for the original 15 European Union members, since the 10 new members joined only recently (2002). The 15 EU member countries are: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Portugal, Spain, Sweden, the Netherlands and the United Kingdom (UK). We compute income convergence from GDP figures for total output, primary, secondary and tertiary sectors. Population figures were all obtained from the World Indicators supplied by the World

Bank. Trade dependency is expressed in exports as a percentage of GDP and domestic demand growth. The figures for these measures were obtained from the Economist Intelligence Unit (EIU).

Unlike Chinese data, the data sets for this experiment were of good quality and easy to find. The Economist Intelligence Unit data sets for trade dependency had no gaps; however, the World Indicators reports were missing data for some indicators or years mainly due to the fact that the country did not supply the relevant information for the specific years. This is a minor problem since the SOM will still characterise the convergence type of any EU country. A list of the missing observations is provided, both for assessing the SOM's performance once more on missing data and consider it when analysing the mappings. The missing observations are:

1. Denmark, Greece, Ireland, Portugal missing observations for 1998 (Portugal also missing observations for 2000). Sweden and Luxembourg missing observations for 1997, 1998, 1999.
2. France missing manufacturing figures for 1999, Germany missing industry figures for 1997 and Spain missing industry figures for 1997 and 1998.

6.1.3 Methodology for Experimental Results

The methodology is as follows:

1. GDP total output data will be used as reference data together with data values for 2002 to check for correct or misleading analysis.
2. Both systems will be tested on the correctness of the cluster formations measured by topographic and quantisation errors for the SOM and the fitness performance in each generation for the GP.
3. The SOM-provincial growth design will be validated using data from 1997 to 2001 on per capita GDP analysis of the primary, secondary and tertiary sectors. Its performance will be tested on how well it characterises convergence types for the EU countries.
4. The GP system will be validated using data from 1997 to 2001 for exports as a percentage of GDP and domestic demand growth. The GP's performance will be tested on its ability to discover the relationship between these two variables which determine trade dependence.

6.2 The GTCA Environment

The GTCA environment uses SOM-based growth design to characterise growth types and genetic programming to discover unknown interactions between the variables of established symbolic mathematical equations. For the validation problem the GTCA has to characterise and identify the types of growth convergence present, and also to discover the relationship for trade dependency among EU countries.

6.2.1 Phase I

In this section we are measuring convergence from the production standpoint where GDP is the summation of the value added of national sectors of agriculture, industry, manufacturing and services. [BarSM92] define convergence as the possible trend of reduction over time of income disparities across countries. We will use this definition for our validation example since we are not comparing business cycles or nominal variables like inflation and interest rates that are required for convergence under the Maastricht Treaty but rather measure convergence of per capita income and product according to the sectoral breakdown of per capita GDP. For this definition we assume that if there is convergence in a given set of countries then poorer economies tend to grow faster than rich ones reducing the income differential between them, and if there is divergence rich countries grow faster, increasing their lead. Convergence in the EU is an ongoing process that started before the beginning of the sample period and is not yet completed for all countries. Consequently, the type and rate of convergence are measured by the type of sector mostly affecting this process and its growth rate.

Per Capita GDP in primary, secondary and tertiary sectors - 2001					
Country Keys	Country	Agriculture	Industry	Manufacturing	Services
AT01	Austria	0.464	7.664	5.109	15.096
BE01	Belgium	0.444	5.998	4.443	15.773
DK01	Denmark	0.894	7.750	5.067	21.163
FI01	Finland	0.698	7.682	6.052	14.665
FR01	France	0.665	5.759	3.987	15.948
DE01	Germany	0.226	6.999	5.419	15.353
GR01	Greece	0.887	2.329	1.331	7.874
IE01	Ireland	1.073	11.269	8.855	14.758
IT01	Italy	0.566	5.469	3.961	12.625
LU01	Luxembourg	0.442	8.389	-	36.645
NL01	Netherlands	0.722	6.494	4.089	16.836
PT01	Portugal	0.427	3.203	2.029	7.047
ES01	Spain	0.581	4.359	2.761	9.590
SWD01	Sweden	0.491	6.623	-	17.416
UK01	United Kingdom	0.239	6.445	4.535	17.186

Table 6.1 – Per capita GDP Sectoral Breakdown

Table 6.1 shows per capita values for the main national sectors for 2001. This is a sample of the actual data set used which was for five years, from 1997 to 2001, found in Appendix B.4. Three different mappings have been computed for this analysis, these are: the main SOM map (figure 6.1) dividing the map in zones and showing graded relationships of growth between the countries; a bar chart (figure 6.2) with the amount each country produces in each sector denoted by the height (percentage of total value from all observations for sector) of each indicator in the bar chart representing each neurone; and the relative importance pie chart (figure 6.3) that shows the level of dependency on each sector and how that affects convergence.

The main SOM in figure 6.1 shows an 8x5 map or 40 neurones representing the data values for the 15 EU countries. The representation of the mapping differs from the main SOM in chapter 3 in that we show information for each of the five years for all countries and all production sectors, instead of taking an average over all years as in Chapter 3. Both map representations are equally important and since the SOM was able to categorise average data for two decades for all provinces and indicators, we thought it would be interesting to evaluate how its performance alters when categorising different years in parallel for all countries and all sectors.

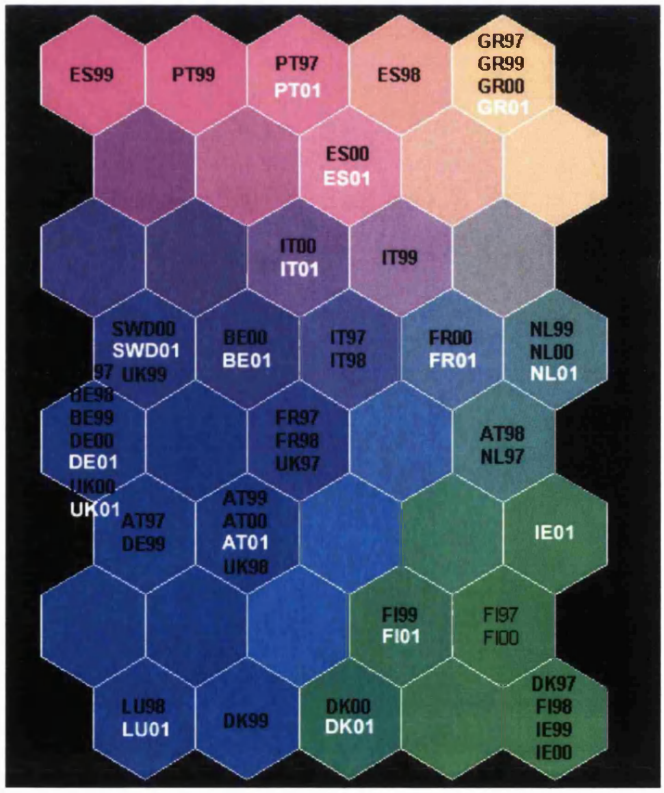


Figure 6.1 – EU Convergence Type Mapping

The clustering dynamics are highlighted by the SOM by categorising income convergence trends for all years. Convergence in 2001 is depicted with white to show the current position of each individual country. As a general observation we can immediately see that there is no single pattern that all countries follow with the map showing a variety of convergence levels. Since we have observations for five years, we can trace each country's convergence or divergence trend on the map. Firstly, we need to understand which sectors affect this trend so we complement this map with a bar chart (figure6.2) for each neurone that shows the amount each sector has in each of the neurones occupied by EU countries and the relative importance pie chart (figure 6.3) – also used in Chapter 3 – to determine the dependency each country has on each national sector, both given below. The neurone positions that the countries occupy in figure 6.1 are the same for both figure 6.2 and figure 6.3.

Figure 6.3 depicts a division of the neurones into three categories, the central section (neurones 16 to 31) where relative importance from the indicators is decreased, the upper section (neurones 1 to 15) and the lower section (neurones 32 to 40) where pie charts size increases. The upper and lower sections of the map depict an increased dependence on the individual sectors with larger relative importance pie charts indicating low convergence levels. In contrast, the countries occupying the central part of the map have limited dependence with the smaller relative importance pie charts indicating a stable consistent growth in all indicators as they tend towards convergence.

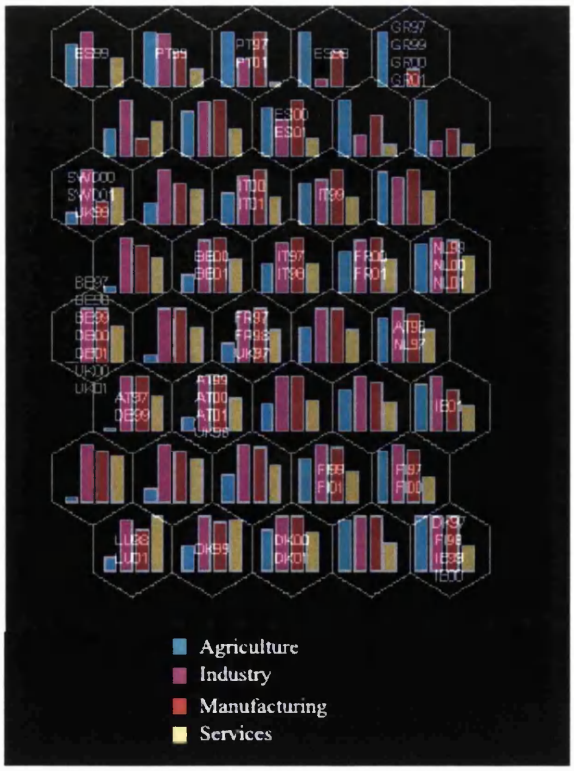


Figure 6.2 – Bar charts of individual sectors

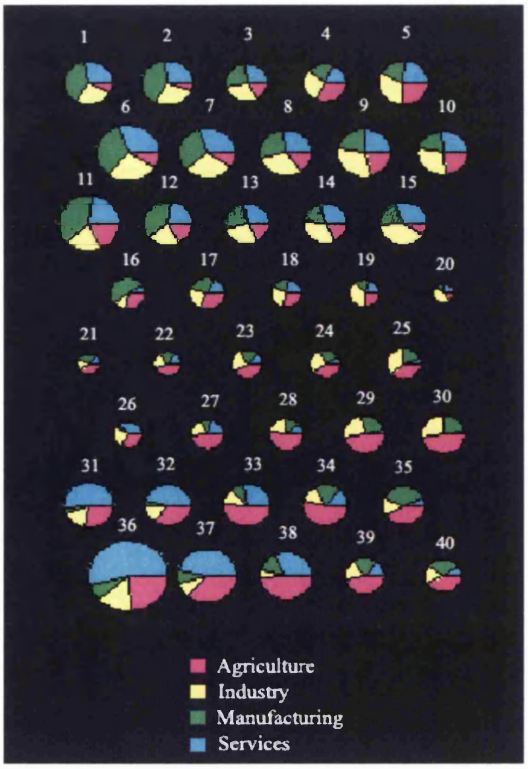


Figure 6.3 – Pie charts of relative importance

Following the general observations about the different clusters is the analysis of individual countries in each of the three convergence types.

Convergence Type 1 *(neurones 1 to 15)*:

This convergence type is represented by the data values in neurones 1, 2, 3, 4, 5, 8, 13, and 14 belonging to Spain, Portugal, Greece, and Italy, depicted by shades of purple and yellow in figure 6.1. Within this cluster the countries that are closer to convergence are Italy and Spain, with Portugal and Greece retaining the same position for the beginning and end of the sample period. Although Italy is the closest to convergence it seems to be moving from convergence type 2 (central section) and neurone 18 from 1997 and 1998 back to this type, with Spain steadily converging from neurone 1 to 8 towards convergence type 2. From figure 6.2 we observe that the sector that characterises this zone is the primary sector (agriculture) with its main representative being Greece. Consequently Greece, Portugal and Spain comprise the zone with an increased primary sector, with Italy moving towards a non-identifiable zone.

Convergence Type 2 *(neurones 16 to 31)*:

This convergence type is represented by the data values in neurones 16, 17, 18, 19, 20, 21, 23, 25, 26, 27 and 30 belonging to Sweden, Belgium, France, Netherlands, the UK, Germany, and Austria depicted by shades of blue in figure 6.1. It should be noted that neurone 18 is occupied by Italy but this was only for the first two values of the sample period. Also Ireland enters this zone in neurone 30, however due to the fact that its other values are in neurone 40 it is not included here. The relative indicator importance position of each country in this type tends to stabilise over time. The main representatives of this trend with central characteristics of the tertiary sector are the UK and Germany with most of their data observations in neurone 21 depicting common converging features with observations for UK97 and UK98 showing this trend.

In contrast, France moves away from the tertiary sector (neurone 23) increasing its industrialised agricultural sector (neurone 19). Belgium leaves the dominant neurone 21 moving inconsistently to a different area of the map to neurone 17. Austria follows the same trend as Belgium, although its move is clearly towards the secondary sector. The Netherlands sustain a balance between secondary and primary sector insisting on an industrialised agriculture explaining most of the country's data observations occupying neurone 20.

Convergence Type 3 *(neurones 32 to 40):*

This convergence type is represented by the data values in neurones 34, 35, 36, 37, 38, and 40 belonging to Finland, Luxembourg, Denmark, and Ireland depicted by shades of green in figure 6.1. Ireland has a strong secondary sector while intensifying the primary sector thus improving its industrialised agriculture moving from neurone 40 to 30 in convergence type 2. Ireland and Denmark are the main representatives of the secondary sector observed clearly in figure 6.2 with the bar charts for industry and manufacturing at their highest. Denmark sustains a strong tertiary sector while developing the secondary sector by moving from neurone 37 in 1999 to neurone 38 for the last two observations 2000 and 2001. Finland clusters all its observations in neurones 34 and 35 thus sustaining its high levels in the tertiary sector. Luxembourg is the special case in all three mappings, in the same manner Hainan province was in the provincial growth mappings. Although it is the main representative of the tertiary sector (services) shown in figure 6.2 characterising countries in convergence type 2 its occupation of neurone 36 with the largest pie chart and highest sector dependence moves it the furthest away from convergence, explaining its positioning far from countries in type 2.

Convergence in 2001

Observing the location of the provinces in 2001 highlighted with white, every country has its own characteristics (there are no two countries in the same neurone) excluding Germany and the UK which seem to form the core of convergence in the EU. Convergence in 2001 is most dominantly formed in neurones 20, 21, 22, and 24, with no data samples located in neurones 20, 22 and 24 and thus neurone 21 occupied by observations for the UK and Germany forming the core of the desirable convergence between the EU countries. Between neurones 16 and 31 forming convergence type 2, the indicator dependence is small with a wider convergence area formed around neurone 21 (UK and Germany) where the following countries are located Austria (27), Ireland (30), The Netherlands (25), France (19) and Belgium (17). Furthest away from convergence are the countries including Greece (5), Portugal (3), Denmark (38), and Luxembourg (36) and less far away are Italy (13), Spain (8), Sweden (16) and Finland (34).

6.2.2 Phase II

In this section the second part of the GTCA is tested that employs GP to find the relationship expressing trade dependence among EU country members. GP genetically breed populations of possible candidate trade dependence solutions and guides the process through which the fitness function selects the symbolic form that best fits the specified observations.

Key	Country	Export % GDP (E%Y)	Domestic Demand (DD%)
AT02	Austria	52,90	1,00
BE02	Belgium	82,40	1,10
DE02	Denmark	35,90	1,00
DK02	Finland	44,80	1,20
ES02	France	28,40	2,60
FI02	Germany	38,70	0,60
FR02	Greece	27,10	1,10
GR02	Ireland	20,50	4,00
IE02	Italy	93,70	2,90
IT02	Luxembourg	24,80	0,80
LU02	Netherlands	145,30	1,00
NL02	Portugal	61,70	1,00
PT02	Spain	30,00	5,00
SWD02	Sweden	43,30	0,60
UK02	United Kingdom	26,10	3,00

Table 6.2 – Export as percentage of GDP and domestic demand growth for 2002

The results from the main functional form determined by the GP are given below. Figure 6.4 depicts the main relationship expressing the data inputs. Secondary equations are used in this application too, since they seem to increase the understanding through highlighting different aspects of the variable that influence the main equation.

The main expression with fitness 1 for a total of 90 fitness cases is given by,

(* (* DD%) (% E%Y DD%) (% E%Y (% (* (* (* E%Y DD%) E%Y) E%Y)

The best-of-run individual gives the following trade dependency relationship for exports as a percentage of GDP (E%Y) and domestic demand growth (DD%):

Trade Dependence = (E%Y) / DD%

which is translated as follows:

Trade dependence increases with an increase in exports percentage of GDP and a decrease in domestic demand.

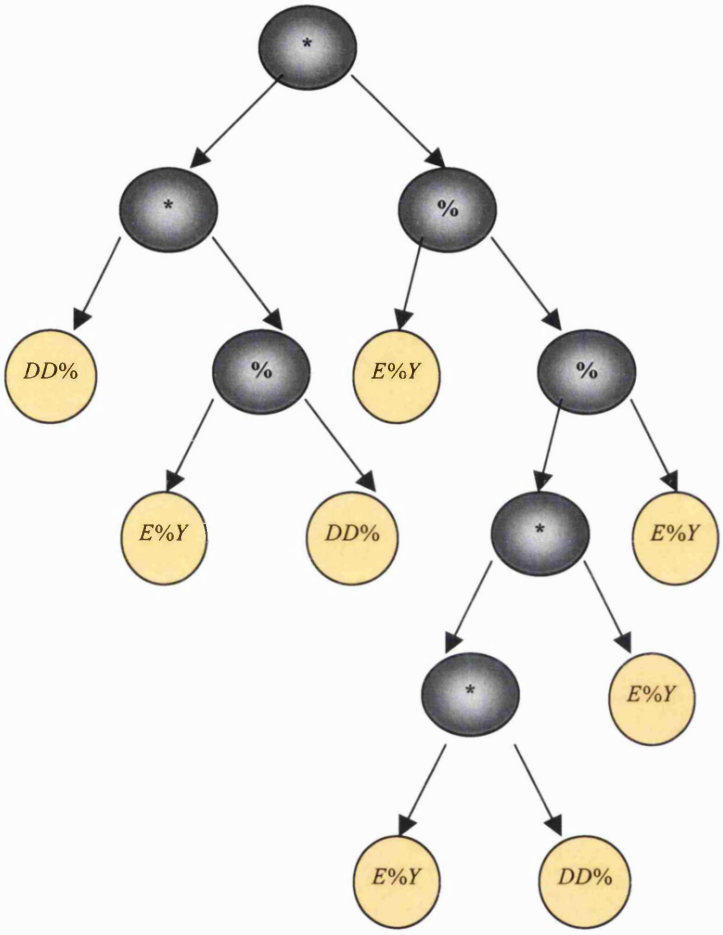


Figure 6.4 –Graphical representation of main trade dependence relationship

The equation for the main relationship is thus given by the ratio between **E%Y** and **DD%** and confirms the general rule that export-driven countries are more dependent on trade than domestic-driven countries. Although the GP found it difficult to find this relationship its dominance was confirmed after several runs. Figure 6.4 shows the best individual for the run with fitness = 1 for all 90 cases. However, the interesting point in this GP test was not the output for the main relationship but rather the evolving format of the secondary relationships. The 3 secondary symbolic expressions with the most striking features for trade dependence given by the GP system have been selected and are discussed below.

1. The first relationship gives $2\left(\frac{E\%Y}{DD\%}+1\right)$ with fitness = 0.92008 and 57 hits. This relationship states that there are factors missing from the current functional form for trade dependence which questions the format and accuracy of the mathematical expression of the variables.

2. In the second relationship $\frac{E\%Y^2}{DD\%}$ is true for 61 hits (fitness = 0.92705). The positive effect that export as a percentage of GDP (E%Y) has on trade dependence has increased immensely (doubled). This is also observed by the fact that this relationship is true for 61 out of 90 trade cases, depicting a higher trade dependency and export-driven countries for these 61 hits.

3. The third relationship $\frac{E\%Y}{DD\%^2}$ with fitness 0.93606 and 62 successful hits depicts an increasing negative effect of domestic demand growth (DD%) on trade dependence. In this case there are 62 cases where domestic demand growth (DD%) has a very high negative impact on trade dependence depicting domestic-driven countries.

These relationships formed by the four different equations introduced above are used by inputting the entire data set for trade dependence which includes exports as percentage of GDP and domestic demand growth for all 15 EU countries for years 1997 to 2002. Each equation gives a resulting value after being tested on the data, the results obtained for each equation are given in table 6.3.

Equation	Fitness	Hits	Results
$\frac{E\%Y}{DD\%}$	0.99971	90	15,8
$2\left(\frac{E\%Y}{DD\%}+1\right)$	0.92008	57	33,6
$\frac{E\%Y^2}{DD\%}$	0.92705	61	771,4
$\frac{E\%Y}{DD\%^2}$	0.93606	62	5,1

Table 6.3 – Results for main and secondary relationships

By calculating the average value for the results that all four equations gave (table 6.3) an average value for exports as percentage of GDP equal to e_{avg} = 48,9 and an average value for domestic demand growth equal to d_{avg} = 3,1, are obtained. By searching all possible solutions and

constantly evolving and improving its search, GP tested and evaluated the relationships between the variables. The results depicted that the strength of the GP methodology is in improving the current functional form of an expression indicating that additional variables are needed or different calculations of the existing variables. In this specific case where the symbolic form of trade dependency was searched, all three secondary equations depicted a deviation from the conventional form of the trade dependency relationship and even tried to propose new formats.

6.2.3 Discussion

We investigated whether the GTCA can characterise income convergence according to production growth variable and determine the relationship between export as a percentage of GDP and domestic demand growth for trade dependence in the European economies. Acquiring such knowledge, for both issues, is significant since policy makers can learn from past policies and improve or alter them.

Observing each component separately, the SOM is successful in interpreting the different types of convergence or divergence of per capita income levels across countries or regions and identifies the sectors that influence these trends most for each EU country. Comparing these results to the ones obtained for Chinese provinces in Chapter 3, it is safe to say that the first module of the GTCA has given consistent results with topographic and qualitative errors either approaching or equal to zero.

The second component of the GTCA not only forms the symbolic relationship between the variables but also seem to be improving the relationships indicating that additional variables are needed or different calculations of the existing variables. By employing genetic programming and obtaining secondary relationships there is a unique opportunity to evaluate the format and accuracy of mathematical expressions of variables for established equations. Assessing this point in our application in chapter 5 we can conclude that the GP confirmed the gravity equation main structure for all partial fitness cases whereas in the validation analysed in the previous section the GP gave different formats for the trade dependency equation observed an inconsistent relationship between the variables.

6.3 GTCA Assessment

In this section we provide an assessment of the work presented in this thesis, in terms of the design and implementation of an intelligent systems model overall performance and its individual stages. Our main aim in this thesis was to introduce a general purpose methodology for developing intelligent country modelling tools for understanding a country's economic potential.

Our hypothesis was to try and achieve a better understanding of structurally different economies through the design, analysis and experimentation of an intelligent systems model testing and hopefully upgrading, in parallel, the proposed intelligent systems abilities and application domain. Following the development and testing of this model we now assess its individual components and discuss the overall approach to intelligent modelling of economic policies. We use the design requirements for both economic and computer assessment from section 1.3.3.

6.3.1 Representation

The increasing objective to develop and refine intelligent techniques to produce more accurate and effective representations of real-world problems has emerged from the belief that these systems perform in an unstable manner when faced with deviations of a dynamic environment. This was a deciding factor for applying intelligent systems to investigate and provide an understanding of China's macroeconomic policies and environment, since it has the most multidimensional, diverse yet integrated economic profile from any other single country. In order for an intelligent system to provide an accurate representation for the specified problem, its features should complement the problem structure and while translating unknown knowledge preserve certain critical properties of the input structure. Assessing the first part of the GTCA we observed that this criterion is met since the SOM not only provides direct visualisation of the numerous Chinese provinces but also goes beyond data ranking used by conventional statistical tools to characterise different types of structures. In the second part of the GTCA, genetic programming was also very successful since it found the dominant features determining China's trade equation and even suggested potential alternative versions of the specified equation by highlighting the influence each variable has on the equation.

6.3.2 Accuracy

Accuracy is determined by the capacity of exploring the data set for an accurate representation and the detection of variables that ones reaction affects that of another. GPs evolved hierarchies of subroutines in order to improve their understanding of the interactions of the variables describing trade actions in the Chinese environment. This feature helped in identifying the relationship that controls these actions but also to decode interim and partial solutions with high fitness thus eliminating information asymmetries caused by this equal weighing of the data used by the conventional form of the gravity equation. The SOM mappings clustered this interaction of the variables locating long-term growth indicators with similar patterns closer together whereas dissimilar ones further apart. In addition, while performing our experiments we realised an interesting process of the SOMs understanding of the interaction of features by grading their relationship along the horizontal and vertical axis of the map.

6.3.3 Reliability

This criterion assesses the procedure of being successful in finding future values and the reliability of these values. In this assessment the results from the intelligent growth type mappings were very satisfactory with map topographic and quantisation errors – discussed in detail in sections 2.3.1 and 3.2.1 – approaching zero and thus giving reliable results in both experimental (Chinese provinces) and validation (EU countries) data. The GP was not as successful in finding future reliable values but it did provide a unique analysis of the data sets by considering and describing variable properties for which it had only the data set and no guidance or information.

6.3.4 Applicability

This criterion relates to the range of conditions under which the system can operate successfully. In Chapter 3 and 5 the first and second part of the GTCA respectively, have been tested on provincial Chinese growth data and trade flow data sets. The GTCA's applicability has also been tested on different sets of data describing income convergence and trade dependence. The systems model was successful for both of these problems. However it would be interesting to test it on another transitional economy, for instance, an economy in Africa or one of the new member countries from Eastern Europe. Another interesting aspect, which is also discussed analytically in the future work section of Chapter 7, would be to apply it to two transitional economies with distinct features converging to the same pattern.

6.3.5 Emergent Properties

This assessment criterion acknowledges that successful research results do not always imply expected outcomes, but rather unexpected features that complement these standard solutions. These unexpected features are usually the incentive for proposal of future work to better understand them, thus increasing the potential of translating unknown knowledge patterns. GPs discovery of alternative versions of the gravity equation by evolution of its descriptive data variables indicated such emergent properties. Exploitation of this potential in GPs understanding of the data seems to be advantageous not just for the application domains used in this thesis but also in areas including medicine and determination of chemical structures for pharmaceuticals. The process of trying to correct or alternatively combine variables relationships as well as stressing the validity or limitations of the specific variables influencing an equation through high or low fitness values is a unique emergent property.

6.4 Research Contributions Revisited

Now that the GTCA has been designed, implemented and evaluated, we can review our research contributions initially introduced in Chapter 1.

The first contribution included the design and evaluation of the GTCA as a whole. The study was successful in discovering previously unknown relationships between Chinese provinces to assist in future policy design. Technical implications including noisy asymmetric trade flows and multivariate provincial growth range data were solved by employing genetic programming in combination with data reconciliation methods, interval coding and formation of trade rules with logic gates. The GTCA components process is automated while the tool does not yet provide an automated process. This is due to the fact that the GTCA is believed to be the first intelligent systems model designed for policy making and so we had to design it from scratch which was very time-consuming as was the collection and adjustments of Chinese data sets and making sure we are using the correct component for each part of the model. In the future work section in 7.2 we suggest ways of automating this process.

The second contribution is the success of the design and implementation of the first intelligent systems model GTCA employing two widely established techniques in providing policy makers with an initial understanding of the effects of country-specific growth and trade past decisions. The importance of such a system is substantial since it evaluates the potential and significance of using intelligent modelling to assist in good decision making on crucial country-specific structural issues. GTCA was successfully in identifying growth types among the numerous Chinese provinces through observation of the entire reform period and exploring the relationship of trade actions in the Chinese trading environment.

The SOM algorithm was successful in characterising inputs it has never seen before and even dealing with missing input data. It classified provincial data observations depicting the underlying policies promoting province-specific growth with relative indicator importance mappings to complement the main map. By observing part performance and focusing on the factors that influence each province in a positive or negative manner policy experts can have an immediate view of a country's internal growth dynamics and thus experiment on alternative policies with this tool.

The second part of the GTCA was also successful since GP did manage to form rules to mimic trading actions for the Chinese trading environment. The use of GP to search for the mathematical relationship shaping a dynamic evolving environment was catalytic. Although our aim was to predict values for trade for 2002, new interesting features of evolving computer programs that

seemed considerably more interesting than a prediction process. The second part of the GTCA does not only form the main relationship for Chinese trade it goes a step forward to improve the existing relationship by reassessing all possible combinations for the variables and even propose alterations to the main established symbolic expression.

The final contribution is the creation of an electronic database of the Chinese provinces for the entire reform period including agriculture, industry, construction, energy and GDP figures.

6.5 Summary

In this chapter we validated and assessed the components forming the basic structure of an intelligent country analyser encapsulating direct visualisation and location of complex unstructured correlations with the search for mathematical relationships determining evolving environments. The first past of the GTCA was used to investigate the levels of income convergence among the 15 original EU member countries. The system formed three main convergence types for the countries with the UK and Germany forming the core of convergence in the EU and the other members forming graded relationships around that core according to production sector dependence. In the second part of the GTCA, genetic programming is employed to find that the relationship determining trade dependence includes a more complex structure than the symbolic expression between exports as a percentage of GDP and domestic demand growth. In the final section we assess the GTCA as a whole as well as the individual components performance and applicability according to a set of criteria. The research contributions are then revisited in order to assess our findings.

Chapter 7

Conclusions and Future Work

The research leads to significant findings for the effectiveness and importance of the model designed and applied, and for the future prospective of its application spectrum. Intelligent country modelling research area was proposed to solve complex real-world macroeconomic structural issues. An intelligent growth and trade country analyser was constructed believed to be the first ever model to analyse and model policies. In chapters 2 and 4 the theoretical basis for the design of the components of this model were thoroughly surveyed. Chapters 3 and 5 provided the experimental work of the thesis which tested individual design stages of the GTCA and applied it to China's provincial national growth and international trade structure. This framework creates prospects for future study in various areas proposed and outlined in the final section of this chapter.

7.1 Conclusions

The main aim of this research was to design and develop the first intelligent systems model, named the Growth and Trade Country Analyser (GTCA), which provides policy makers with an initial understanding of the effects of country-specific growth and trade strategies thus assisting in good policy design on crucial country-specific structural issues. The high level of complexity of the application gave rise to significant design achievements for crucial real-world applications. These included characterisation of the numerous Chinese provinces according to growth type despite missing and inconsistent data as well as different definitions for indicators, the symbolic relationship of China's trade environment and the potential of improving the symbolic expression of established trade equations. The overall research achievement is the use of two widely established intelligent systems, namely Self-Organising Maps (SOMs) and Genetic Programming (GP) for the development of an intelligent country analyser.

The first part of the GTCA used SOM-based design in providing direct visualisation and location of the numerous Chinese provinces according to their distinct sector-specific growth identities, assessing the strategic importance of using the specific intelligent system. The SOM dealt with inaccurate and missing elements and formed provincial growth patterns while examining the effect of the growth process of specific indicators to each province's specific growth structure. We discovered that the SOM's ability to successfully investigate the impact macroeconomic policies had on each of the Chinese provinces over the reform period might also be due to its formation of horizontal and vertical scaling reflecting different levels of transition, poverty and inequality.

In the second part of the GTCA we designed trade rules to mimic trading actions for the Chinese trading environment using genetic programming to determine the symbolic relationship expressing an evolving environment. By evolving computer programs of our problem the GP showed interesting distinct features resembling the workings of Koza's GP for Kepler's 3rd law and partial relationships. These step features included the improvement of existing relationships, the proposal of alternative formats and extended versions of established symbolic expressions.

7.2 Future Work

Future work in this area is limitless, since intelligent country modelling has just been introduced in this thesis. Some proposed future work includes:

- The creation of a link between the two independent components to achieve a fully automated intelligent country analyser.
- The further investigation of the GP features for improving the format of the expression of variables in an equation.
- The application of the trading rules created for the GP to neuro-fuzzy systems to compare their results with GP.
- The investigation of the relevance – if it exists - of schemata formations for these GP features.
- The development of individual fitness functions for each variable in the mathematical relationship linked together with the fitness function for the equation similar to island-injection genetic algorithms (iiGA) [EbAPG99].
- Introducing other established techniques to the analyser or building new ones to complement them.

- The use of many other indicators for an analytical view of China. The application of growth and trade information for transitional economies in Africa and Eastern Europe to observe the effect and evolution of policy design.

References

[AN7898] *Anhui Statistical Yearbook*, (1978 -1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[AlKar99] Allen F., Karjalainen R., (1999), *Using Genetic Algorithms to find Technical Trading Rules*, Journal of Financial Economics, p.245 -71.

[Ander73] Anderberg, M.R., (1973), *Cluster Analysis for Applications*, Academic Press.

[Ander79] Anderson, J.E. (1979), *A Theoretical Foundation for the Gravity Equation*, American Economic Review, p.106 -116.

[Andre94] Andre, D., (1994), *Automatically Defined Features: The Simultaneous Evolution of 2D Feature Detectors and an Algorithm for using them,* Advances in Genetic Programming, Ed. Kinnear Jnr, p.477-494, The MIT Press.

[BJ7898] *Beijing Statistical Yearbook*, (1978 -1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[Banz93] Banzhaf, W., (1993), *Genetic Programming for Pedestrians*, Proceedings of the 5th International Conference on Genetic Algorithms ICGA'93, p.628-649, Morgan Kaufmann.

[BarDu01] Barrell, R., Dury, K., (2001), *Asymmetric Labour Markets in a Converging Europe: Do Differences matter?,* Working Paper No.2, European Network of Economic Policy Research Institutes (ENEPRI).

[BarrMa92] Barro, R.J., Sala-Martin, X., (1992), *Convergence*, Journal of Political Economy, p.223-251.

[Bau94] Bauer, R.J.Jnr, (1994), *Genetic Algorithms and Investment Strategies*, Wiley.

[Berg85] Bergstrand, J.H., (1985), *The Gravity Equation in International Trade: Some Microeconomic Foundations and Empirical Evidence*, Review of Economics and Statistics, p.474 - 481.

[Berg89] Bergstrand, J.H. (1989), *Trade Generalised Gravity Equation, Monopolistic Competition, and Factor-proportions Theory in International Trade*, Review of Economics and Statistics, p.143 - 153.

[Blic96] Blickle, T., (1996), *Evolving Compact Solution in Genetic Programming: A Case Study*, Parallel Problem Solving From Nature IV - Proceedings of the International Conference on Evolutionary Computation, Eds. Voigt, Ebeling, Rechenberg, and Schwefel, vol.1141, p.564-573, Springer-Verlag.

[BoHSS94] Bolsover, S.R., Hyams, J.S., Jones, S., Shephard, E.A., White, H.A., (1994), *From Genes to Cells*, Wiley.

[ChLeTr00] Chidambaran, N., Lee, C., Trigueros, J., (2000), *Option Pricing via Genetic Programming*, Computational Finance - Proceedings of 6th International Conference, Eds. Abu-Mostafa, LeBaron, Lo, and Weigend, The MIT Press.

[Chat89] Chatfield, C., (1989), *Analysis of Time Series: An Introduction*, 4th edition, Chapman and

Hall.

[Chen01] Chen, S., (2001), *On the Relevance of Genetic Programming to Evolutionary Economics*, Evolutionary Controversy in Economics towards a New Method in Preference of Trans Discipline, Eds. Anuka, *electronic version.*

[ChidTr00] Chidambaran N.L., Trigueros, J., (2000), *Option Pricing via Genetic Programming*, Computational Finance – Proceedings of the 6th International Conference, Eds. Abu-Mostafa, LeBaron, Lo, Weigend, The MIT Press.

[Chom96] Chomsky, N., (1996), *Powers and Prospects: Reflections on Human Nature and the Social Order*, Pluto Press.

[Chow94] Chow, G.C., (1994), *Understanding China's Economy*, World Scientific, Singapore.

[DaBRV96] Daida J.M., Bersano-Begey T.F., Ross S.J., Vesecky J.F.,(1996), *Computer-assisted Design of Image Classification Algorithms: Dynamic and Static Fitness Evaluations in Scaffolded Genetic Programming Environment*, Proceedings of the 1st Annual Conference, Eds. Koza, Goldberg, Fogel, Riolo, pp.279-284, The MIT Press.

[Darw59] Darwin, C., (1859), *On the Origin of Species*, John Murray.

[deFue00] de la Fuente, A., (2000), *Convergence across Countries and Regions : Theory and Empirics,* Instituto de Análysis Económico (CSIC), Barcelona.

[Deb00] Deboeck, G.J., (2000), *Trading on the Edge: Neural, Genetic, and Fuzzy Systems for Chaotic Financial Markets*, Wiley.

[DebKo98] Deboeck, G.J., Kohonen, T., (1998), *Visual Explorations in Finance using Self-Organising Maps*, Eds. Deboeck and Kohonen, Springer - Verlag.

[DoTS] *Direction of Trade Statistics*, (various years), International Monetary Fund (IMF), Washington D.C.

[Droz01] Drozdek, A., (2001), *Data Structures and Algorithms in Java*, Eds. Brooks and Cole.

[EbAPG99] Eby, D., Averill, R.C., Punch II, W.F., Goodman, E.D., (1999), *The Optimisation of Flywheels using the Injection Island Genetic Algorithm*, Evolutionary Design by Computers, Ed. Bentley, p167-190, Morgan Kaufmann.

[FJ7898] *Fujian Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[FZZ02] Fan, S., Zhang, L., Zhang X., (2002), *Growth, Inequality and Poverty in China: The Role of Public Investments*, Research Report 125, The International Food Policy Research Institute.

[Fer02] Ferreira, C., (2002), *Function Finding and the Creation of Numerical Constants in Gene Expression Programming*, 7th Online World Conference in Soft Computing in Industrial Applications.

[FoOW66] Fogel, L.J., Owen A.J., Walsh M.J., (1966), *Artificial Intelligence through Simulated Evolution*, Wiley.

[FuLa98] Fung, K.C., Lau L., (1998), *The China-US Bilateral Trade Balance: How big is it really?* Pacific Economic Review, p.33 - 47.

[FuLa03] Fung, K.C., Lau L., (2003), *Adjusted Estimates of US-China Bilateral Trade Balances: 1995-2002*, Japan Institute Report.

[GA7898] *Gansu Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[GD7898] *Guangdong Statistical Yearbook* (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[GU7898] *Guizhou Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[GX7898] *Guangxi Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[GathRo96] Gathercole, C., Ross, P., (1996), *An Adverse Interaction between Crossover and Restricted Tree Depth in Genetic Programming*, Proceedings of the 1st Annual Conference, p.291-296, The MIT Press.

[Gold89] Goldberg, D., (1989), *Genetic Algorithms In Search, Optimisation and Machine Learning.* Addison - Wesley.

[Good97] Goodman, D.S.G., (1997) *China's Provinces in Reform: Class, Community, and Political Culture*, Routledge.

[GooSe94] Goodman, D.S.G., Segal, G., (1994), *China Deconstructs: Politics, Trade and Regionalism*, Routledge.

[Guj03] Gujarati, D.N., (2003), *Basic Econometrics*, 4th edition, Mc Graw Hill.

[HA7898] *Hainan Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[HB7898] *Hubei Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[HN7898] *Henan Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[HU7898] *Hunan Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[HaBM94] Hampo, R.J., Bryant B.D., Marko K.A., (1994), *IC Engine Misfire Detection Algorithm Generation using Genetic Programming*, EUFIT'94, p.1674-1678.

[Handl93] Handley, S., (1993), *Automatic Learning of a Detector for Alpha-helices in Protein Sequences via Genetic Programming*, Proceedings of the 5th International Conference on Genetic Algorithms, ICGA'93, p.271-278, Morgan Kaufmann.

[HasSt89] Hastie, T., Stuetzle, W., (1989), *Principal Curves*, Journal of the American Statistical Association, p.502-516.

[Hebb49] Hebb, D.O., (1949), *The Organisation of Behaviour: A Neuropsychological Theory*, Wiley.

[Hend99] Hendrischke, H., (1999), *The Political Economy of China's provinces: Comparative and Competitive Advantage*, Eds. Hendrischke and Chongyi, Routledge.

[Hill97] Hiller, B., (1997), *The Macroeconomic Debate: Models of the Closed and Open Economy*, 3rd edition, Blackwells.

[HoeKo95] Hoekman, B., Kosteki, M., (1995), *The Political Economy of the World Trading System: From GATT to WTO*, Oxford University Press.

[Holl73] Holland, J.H., (1973), *Genetic Algorithms and the Optimal Allocation of Trials*, SIAM Journal of Computation, p.88-105.

[Holl75] Holland, J.H., (1975), *Adaptation in Natural and Artificial Systems*, University of Michigan Press.

[Hotel33] Hotelling, H., (1933), *Analysis of a Complex of Statistical Variables into Principal Components*, Journal of Educational Psychology, p.498-520.

[HuBr98] Huang, C., Broadbent, S., (1998), *Trade with China: Do the Figures add up?*, International Review of Applied Economics, vol.12, no.1, p.107-127.

[Hug03] Hughes, J., (2003), *US Trade Gap with China at New High*, Financial Times, International News section, p7.

[IM7898] *Inner Mongolia Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[IaMar01] Ianchovichina, E., Martin, W., (2001), *Trade Liberalisation in China's accession to the World Trade Organisation*, World Bank Report.

[JL7898] *Jilin Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[JS7898] *Jiangsu Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[JX7898] *Jiangxi Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[JaiDu88] Jain, A.K., Dubes, R.C., (1988), *Algorithms for Clustering Data*, Prentice Hall.

[JarSi71] Jardine, N., Sibson, R., (1971), *Mathematical Taxonomy*, London.

[KaVK00] Kaski, S., Venna, J., Kohonen, T., (2000), *Colouring that Reveals Cluster Structures in Multivariate Data*, Australian Journal of Intelligent Information Processing Systems, p.82-88.

[Kalb01] Kalbasi, H., (2001), *The Gravity Model and Global Trade Flows*, Policy Modelling for European and Global Issues Conference, Brussels.

[Kaski97] Kaski, S., (1997), *The Self-Organising Map*, PhD Thesis, Helsinki University Technology.

[KasKo95] Kaski,S., Kohonen, T., (1995), *Structures of Welfare and Poverty in the World Discovered by the Self-Organising Map,* Technical Report A24, Helsinki University of Technology, Finland.

[Keb99] Keber, C., (1999), *Genetically Derived Approximations for Determining the Implied Volatility*, OR Spektrum 21, p.205-238.

[Keb00] Keber, C., (2000), *Option Valuation with the Genetic Programming Approach*, Computational Finance - Proceedings of the 6th International Conference, Abu-Mostafa, LeBaron, Lo, Weigend, The MIT Press.

[Kinn94] Kinnear, K.E., (1994), *Advances in Genetic Programming*, The MIT Press.

[KinJ97] Kingdon, J., (1997), *Intelligent Systems and Financial Forecasting,* Perspectives in Neural Computing, Springer-Verlag.

[Kivi96] Kiviluoto, K., (1996) *Topology Preservation in Self-Organising Maps*, Proceedings of IEEE International Conference on Neural Networks, p.294-299.

[Klein01] Klein, N., (2001), *No Logo*, Flamingo.

[KohHa99] Kohonen, T., Hari, R., (1999), *Where the Abstract Feature Maps of the Brain might come from*, Trends in Neurosciences, p.135-139.

[KoOS96] Kohonen, T., Oja, E., Simula, O., Visa, A., Kangas, J., (1996), *Engineering Applications of the Self-Organising Map*, Proceedings of the IEEE, p.1358 - 1384.

[Koh82] Kohonen, T., (1982), *Self-Organised Formation of Topologically Correct Feature Maps*, Biological Cybernetics, p.59-69.

[Koh89] Kohonen, T., (1989), Self-Organisation and Associative Memory, 3rd edition, Springer-Verlag.

[Koza89] Koza, J.R., (1989), *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, The MIT Press.

[Koza91] Koza, J.R., (1991), *Concept Formation and Decision Tree Induction using the Genetic Programming Paradigm,* Parallel Problem Solving from Nature, Eds. Schwefel and Maenner, Springer-Verlag.

[Koza92a] Koza J.R., (1992), *Genetic Programming: On the Programming of Computers by Means of Natural Selection,* The MIT Press.

[Koza92b] Koza J.R., (1992), *A Genetic Approach to Econometric Modelling*, Economics and Cognitive Science, Eds. Bourgine and Walliser, p.57-75, Pergamon Press.

[KzABK96] Koza, J.R., Andre D., Bennett III, F.H., Keane, M.A., (1996), *Use of Automatically Defined Functions and Architecture-altering Operations in Automated Circuit Synthesis using Genetic Programming*, Proceedings of the 1st Annual Conference, p.132-149, The MIT Press.

[Kru99] Krugman, P., (1999), *The Role of Geography in Development*, Annual Wold Bank

Conference on Development Economics, The World Bank, p.89-107.

[KruWi78] Kruskal, J.B., Wish, M., (1978), *Multidimensional Scaling*, Sage University paper series on Qualitative Applications in the Social Sciences, Sage Publications.

[LHKK00] Lagus, K., Honkela, T., Kaski, S., Kohonen, T., (2000), *WEBSOM for Textual Data Mining*, Artificial Intelligent Review, p.345-364.

[LN7898] *Liaoning Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[LaPol02] Langdon, W.B., Poli, R., (2000), *Foundations of Genetic Programming*, Springer.

[Lan98] Langdon, W.B., (1998), *Genetic Programming and Data Structures*, Kluwer Academic Publishers.

[Lan00] Langdon, W.B., (2000), *Size Fair and Homologous Tree Crossover for Tree Genetic Programming*, Genetic Programming and Evolvable Machines, Ed. Banzhaf, p.95-120.

[Lar94] Lardy, N., (1994), *China in the World Economy*, Institute for International Economics, Washington, D.C.

[Leen96] Leendert, A., (1996), *Algorithms and Data Structures in C++*, Wiley.

[Levy92] Levy, S., (1992), *Artificial Life: A Report from the Frontier where Computers Meet Biology*, Vintage.

[Linn66] Linnemann, H., (1966), *An Economic Study of International Trade Flows,* North–Holland Publishing Company.

[LloZh00] Lloyd, P.J., Zhang, X., (2000), *China in the Global Economy*, Edward Elgar Publications.

[LugKo03] Luginbuhl, R., Koopman, S.J., (2003), *Convergence in European GDP Series*, Tinbergen Institute Discussion Paper, Vrije Universiteit Amsterdam.

[MaSev96] Matyas, L., Sevestre, P., (1996), *The Econometrics of Panel Data: A Handbook of the Theory with Applications*, Ed. Matyas, 2nd edition.

[Madd01] Maddison, A.,(2001),*The World Economy: A Millennial Perspective*, OECD Development Centre Studies.

[Mank97] Mankiw, N.G., (1997), *Macroeconomics*, 3rdedition, Worth.

[McCulP43] McCulloch, W., Pitts, W., (1943), *A Logical Calculus of the Ideas Immanent in Nervous Activity*, Bulletin of Mathematical Biophysics, p.115-137.

[McQue67] McQueen, J., (1967), *Some Methods for Classification and Analysis of Multivariate Observations*, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability Volume I: Statistics, p281-297, University of California Press.

[MiCM86] Michalski, R.S., Carbonell, J.G., Mitchell, T.M., (1986), *Machine Learning: An Artificial Intelligence Approach*, Morgan Kaufman.

[MiJV00] Miller, J.F., Job, D., Vassilev, V.K., (2000), *Principles in the Evolutionary Design of Digital Circuits Part I*, Genetic Programming and Evolvable Machines, Ed. Banzhaf, p.259-288.

[MitchM96] Mitchell, M., (1996), *An Introduction to Genetic Algorithms*, The MIT Press.

[MitchT97] Mitchell, T. (1997), *Machine Learning*, McGraw Hill International Editions.

[MulCh95] Mulier, F., Cherkassky, V., (1995), *Self-Organisation as an iterative kernel smoothing process*, Neural Computation, p.1165-1177.

[NX7898] *Ningxia Statistical Yearbook,* (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[Nach00] Nachbar, R.B., (2000), *Molecular Evolution: Automated Manipulation of Hierarchical Chemical Topology and its Application to Average Molecular Structures*, Genetic Programming and Evolvable Machines, Ed. Banzhaf, p.57-94.

[Nas99] Nasution, A., (1999), *Recent Issues in the Management of Macroeconomic Policies: the PRC, in Rising to the Challenge in Asia: A Study of Financial Markets*, Asian Development Bank, p.2-31.

[Neel97] Neely C., Weller P., Ditmar R., (1997), *Is Technical Analysis in the Foreign Exchange Market Profitable? A Genetic Programming Approach*, Journal of Economic Dynamics and Control, p.329-373.

[OReil95] O'Reilly, U., (1995), *An Analysis of Genetic Programming*, PhD Thesis, Carleton University, Ottawa-Carleton Institute of Computer Science, Canada.

[Paton94] Paton, R. (1994), *Computing with Biological Metaphors*, Ed. Paton, Chapman and Hall.

[Poli01] Poli, R., (2001), *Exact Schema Theory for Genetic Programming and variable-length Genetic Algorithm with One-point Crossover, Genetic Programming and Evolvable Machines*, Ed. Banzhaf, p.123-164.

[Price70] Price, G.R., (1970), *Selection and Covariance*, Nature, vol.227, p.520-521.

[Rech73] Rechenberg, I., (1973), *Evolutionsstrategie: Optimierung Technisher Systeme nach Prinzipien der Biologischen Evolution*, Frommann-Holzboog.

[RittMa92] Ritter, H., Martinetz, T., Schulten, K., (1992), *Neural Computation and Self-Organising Maps: An Introduction*, Addison-Wesley.

[Rosen62] Rosenblatt, F., (1962), *Principles of Neurodynamics*, Spartan Books.

[Rusk03] Ruskin, A., (2003), *A truer measure of China's trade surplus*, Financial Times.

[SD7898] *Shandong Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[SAX7898] *Shaanxi Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[SX7898] *Shanxi Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[Sam69] Sammon, J.W., (1969), *A Nonlinear Mapping for Data Structure Analysis*, IEEE Transactions on Computers, p.401-409.

[SneaSo73] Sneath, P.H.A., Sokal R.R., (1973), *Numerical Taxonomy*, Freeman.

[SerMa93] Serrano-Cinca, C., Martin-del-Brio, B., (1993), *Self-Organising Neural Networks of the Analysis and Representation of Data: Some Financial Cases*, Neural Computing and Applications, p.193-206, Springer-Verlag.

[SAHV96] Simula, O., Alhoniemi., E., Hollmén, J., Vesanto, J., (1996), *Monitoring and Modelling Complex Processes using Hierarchical Self-organising Maps*, Proceedings of the IEEE International Symposium on Circuits and Systems, p.73-76.

[SiVV99] Simula, O., Vasara, P., Vesanto, J., Helminen, R., (1999), *The Self-Organising Map in Industry Analysis,* Intelligent Techniques in Industry, Eds. Jain and Vemuri, p.87-112, CRC Press LLC.

[Smith98] Smith, J.M., (1998), *Evolutionary Genetics*, Oxford University Press, 2nd edition.

[SoSu02] Sonntag, B., Sun, H., (2002), *Agriculture Development and Environment in Critical Areas of China,* Science Press Beijing.

[StatCHN] *Statistical Yearbook of China*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[StatEN] *Energy Statistical Yearbook of China* (1978-1998), China Statistical Bureau, Beijing.

[StiHS00] Stiglitz, J., Hussain, A., Stern, N., (2000), *Chinese Reforms from a Comparative Perspective*, Incentives, Organisation, and Public Economics: Papers in honour of Sir James Mirrlees, Eds. Hammond and Myles, p.243-276, Oxford University Press.

[Stig02] Stiglitz, J.,(2002),*Globalisation and its Discontents*, Ed. Allen Lane, The Penguin Press.

[StuVar00] Stuphens, C.R., Vargas, J.M., (2000), *Effective Fitness as an Alternative Paradigm for Evolutionary Computation I*, Genetic Programming and Evolvable Machines, Ed. Banzhaf, p.363-378.

[StuVar01] Stuphens, C.R., Vargas, J.M., (2001), *Effective Fitness as an Alternative Paradigm for Evolutionary Computation II: Examples and Applications*, Genetic Programming and Evolvable Machines, Ed. Banzhaf, p.7-32.

[Sung91] Sung, Y., (1991), *The China-Hong Kong Connection: The key to China's Open Door Policy*, Cambridge University Press.

[TJ7898] Tianjin Statistical Yearbook, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[Todar99] Todaro, M.P., (1999), *Economic Development*, 7th edition, Addison Wesley Longman.

[Tyb92] Tybout, J.(1992), *Making Noisy Data Sing: Estimating Production Technologies in Developing Countries*, Journal of Econometrics, p.25-44.

[Ults93] Ultsch, A., (1993), *Self-Organising Feature Maps for Monitoring and Knowledge Acquisition of a Chemical Process*, Proceedings of International Conference on Artificial Neural Networks (ICANN), p.864-867.

[UltGK93] Ultsch, A., Guimaraes, G., Korus, H., (1993), *Knowledge Extraction from Artificial Neural Networks and Applications*, Proceedings of Tranputer-Anwender-Treffen / World Transputer-Congress, p.194-203, Springer Verlag.

[Vapn95] Vapnik, V.N., (1995), *The Nature of Statistical Learning Theory*, Springer-Verlag.

[Ves97] Vesanto, J., (1997), *Data Mining Techniques based on the Self-Organising Map*, Master's thesis, Helsinki University of Technology (HUT), Finland.

[Ves00] Vesanto, J., (2000), *Neural Network Tool for Data Mining: SOM Toolbox*, Proceedings of Symposium on Tool Environments and Development Methods for Intelligent Systems, Finland, p.184-196.

[VeKa01] Venna, J., Kaski, S., (2001), *Neighbourhood Preservation in Nonlinear Projection Methods: An Experimental Study*, 11th International Conference on Artificial Neural Networks (ICANN01).

[Verb97] Verburg, P., (1997), *Exploring the Spatial and Temporal Dynamics of Land Use*, PhD Thesis, Wageningen University, The Netherlands.

[WolpMa97] Wolpert, D.H., Macready, W.G., (1997), *No Free Lunch Theorems for Optimisation*, IEEE Transactions of Evolutionary Computation.

[XJ7898] *Xinjiang Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[YN7898] *Yunnan Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[ZJ7898] *Zhejiang Statistical Yearbook*, (1978-1998), China Statistical Bureau, Statistical Publishing House, Beijing.

[Zhao96] Zhao, C., (1996), *China in the New World Trading Environment after the Uruguay Round*, Country Study Report, Asian Development Bank.

[Zhao97] Zhao, H., (1997), *Foreign Trade in the People's Republic of China: Past Performance and Future Challenges*, Asian Development Review, p.88-110.

Appendix A - Implementation

In Appendix A we introduce the software used for implementation of the two different stages of the Growth and Trade Country Analyser (GTCA). The first part of the GTCA used the Self-Organising Map (SOM) to classify and locate provinces according to their growth type. The SOM Toolbox created by Teuvo Kohonen's team at Helsinki University of Technology (HUT) was used as the basis for all our SOM workings. The second part of the GTCA employed Genetic Programming (GP) to identify China's bilateral trade environment. SIMPLE GP created by Bill Langdon at University College London was used as the basis for GP. Section A.1 introduces the features of the SOM Toolbox package and section A.2 introduces SIMPLE GP features.

A.1 The SOM Toolbox

The SOM Toolbox is a public domain program package created by Teuvo Kohonen's research group, namely Juha Vesanto, Johan Himberg, Elsa Alhoniemi and Juha Parhankangas at the in Helsinki University of Technology (HUT), for undertaking SOM applications using a high-level programming language computing environment, MATLAB. It requires no other toolboxes, just the basic MATLAB functions. The Toolbox was designed and built to assist researchers by providing a good user-friendly implementation of the SOM in MatLab. It contains powerful visualisation functions and is an excellent basis for data mining since the researchers that built it specialise in this area. For reasons of compatibility with the SOM Toolbox requirements our work for the SOM is primarily coded in C++. The Toolbox can used to pre-process data, initialise and train SOMs using different kinds of topologies, visualise SOMs in various ways and analyse the properties of the SOMs and data.

The SOM Toolbox takes full advantage of MATLAB's strong support for graphics and visualisation. Important SOM Toolbox features include:

- *Modular Programming Style:* The user can tailor the code to their specific needs since the toolbox utilises MatLab structures and functions are constructed in a modular manner.
- *Component Weights and Names:* The input vector components may be given different weights according to their relative importance, and the components can be given names to make figures easier to read.

- *Data Pre-processing Tools:* The data pre-processing tools included in the package are variance normalisation and histogram equalisation.
- *Batch or Sequential Training:* The speed of training can be immensely improved in data analysis applications by using the batch training version.
- *Map Dimensions:* Although visualisation can be problematic when the dimension is higher than 2, high-dimensional maps are made available in the toolbox.
- *Graphical User Interface:* Graphical user interface (GUI), guides through the initialisation and training procedures of the map and offers several different methods of data visualisation on the trained map.
- *Advanced Graphics:* The SOM toolbox builds on MATLAB's strong graphics capabilities and can be used to produce very appealing and extraordinary figures.

A.2 GP Implementation

For the GP implementation we use simple GP written by Bill Langdon at University College London as our basis and added features specific to our application. As in section A.1 we again code the functions in C++. We used simple GP for the following reasons:

- It is a public domain software
- Since it is written in C++ code, we made negligible changes to our functions created in earlier programs.
- It is stable.
- It has a flexible structure so we could easily add new features to it.

While using the software we made numerous changes, listed below:

- Extended selection of functions to include logic gates
- Changed the function that calculates fitness to count the number of individuals and then give a result
- Extended selection of terminals to include user-defined codings.

Appendix B – Data

In Appendix B we provide the data sets we used in this work. These include:

- Sector-specific long-term indicator data sets for the entire reform period (1978 to 1998) for each individual province. The province and indicator names are in also in Chinese. These data sets were used for the main SOM map (figure 3.8a), the relative importance chart (figure 3.8b) and the China geographical map representation (figure 3.9) and are presented in Appendix B.1; Data for each indicator and each province for individual years 1978, 1988 and 1998 as well as data for the analysis of individual and pair-wise indicator profiles in figures 3.11, 3.12 and 3.13 for 1998 are also attained from this data;
- In Appendix B.2 we provide data sets used for the Huang-Broadbent model to deal with export flow inconsistencies, these include partners data on imports from China and Hong Kong (HK), HK domestic exports to partner country, total HK re-exports to partner country, Chinese exports and HK re-exports to partner country from China;
- In Appendix B.3 we present data sets from which we constructed the GP trading rules. We will provide full data sets for the start (1997) and end (2001) years for GDP, GDP per capita, geographical distance and import flows for 50 partner countries;
- Validation data from 1997 to 2001 for GDP sectors, primary (agriculture), secondary (industry and manufacturing) and tertiary (services) and also for exports as % of GDP and domestic demand growth for each European Union member country are all presented in Appendix B.4.

Appendix B.1

This section of Appendix B presents data inputs for each individual province for indicators including agriculture, construction, energy, industry, population and GDP for years 1978 to 1998. GDP is given because it is a central measured that includes and is expressed by all the above indicators. We provide more data information for the earlier years since it is harder to find and also give the names of the provinces and indicators in Chinese to assist researchers looking for these particular indicators in the provincial Statistical Yearbooks.

Chinese notations	Indicators
国内生产总值	GDP
总人口	Population
工业总产值	Industry
建筑业总产值	Construction
能源生产总量	Energy
农林牧渔业	Agriculture

EASTERN PROVINCES:

BEIJING	1978	1980	1982	1984	1986	1988	1992	1994	1998
国内生产总值	104	148	178	217	285	410	685	1084	2011
总人口	872	904	935	966	1032	1080	1154	1208	1246
工业总产值	152	182	213	282	372	569	1086	1788	2017
建筑业总产值	105	147	173	334	513	816	1254	3015	6230
能源生产总量	832	674	773	902	993	870	1015	1000	1098
农林牧渔业	11	15	18	22	28	52	85	144	177

FUJIAN	1978	1980	1982	1984	1986	1988	1992	1994	1998
国内生产总值	66	88	118	158	224	386	798	1706	3330
总人口	2452	2517	2604	2676	2749	2845	3066	3126	3299
工业总产值	64	81	97	132	191	313	911	1973	4509
建筑业总产值	65	72	85	102	148	196	373	1305	2446
能源生产总量	604	691	757	799	784	924	910	1000	1349
农林牧渔业	36	45	64	81	107	182	301	591	973

GUANGDONG	1978	1980	1982	1984	1986	1988	1992	1994	1998
国内生产总值	185	246	331	442	638	1099	2294	4241	7919
总人口	5064	5227	5396	5532	5729	5928	6463	6691	7143
工业总产值	200	236	295	396	585	1051	3353	6220	13575
建筑业总产值	123	153	218	431	946	1125	2165	4890	7886
能源生产总量	955	908	928	951	940	1028	1240	1384	3022
农林牧渔业	74	82	127	142	162	191	664	720	1615

HAINAN	1978	1980	1982	1984	1986	1988	1992	1994	1998
国内生产总值	18	22	26	35	48	77	182	331	439
总人口	528	552	571	589	605	627	671	691	753
工业总产值	76	69	75	111	155	241	722	1175	255
建筑业总产值	20	23	31	43	58	83	102	222	308
能源生产总量	0	0	0	0	0	0	2	1	4
农林牧渔业	96	102	175	218	246	267	796	1010	2425

JIANGSU	1978	1980	1982	1984	1986	1988	1992	1994	1998
国内生产总值	568	751	898	1239	1960	3320	6863	14164	7199
总人口	5834	5938	6083	6150	6270	6438	6911	7021	7182
工业总产值	337	468	535	745	1235	2153	4674	9826	13186
建筑业总产值	84	138	182	243	303	396	1142	3609	11962
能源生产总量	1428	1765	1905	1989	2065	2442	2549	2603	2861
农林牧渔业	106	139	188	254	332	498	674	1335	1849

LIAONING	1978	1980	1982	1984	1986	1988	1992	1994	1998
國内生产总值	228	281	312	429	578	839	1298	2584	3882
总人口	3394	3487	3592	3655	3726	3825	3957	4007	4157
工业总产值	397	466	490	609	791	1032	2180	3545	6674
建筑业总产值	142	162	193	348	553	996	2215	3852	4189
能源生产总量	4654	4436	5043	5382	5988	6402	6783	7002	7803
农林牧渔业	49	73	85	117	142	227	341	602	969

SHANGHAI	1978	1980	1982	1984	1986	1988	1992	1994	1998
國内生产总值	273	312	337	391	491	648	1114	1972	3688
总人口	1098	1146	1180	1205	1232	1262	1289	1373	1464
工业总产值	263	395	523	744	952	1296	2429	4289	5848
建筑业总产值	112	149	184	247	318	406	1176	3096	5859
能源生产总量	0	0	0	0	0	0	0	0	0
农林牧渔业	18	19	24	27	34	53	80	140	207

SHANDONG	1978	1980	1982	1984	1986	1988	1992	1994	1998
國内生产总值	225	292	395	582	742	1118	2197	3872	7162
总人口	7160	7296	7494	7637	7818	8061	8610	8671	8838
工业总产值	297	340	393	535	784	1455	3115	7023	10579
建筑业总产值	88	133	179	239	299	397	986	2064	7001
能源生产总量	1547	1710	1948	2054	2216	2551	3036	3757	5107
农林牧渔业	102	161	219	310	361	495	841	1387	2175

TIANJIN	1978	1980	1982	1984	1986	1988	1992	1994	1998
國内生产总值	83	104	114	148	195	260	411	725	1336
总人口	796	806	826	846	869	889	921	935	957
工业总产值	158	196	218	259	345	521	998	1838	2563
建筑业总产值	69	84	109	142	212	304	495	1239	2111
能源生产总量	309	299	338	368	379	477	492	589	757
农林牧渔业	7	9	13	17	27	44	62	95	156

ZHEJIANG	1978	1980	1982	1984	1986	1988	1992	1994	1998
國内生产总值	124	179	233	322	500	765	1365	2667	4987
总人口	3751	3826	3924	3993	4070	4169	4286	4315	4456
工业总产值	132	202	251	383	674	1141	2447	5784	11338
建筑业总产值	91	111	161	201	257	306	656	2213	9425
能源生产总量	219	221	149	172	192	216	144	100	121
农林牧渔业	65	93	118	147	192	282	405	707	1004

CENTRAL PROVINCES:

ANHUI	1978	1980	1982	1984	1986	1988	1992	1994	1998
国内生产总值	114	141	187	266	383	547	801	1489	2805
总人口	4693	4835	4987	5108	5242	5443	5834	5955	6184
工业总产值	107	130	153	185	324	518	999	2478	3852
建筑业总产值	96	103	116	148	232	301	554	1145	2356
能源生产总量	1836	1801	2098	2537	2921	2803	3376	4106	6003
农林牧渔业	81	98	127	162	222	315	390	774	1202

GUANGXI	1978	1980	1982	1984	1986	1988	1992	1994	1998
国内生产总值	76	97	129	150	205	313	647	1242	1903
总人口	3402	3538	3684	3806	3946	4088	4359	4493	4675
工业总产值	70	79	93	109	165	272	583	1381	1698
建筑业总产值	57	69	83	98	131	189	397	849	1344
能源生产总量	1068	845	913	902	998	1202	1096	1203	1307
农林牧渔业	47	57	74	94	119	169	333	537	866

HEILONGJIANG	1978	1980	1982	1984	1986	1988	1992	1994	1998
国内生产总值	175	221	253	303	397	492	964	1618	2833
总人口	3129	3203	3281	3331	3385	3466	3608	3672	3773
工业总产值	212	244	282	338	452	588	1103	1797	2688
建筑业总产值	118	152	197	251	319	399	1194	1802	2739
能源生产总量	8576	8484	9520	10199	11813	12701	13961	13301	14502
农林牧渔业	61	86	98	122	137	152	285	538	736

HENAN	1978	1980	1982	1984	1986	1988	1992	1994	1998
国内生产总值	163	229	263	370	503	749	1280	2224	4357
总人口	7067	7285	7519	7737	7985	8317	8861	9027	9315
工业总产值	171	209	246	308	478	780	1629	3437	5830
建筑业总产值	102	141	184	229	273	345	703	1455	3049
能源生产总量	5475	5542	6527	8073	8860	9751	9837	10288	13251
农林牧渔业	95	135	151	209	259	371	574	883	1823

HUBEI	1978	1980	1982	1984	1986	1988	1992	1994	1998
国内生产总值	151	199	242	328	442	627	1088	1879	3704
总人口	4574	4684	4796	4887	4989	5144	5513	5718	5907
工业总产值	167	228	276	359	538	835	1374	3025	6405
建筑业总产值	120	162	203	246	294	372	803	1693	3436
能源生产总量	826	646	772	1015	1069	1131	962	1287	1729
农林牧渔业	84	95	118	180	219	298	435	787	1148

HUNAN	1978	1980	1982	1984	1986	1988	1992	1994	1998
国内生产总值	147	192	233	287	398	584	998	1694	3211
总人口	5165	5280	5452	5561	5696	5915	6207	6302	6502
工业总产值	128	166	202	229	369	582	1007	1925	4074
建筑业总产值	91	110	113	145	237	305	560	1154	3273
能源生产总量	246	277	296	317	351	405	360	490	670
农林牧渔业	94	104	129	185	227	303	471	838	1232

I.MONGOLIA	1978	1980	1982	1984	1986	1988	1992	1994	1998
国内生产总值	46	68	74	128	182	271	422	682	1192
总人口	1823	1876	1941	1993	2040	2093	2207	2260	2345
工业总产值	219	221	238	274	329	373	504	605	1137
建筑业总产值	76	95	109	137	210	304	507	801	1017
能源生产总量	1335	2436	3015	3529	4011	4295	5139	6145	8295
农林牧渔业	33	40	59	76	95	156	262	563	534

JIANGXI	1978	1980	1982	1984	1986	1988	1992	1994	1998
国内生产总值	75	104	121	144	208	263	476	745	1852
总人口	31828	32701	33483	34578	35757	36838	39130	40154	41913
工业总产值	74	94	107	137	214	345	646	1571	1631
建筑业总产值	56	69	87	99	150	188	351	673	894
能源生产总量	1578	1518	1844	1893	2062	2200	2088	2300	3504
农林牧渔业	49	60	77	98	125	174	298	528	735

JILIN	1978	1980	1982	1984	1986	1988	1992	1994	1998
国内生产总值	82	99	122	174	227	369	558	969	1558
总人口	2149	2210	2257	2284	2315	2357	2474	2516	2644
工业总产值	113	135	149	209	282	454	768	1281	1708
建筑业总产值	98	121	172	216	269	313	741	1163	1460
能源生产总量	2264	2189	2471	2650	2762	2835	2848	2832	3226
农林牧渔业	38	47	61	89	98	141	204	405	666

SICHUAN	1978	1980	1982	1984	1986	1988	1992	1994	1998
国内生产总值	245	322	429	518	655	923	1625	2778	3580
总人口	9707	9819	9977	10112	10319	10589	10943	11084	11639
工业总产值	226	274	340	391	599	964	1850	3917	3689
建筑业总产值	97	149	179	339	516	821	1377	2884	5977
能源生产总量	4288	4448	5123	5537	6344	6853	7103	8817	10646
农林牧渔业	148	181	233	282	339	476	745	1229	1394

SHANXI	1978	1980	1982	1984	1986	1988	1992	1994	1998
國内生产总值	88	109	139	197	235	317	570	854	1601
总人口	2423	2476	2551	2632	2713	2807	2979	3045	3172
工业总产值	98	115	134	171	218	321	567	887	2318
建筑业总产值	92	118	167	211	261	311	665	1325	2178
能源生产总量	8035	9759	14898	19024	20963	23907	29687	32400	41137
农林牧渔业	29	38	45	65	59	87	131	219	359

WESTERN PROVINCES:

GANSU	1978	1980	1982	1984	1986	1988	1992	1994	1998
國内生产总值	65	74	77	103	141	192	318	452	870
总人口	1870	1918	1974	2015	2071	2137	2288	2352	2519
工业总产值	77	77	80	99	144	204	369	657	1081
建筑业总产值	49	65	78	107	145	189	325	515	1133
能源生产总量	1251	1257	1319	1256	1457	1599	1740	2257	2930
农林牧渔业	21	29	29	38	56	85	123	142	336

GUIZHOU	1978	1980	1982	1984	1986	1988	1992	1994	1998
國内生产总值	47	60	79	108	140	211	340	521	842
总人口	2779	2831	2904	2965	3042	3140	3340	3402	3658
工业总产值	42	45	54	80	95	126	300	382	796
建筑业总产值	43	45	53	75	112	152	264	486	832
能源生产总量	2072	1835	2223	2499	3000	3397	4162	5100	7169
农林牧渔业	27	28	48	55	60	63	170	183	402

NINGXIA	1978	1980	1982	1984	1986	1988	1992	1994	1998
國内生产总值	13	16	18	25	35	50	83	134	227
总人口	3555	3737	3907	4068	4243	4445	4822	5038	5382
工业总产值	138	139	147	191	282	440	894	1506	228
建筑业总产值	12	22	30	42	55	69	135	186	393
能源生产总量	1124	1032	1127	1160	1242	1346	1400	1412	1774
农林牧渔业	48	64	80	109	142	197	284	458	79

SHAANXI	1978	1980	1982	1984	1986	1988	1992	1994	1998
國内生产总值	294	338	389	508	690	1006	1605	2440	1382
总人口	2663	2754	2851	2916	2987	3093	3286	3356	3596
工业总产值	96	110	118	151	219	332	600	1010	1295
建筑业总产值	63	91	106	139	215	307	485	980	1879
能源生产总量	2062	2355	2748	3004	3237	3606	3524	3850	5180
农林牧渔业	36	42	57	74	87	131	205	302	479

YUNNAN	1978	1980	1982	1984	1986	1988	1992	1994	1998
國内生产总值	69	84	110	140	182	301	619	974	1794
总人口	3091	3173	3283	3372	3480	3594	3832	3939	4144
工业总产值	59	65	84	109	128	176	464	637	1503
建筑业总产值	54	68	88	102	132	191	383	843	2517
能源生产总量	1553	1352	1506	1632	1700	2109	2391	2615	3413
农林牧渔业	40	48	62	77	96	135	250	357	615

XINJIANG	1978	1980	1982	1984	1986	1988	1992	1994	1998
國内生产总值	39	53	65	90	129	193	402	674	1117
总人口	1233	1283	1314	1345	1384	1426	1581	1632	1747
工业总产值	25	34	42	60	98	149	317	638	708
建筑业总产值	59	72	87	134	178	234	490	739	1435
能源生产总量	5332	5059	5739	6808	7226	8274	9696	10890	12993
农林牧渔业	23	30	37	50	66	108	172	306	498

Appendix B.2

Data for imports from China, imports from Hong Kong (HK), HK domestic exports to partner country, total HK re-exports to partner country, Chinese exports and HK re-exports to partner country from China. We used these data sets in Chapter 5 to deal with export flow inconsistencies using the Huang-Broadbent model.

Raw Data JAPAN (US$)	$M_{CN,JP}$	$M_{HK,JP}$	$DX_{HK,JP}$	$RX_{HK,JP}$	$X_{CN,JP}$	$RX_{CN,JP}$
1997	46217	2485	1364	9964	31816	8780
1998	44231	2065	836	8336	29692	7346
1999	44632	1855	709	8767	32399	7560
2000	54391	1645	660	10656	41654	9388
2001	64326	1621	527	10851	45078	9731

Raw Data US (US$)	$M_{CN,US}$	$M_{HK,US}$	$DX_{HK,US}$	$RX_{HK,US}$	$X_{CN,US}$	$RX_{CN,US}$
1997	62600	10300	7147	33944	32702	31460
1998	71200	10500	6789	33748	37975	31080
1999	70000	10600	6670	34993	41945	32235
2000	100100	9400	7070	40396	52104	36855
2001	102300	9700	6180	36648	54318	33678

Appendix B.3

Data from 1997 to 2001 for GDP, GDP per capita, geographical distance and import flows for 75 countries used to construct the trading rules.

1997	Exports	GDP	GDPpc	DST	Imports
Argentina97	465	325	9	11957	721
Australia97	2054	394	21	5548	3245
Austria97	199	206	25	4646	256
Belgium97	1361	243	24	4959	916
Brazil97	1051	820	5	10720	1486
Canada97	1906	608	20	6510	2001
Chile97	562	77	5	11842	415
Denmark97	372	170	32	4488	347
Finland97	271	120	23	3939	687
France97	2331	139	2	5119	3239
Germany97	6491	209	3	4585	6184
HK97	43797	171	26	1254	6997
India97	934	382	0	2341	897
Indonesia97	1841	215	1	3227	2673
Iran97	497	90	1	3482	536
Ireland97	128	75	20	5154	67
Israel97	256	98	17	4441	99
Italy97	2239	114	2	5057	2449
Japan97	31816	419	3	1307	28988
Kazakhstan97	95	22	1	2037	433
Malaysia97	1921	98	5	2693	2485
Mexico97	415	403	4	7750	184
N.Zealand97	282	65	17	6681	347
Netherlands97	4406	360	23	4874	1072
Nigeria97	316	40	0	7138	11
Norway97	568	153	35	4377	379
Pakistan97	691	62	0	2416	379
Philippines97	1335	82	1	1764	327
Poland97	673	136	4	4308	32
Romania97	176	35	2	4396	73
Russia97	2033	447	3	3608	4084
S.Africa97	785	129	3	7255	789
Saudi Arab97	854	140	7	4109	825
Singapore97	4321	96	31	2769	4385
S.Korea97	9122	443	10	598	14884
Spain97	1245	532	14	5742	555
Sudan97	111	10	0	5216	23
Sweden97	527	228	26	4178	1297
Switzerland97	615	255	36	5112	873
Taiwan97	3397			1070	16433
Thailand97	1502	154	3	2039	2004
Turkey97	558	190	3	4392	64
UAE97	1301	39	15	3623	84
UK97	3815	128	2	5071	1977
US97	32705	783	3	6941	16288
Venezuela97	119	87	4	8955	32
Vietnam97	1078	25	0	1442	357
Yemen97	101	6	0	4676	651

2001	Exports	GDP	GDPpc	DST	Imports
Argentina01	574	269	7	11957	1281
Australia01	3573	369	19	5548	5431
Austria01	354	189	23	4646	662
Belgium01	2547	230	22	4959	1721
Brazil01	1363	503	3	10720	2347
Canada01	3349	694	22	6510	4029
Chile01	816	66	4	11842	1304
Denmark01	898	162	30	4488	626
Finland01	912	121	23	3939	2376
France01	3692	131	2	5119	4105
Germany01	9759	185	2	4585	13695
HK01	46503	162	24	1254	9424
India01	1903	477	1	2341	1701
Indonesia01	2847	145	1	3227	3888
Iran01	900	114	2	3482	2424
Ireland01	530	103	27	5154	613
Israel01	833	108	17	4441	483
Italy01	4005	109	2	5057	3784
Japan01	45078	414	3	1307	42811
Kazakhstan01	328	22	2	2037	961
Malaysia01	3223	88	4	2693	6206
Mexico01	1802	618	6	7750	761
N.Zealand01	435	50	13	6681	737
Netherlands01	7293	380	24	4874	1456
Nigeria01	919	41	0	7138	228
Norway01	412	166	37	4377	571
Pakistan01	820	59	0	2416	581
Philippines01	1622	71	1	1764	1945
Poland01	1017	176	5	4308	226
Romania01	255	39	2	4396	104
Russia01	2715	310	2	3608	7959
S.Africa01	1051	113	3	7255	1173
Saudi Arab01	1356	186	9	4109	2723
Singapore01	5795	86	21	2769	5143
S.Korea01	12544	422	9	598	23396
Spain01	2264	582	14	5742	714
Sudan01	227	13	0	5216	938
Sweden01	932	210	24	4178	2173
Switzerland01	652	247	34	5112	1731
Taiwan01	5006			1070	27344
Thailand01	2504	115	2	2039	4713
Turkey01	676	148	2	4392	231
UAE01	2381	46	15	3623	448
UK01	6784	142	2	5071	3525
US01	54319	106		6941	26204
Venezuela01	444	125	5	8955	146
Vietnam01	1805	33	0	1442	1009
Yemen01	211	9	1	4676	451

Appendix B.4

In this section we provide the entire validation data sets used in Chapter 6, these include GDP sectoral breakdown from 1997 to 2001 for primary (agriculture), secondary (industry and manufacturing) and tertiary (services) production. Trade dependence data including exports as % of GDP and domestic demand growth is also provided in this section.

AUSTRIA UK (£)	Agriculture	Industry	Manufacturing	Services	Exp%GDP	DD%
1997	0,256	7,682	5,121	17,411	41,80	1,20
1998	0,791	6,854	3,690	18,716	43,40	3,00
1999	0,518	7,507	4,919	17,862	45,50	3,00
2000	0,469	7,738	4,924	15,241	50,20	2,60
2001	0,464	7,664	5,109	15,096	52,50	-0,20
2002	-	-	-	-	52,90	-1,00

BELGIUM UK(£)	Agriculture	Industry	Manufacturing	Services	Exp%GDP	DD%
1997	0,242	6,539	4,360	17,438	74,70	2,90
1998	0,248	6,935	4,458	17,585	75,20	3,20
1999	0,246	6,157	4,433	17,977	75,70	2,40
2000	0,447	6,032	4,468	16,086	85,60	3,30
2001	0,444	5,998	4,443	15,773	85,60	0,50
2002	-	-	-	-	82,40	1,10

DENMARK UK(£)	Agriculture	Industry	Manufacturing	Services	Exp%GDP	DD%
1997	1,282	8,652	6,088	22,110	36,40	4,90
1998	-	-	-	-	35,80	4,00
1999	0,652	6,843	4,562	24,765	38,10	0,10
2000	0,892	7,729	5,054	21,106	44,20	1,90
2001	0,894	7,750	5,067	21,163	45,00	0,90
2002	-	-	-	-	44,80	1,20

FINLAND UK(£)	Agriculture	Industry	Manufacturing	Services	Exp%GDP	DD%
1997	0,945	8,029	5,904	14,641	38,80	5,10
1998	1,012	8,599	6,323	15,681	38,60	4,30
1999	0,743	6,937	5,203	16,846	37,70	2,30
2000	0,927	7,882	5,796	14,373	42,90	3,10
2001	0,698	7,682	6,052	14,665	40,00	1,70
2002	-	-	-	-	38,70	0,60

FRANCE UK(£)	Agriculture	Industry	Manufacturing	Services	Exp%GDP	DD%
1997	0,485	6,306	4,608	17,461	25,50	0,70
1998	0,498	6,480	4,735	17,945	26,10	4,20
1999					25,90	3,70
2000	0,663	5,749	4,201	15,698	28,60	4,50
2001	0,665	5,759	3,987	15,948	28,00	2,00
2002	-	-	-	-	27,10	1,10

GERMANY UK(£)	Agriculture	Industry	Manufacturing	Services	Exp%GDP	DD%
1997	0,259	-	-	-	27,90	0,60
1998	0,259	-	-	-	29,00	2,40
1999	0,257	7,204	5,403	18,268	29,60	2,80
2000	0,228	7,077	5,251	15,524	33,80	1,80
2001	0,226	6,999	5,419	15,353	35,30	-0,80
2002	-	-	-	-	35,90	-1,60

GREECE UK(£)	Agriculture	Industry	Manufacturing	Services	Exp%GDP	DD%
1997	1,158	1,896	1,053	7,477	19,70	3,60
1998	-	-	-	-	19,80	4,70
1999	0,802	2,290	1,260	8,244	20,50	2,90
2000	0,853	2,558	1,279	7,249	24,10	4,30
2001	0,887	2,329	1,331	7,874	22,70	4,40
2002	-	-	-	-	20,50	3,10

IRELAND UK(£)	Agriculture	Industry	Manufacturing	Services	Exp%GDP	DD%
1997	-	-	-	-	79,70	10,00
1998	-	-	-	-	85,80	9,40
1999	1,285	8,738		15,934	87,60	8,70
2000	1,006	9,050	7,039	15,084	97,40	8,60
2001	1,073	11,269	8,855	14,758	98,40	4,40
2002	-	-	-	-	93,70	2,90

ITALY UK(£)	Agriculture	Industry	Manufacturing	Services	Exp%GDP	DD%
1997	0,608	6,286	4,055	13,383	24,10	2,70
1998	0,626	6,469	4,174	13,982	24,10	3,00
1999	0,615	5,328	3,894	14,550	23,40	3,00
2000	0,559	5,591	3,913	12,672	25,90	2,50
2001	0,566	5,469	3,961	12,825	26,10	1,70
2002	-	-	-	-	24,80	0,80

LUXEMBOURG UK(£)	Agriculture	Industry	Manufacturing	Services	Exp%GDP	DD%
1997					119,30	6,50
1998	0,444	9,767		34,186	127,30	7,20
1999					135,90	6,30
2000					151,90	3,60
2001	0,442	8,389		36,645	152,40	3,90
2002					145,30	0,10

NETHERLANDS UK(£)	Agriculture	Industry	Manufacturing	Services	Exp%GDP	DD%
1997	0,730	6,566	4,377	17,022	61,10	3,90
1998					61,00	4,80
1999	0,760	6,077	4,051	18,738	60,30	4,30
2000	0,704	6,332	3,987	16,415	67,30	2,80
2001	0,722	6,494	4,089	16,836	65,30	1,40
2002					61,70	-0,10

PORTUGAL UK(£)	Agriculture	Industry	Manufacturing	Services	Exp%GDP	DD%
1997	0,424	3,395	2,334	6,790	30,40	5,30
1998					30,80	6,60
1999	0,462	3,117		7,967	29,70	5,70
2000					31,50	3,00
2001	0,427	3,203	2,029	7,047	30,80	1,30
2002					30,00	-0,50

SPAIN UK(£)	Agriculture	Industry	Manufacturing	Services	Exp%GDP	DD%
1997					26,70	3,50
1998	0,450		2,699	3,749	27,20	5,70
1999	0,609	4,262	0,000	10,504	27,50	5,60
2000	0,564	4,374	2,822	9,313	30,10	4,50
2001	0,581	4,359	2,761	9,590	29,90	3,00
2002					28,40	2,60

SWEDEN UK(£)	Agriculture	Industry	Manufacturing	Services	Exp%GDP	DD%
1997					41,40	1,30
1998					42,50	4,30
1999					42,60	3,40
2000	0,539	7,817	4,531	18,598	45,80	4,00
2001	0,491	6,623	4,334	17,416	45,20	0,10
2002					43,30	0,60

UK UK(£)	Agriculture	Industry	Manufacturing	Services	Exp%GDP	DD%
1997	0,450	6,974	4,725	15,074	28,70	3,60
1998	0,481	7,449	5,046	16,099	26,80	4,80
1999	0,246	6,145		18,188	26,40	3,80
2000	0,241	6,985	4,335	16,860	28,10	3,80
2001	0,239	6,445	4,535	17,186	27,30	2,70
2002					26,10	3,00