ISSN: 1940-9893 (Print), 1940-9907(Online) Volume 2, Number 1, March 2009

Journal of
Service Science and Management

Chief Editor : Samuel Mendlinger

www.scirp.org/journal/jssm

Scientific
Research
Publishing

Journal Editorial Board

ISSN: 1940-9893 (Print), 1940-9907(Online)
http://www.scirp.org/journal/jssm

***J. Serv. Sci. & Management*, 2009, 2**
Published Online March 2009 in SciRes (www.SciRP.org/journal/jssm)

CONTENTS

Volume 2 Number 1 March 2009

Journal of Service Science and Management (JSSM)

SUBSCRIPTIONS

The *Journal of Service Science and Management* (Online at Scientific Research Publishing, **www.SciRP.org**) is published quarterly by Scientific Research Publishing, Inc. 5005 Paseo Segovia, Irvine, CA 92603–3334, USA.

E-mail: **jssm@scirp.org**

Subscription rates: Volume 2 2009
Print: $50 per copy.
Electronic: freely available on **www.SciRP.org**.
To subscribe, please contact Journals Subscriptions Department, E-mail: **jssm@scirp.org**

Sample copies: If you are interested in subscribing, you may obtain a free sample copy by contacting Scientific Research Publishing, Inc. at the above address.

SERVICES

Advertisements
Advertisement Sales Department, E-mail: **jssm@scirp.org**

Reprints (minimum quantity 100 copies)
Reprints Co-ordinator, Scientific Research Publishing, Inc. 5005 Paseo Segovia, Irvine, CA 92603-3334, USA.
E-mail: **jssm@scirp.org**

COPYRIGHT

Statements and opinions expressed in the articles and communications are those of the individual contributors and not the statements and opinion of Scientific Research Publishing, Inc. We assume no responsibility or liability for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained herein. We expressly disclaim any implied warranties of merchantability or fitness for a particular purpose. If expert assistance is required, the services of a competent professional person should be sought.

PRODUCTION INFORMATION

For manuscripts that have been accepted for publication, please contact:
E-mail: jssm@scirp.org

J. Serv. Sci. & Management, **2009, 2: 1-9**
Published Online March 2009 in SciRes (www.SciRP.org/journal/jssm)

A Virtualization-Based Service System Development Method

Tong Mo[1]*, Zhongjie Wang[1], Xiaofei Xu[1], Xianzhi Wang[1]

[1]School of Computer Science and Technology, Harbin Institute of Technology, 150001, China.
Email: motong_hit@126.com*; rainy@hit.edu.cn; xiaofei@hit.edu.cn; wangxianzhi@hit.edu.cn

Received December 30th, 2008; revised January 24th, 2009; accepted February 18th, 2009.

ABSTRACT

Amazon and Taobao publish product introduction online for customers' personal choice and become the successful examples of modern service based on information and communication technologies. However, if they want to achieve a complex service through composing some simple services, the research of service composition platform is still rudimental. In order to achieve this goal, a virtualization-based service system development method is proposed. The service elements are described standardized as service component at first. Then the service component is virtualized to be a web service and is deployed on a service platform, which is similar to Amazon. Customers put forward their demands on this platform and the platform will auto-match and dispatch task to the component to build a real-world service system. Finally, an application in ocean logistics service is briefly introduced.

Keywords: *service description, service system development, service composition, service component, online publish-choice, ocean logistics*

1. Introduction

The services industry has increasingly grown over the last 50 years to dominate economic activity in most advanced industrial economies. According to relating statistics, in some developed countries above 90%'s labor force are working for service industry and it occupies over 70% of the Gross Domestic Product (GDP) [1]. The era of service economy has arrived.

Assisted by development of information and communication technologies, service pattern has been transformed from provider-customer traditional mode into a complex social technical ecosystem [2]. More and more service links and participators increase the cost of service interoperation significantly. In order to reduce the cost, a popular way of these new service systems is using a virtualization-based service for composition. Providers display what they can provide through virtualized information on the web, such as pictures and some textual descriptions, to replace the physical display. And customers can select and get what they want (or inversed) [3]. The virtualization-based service can ignore the space-time distance and make information transformation, service composition and data collection more conveniently. On one hand, it can reduce service cost such as time, money, energy, etc; on the other hand, the original data that is saved automatically provide strong support for analysis and management. Some successful examples are Amazon [4], Taobao [5], etc.

Though there are such a lot of effective applications, most of them are "composition for physical services". What are published by provider or customer are physical things, such as books, commodities, etc. The services, for example, transportation, healthcare, consultation, etc., are also remained in state of single publish-choice. But with increasingly fining of social division, service has become more and more complex and cross-domain. If customers want to get a complex service that is composed by a lot of service elements that are provided by multi-providers, this kind of composition is always done by manpower. But on the internet, the number of providers is much larger than the number in local range. So we need a computer-assistant method that can develop a complex service system by service composition.

The complex service system in this paper is not only an information system, but also includes non-software elements such as people and hardware. But the main idea of the method is similar with that we use software services building business information system based on the idea of Software as a service (SaaS) in software engineering [6]. Travel service gives us a good reference in this domain. A travel program can be seen as a composition of the single service elements such as showplace, catering, accommodation, transport, etc, and they are still often composed by the travel agents or tourists themselves.

- The benefits of these new service system development methods are as follows:
- Fast service query on-line and selection.
- Computer-assistant composition based on customers'

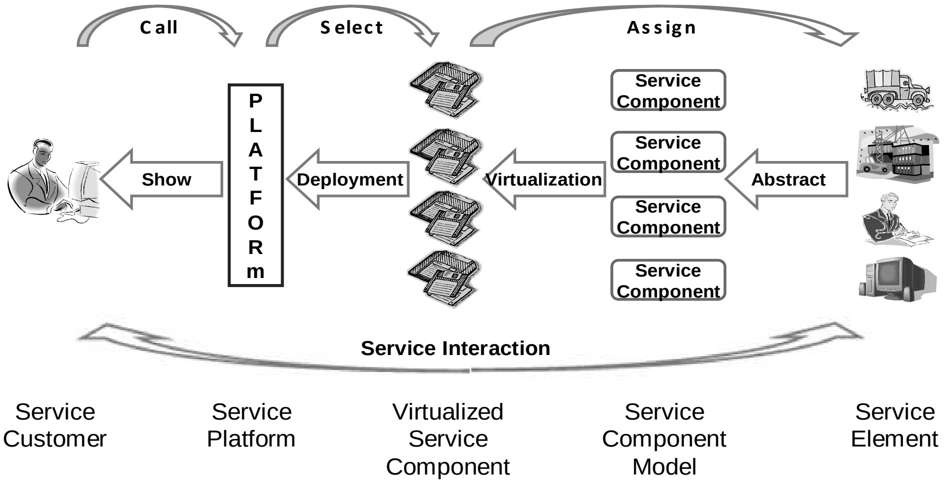


Figure 1. Big picture of building complex service based on virtualized service elements

voice and individual adaptation.

- Assign the service element in real world with the invoking and controlling service on web.

The rest of this paper is organized as follows. The big picture of the method is shown in Section 2. In Section 3, how to describe service elements is discussed and the virtualization method is proposed in Section 4. In Section 5, the arithmetic of service matching is briefly introduced and an application prospect is discussed based on a case of ocean logistic in Section 6. Finally the paper comes to the conclusions.

2. Big Picture: Virtualization-Based Service System Development

The big picture of building a complex service based on virtualized service elements is shown in Figure 1.

The process of this method includes three main phases and can be divided into eight steps.

Deployment phase:

Step1. The key elements of service such as behavior, actor, and resources are packaged as service components.

Step2. The service components are virtualized as software.

Step3. The virtualized service components are deployed onto the service platform.

Matching phase:

Step4. Customers log in the platform and input their individual demands through the interfaces that fits the habits of people. These demands are transformed into a standards XML format file.

Step5. According to the customer's demand on specific steps of service, platform will filter the relevant components, and get a smaller set of candidates. If some of the demands can find eligible components, the components closest to the demand will be reserved as a substitute.

Step6. According to the partial and overall demand of service, platform will match the candidate's components and feedback all the results in line with the demand. The result will be ranked with customer-defined policies to facilitate customer's choice.

Step7. Customer can self-edit the service based on the results, and submit a satisfactory outcome of the choice.

Scheduling phase:

Step8. Platform will assign the work to the software component that is selected by customer. And the *tasklist* of the component will add a record. Then the actor will execute service interaction following the *tasklist* with customer.

The key techniques and challenges in three phases are as follows:

- Standardized expression of the elements

In order to find what the customer want easily, the published things should comply with certain standards. The providers and customers often have different understanding and description with the same service element based on their own backgrounds and angles of view. This gap can be merged easily by similarity judgment and further communication likes a phone call or face to face talk in single-service-choice. But in complex service composition, cumulative effect of these gaps may make the result meaningless. Both of the above two successful composition platforms have a standard mode for each owner to show their goods and payment. But in this situation, a standardized expression for service is more complex than for physical goods. It not only contains the functional contents (input, output, precondition, quality index, etc), but also some other related elements. For example, the actor may be a very important factor that affects the customer satisfaction, and what they use is also need to be shown.

- Invoking service on web and assign the service element in real world

The service that published on web is not just a literal description of a variety of formats. It also can be called just as we operate software. A similar scene is that if we use telephone to appoint a service, there must be a call center (front) as a kind of interface. So for each service behavior, it also needs a software interface which is deployed on web for us to call. Though we publish and choose our service on web; it finally should be mapped and implemented in the real world. In Amazon, the book that a customer selects and chooses is just a virtualized image of real book on-line, but what the customer finally gets is that real book.

- Matching service based on customers' voice

Based on customers' voice, the most suited groups of services should be selected automatically. These results are collated by the integral sufficiency and key Quality of Service (QoS) parameters.

These three techniques and challenges will be discussed in the following sections.

3. Service Component: A Uniform Description of Service Elements

3.1 Information that Needs to be Described

Because there is still lacking a recognized definition for behavior that applies to all service fields, it is very hard to give a complete description of it. However, we describe it in order to establish a unified environment for customer and provider to choose and publish. So we only need to focus on the information that is used frequently in establishing the relationship between service demand and supply. A comparison with product behavior will help us to understand it more clearly.

The most important information is the function of behavior. It means what this behavior can do and what is the result of it. Though we can still use a binary group ＜ input, output ＞ as we describe the function of production behavior, the input and output of service behavior are mostly not physical things. For example, transformation behavior changes customers' position, so the input and output are the start points and end points respectively. In addition to input and output, quality index is another important part of service behavior function. Sometimes, for the same input and output, different performance of quality may means different services. In precedent, for the same transformation, the service may either be a normal one or an express one according to different time that should be spent on it.

Though we talk about service behavior and emphasize the difference from production, resource is also an important factor that impacts customers' choice. This resource is a kind of support for service behavior rather than the result. That means this resource is the thing that will be used during the service and impact customer's service experience.

In the field of manufacturing, customers just concern about the final product while don't care who made it. But because interaction is a major feature of service behavior, there could be some special request on the behavior actor. For example, a pretty miss front may reduce customers' complaints and an automatic response system could offer make them angry. Customers often put forward a number of requests to the actor to guarantee the behavior can be implemented smoothly.

3.2 Definition of Service Component

Based on the above analysis, we have gotten all the key elements that are needed for the establishment of the relationship between service demand and supply, and we use service component to depict it. Service component is a package of service actor and a behavior performed by this actor including all the supportive elements in conceptual level. It is the basic unit of service system offering a predefined functionality and able to communicate with other components. Reusability and independent execution are two important characteristics of service component and it can be plugged into different service systems. The structure is shown in Figure 2.

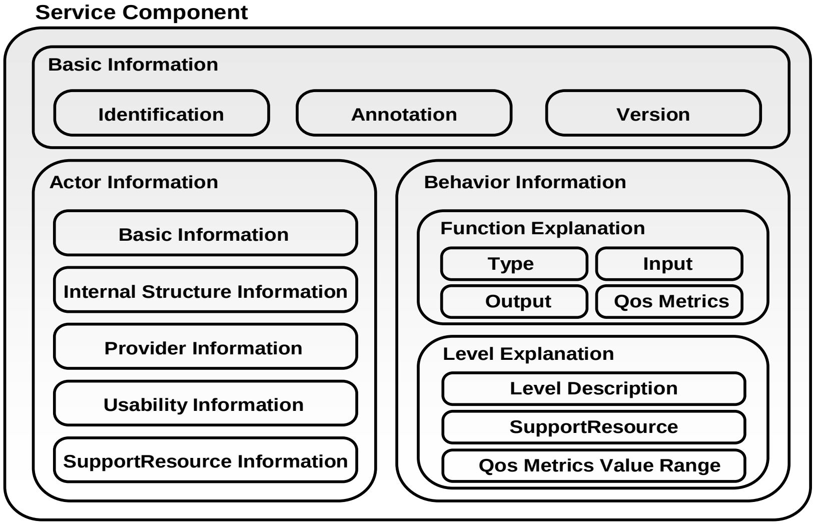


Figure 2. The structure of service component

As it is shown in Figure 2, service component gives us a unified format to describe the behavior that is needed in service by both customer and provider. It can be easily described by formal description language and here we use XML as a description language. The details are shown as follows:

• Basic Information: It involves the most basic information of a component, such as identification, annotation and version. Identification includes ID No. and name. Annotation is a text of narration. Version often includes Version No., creator and creating date.

• Actor Information: Actor is one of the two cores of service component. It describes the main body that provides service function in this component. In software domain, the main body of a traditional web service is a section of code and don't need to be further introduced. In service domain, the main may be human, software, hardware or a group of these things. It includes the following ingredients:

★ Basic Information. Such as ID No., name, etc.

★ Internal Structure Information. Sometimes, the actor of a behavior may be not executed by one entity but by a team or a group of entities. So the composition of actors is shown in this subsection.

★ Provider Information. Provider is the organization that the actor belongs to.

★ Usability Information. It shows the time when the actor can be used.

★ *SupportResource* Information. *SupportResource* is the thing that is needed and used by actor during the execution of service behavior. Sometimes, it is a little hard to distinguish a resource to be a *SupportResource* or a part of actor. Here we use an example to illustrate the nuance. In logistics service, a transport function is provided by a component. In this component, there is a truck driver and a truck that belongs to a cargo. In this cargo, if the truck driver just drives that truck, then they can be looked as a group, the truck is a part of actor. If the truck driver and truck don't have a binding relation and just many-to-many, the truck can be seen as a *SupportResource* for the truck driver.

• Behavior Information: Behavior is the other core of service component. As the same with traditional software behavior description; it also focus on the function description but adds one unique service character-service level.

Table 1. A case of service component package real service element

Natural language description of service element	*Service component description*

The name of truck driver is ZhangSan. He is a self-employed truck driver. His cell phone No. is 13936831568.	<tns:actor>
	    <tns:provider>
	    <tns:orgnizationName>ZhangSan</tns:orgnizationName>
	        <tns:description>Self employed</tns:description>
	        <tns:telephone>13936831568</tns:telephone>
	        <tns:contactPerson>ZhangSan</tns:contactPerson>
	    </tns:provider>
	</tns:actor>

He works from 9:00 AM to 7:00 PM everyday.	<tns:actor>
	    <tns:availableTime>
	        <tns:type>Everyday</tns:type>
	        <tns:startTime>9:00</tns:startTime>
	        <tns:endTime>19:00</tns:endTime>
	    </tns:availableTime>
	</tns:actor>

His business is truck transformation	<tns:behavior> <tns:functionType>transormation_truck</tns:functionType> </tns:behavior>

The normal speed of delivery is an average of 80km/h and carrying capacity is 5t.	<tns:behavior>
	    <tns:levels>
	        <tns:levelInfo>
	            <tns:levelName>normal</tns:levelName>
	            <tns:levelDescription/>
	        </tns:levelInfo>
	        <tns:QoSMetrics>
	
	    <tns:parameterName>speed</tns:parameterName>
	            <tns:minimumValue/>
	            <tns:maximumValue>80</tns:maximumValue>
	            <tns:parameterUnit>km/h</tns:parameterUnit>
	        </tns:QoSMetrics>
	        <tns:QoSMetrics>
	    <tns:parameterName>carrying_capacity</tns:parameterName>
	            <tns:minimumValue/>
	            <tns:maximumValue>5</tns:maximumValue>
	            <tns:parameterUnit>t</tns:parameterUnit>
	        </tns:QoSMetrics>
	    </tns:levels>
	</tns:behavior>

★ Function Explanation.

◎ Type. It is a quote to a concept of service domain ontology. The subjectivity of service naming has brought a lot of inconvenience to service query. So the ontology will give us a standardized solution. Unfortunately, it is very hard to build industry recognized domain ontology, so we often use a conventional concept for substitution.

◎ Input & output (I/O). The same with software, the I/O is the main reflection of function. Sometimes, the I/O is information, but the type may not only be an electronic data, but also some papery stuff such as documents, graph, table, etc. In most cases of service behaviors, resource is another kind of I/O. For example, patients may get some drugs from a medical care; a repair service is another typical case.

◎ QoS Metrics. A set of quality parameters are used to evaluate the result of service behavior. Some of them can be obtained directly from the basis data. Some of them are gained via further computing, statistics or analysis. Some typical parameters are time, cost, customer satisfaction, etc. For each parameter, it needs to show the related data, data's acquiring method and the formula.

★ Level Explanation. In traditional web service, each function of a software component has only one performance. But in service domain, with the same terms of be havior and the actor, the service component may have different performance. This distinction is reflected by different *SupportResource* and the value of QoS parameters. For example, in logistics service, a transport component's actor is a person and the behavior is transport. For the same task, if the actor does it by a car, the time will be less than an hour and the cost will be 100 RMB; on the contrast, if the actor does it on foot, the time will be more but the cost will be less. So we use service level with different *SupportResource* and QoS metrics value range to make a distinction. It is a unique character of service component and what makes it most different from software component.

◎ Level Description. Including identifying information and a text of narration.

◎ SupportResource. The *SupportResource* will be used in this level. It is a subset of the *SupportResource* in actor information.

◎ QoS Metrics Value Range. This is the value bound of quality parameters.

A case of service component is shown in table 1, and it shows a service component of transformation. The natural language description of service is in the left and the right is the XML based service component description follows the above structure. For example, the normal speed of delivery is an average of 80km/h and weight is 5t. So it has a service level named normal and has two QoS parameters: speed and weight. Of course, it may have other levels, such as a high-speed which may load less. As a complete service component description is too long to show, we just excerpt some critical segments to this paper.

4. Virtualization Method of Service Component

Service component is only a standard document for provider to publish his service and for customer to query easily. In contrast, the software virtualized version provides a user interface to show how to use the real service element packaged by component and record the running information. It is similar as the relationship between web service and the functional software system. According to the difference of actor and business level, the service component can be divided into human component, software component, hardware component and composite component.

4.1 Virtualization Method for Human Component

Human component represents the behavior that is executed by people or mainly depends on people. It is the dominant component and the difficulty of service description. But software human behavior is not a new topic and there has been already a well-formed specification: WS-HumanTask [7]. It is used with another specification: BPEL4People [8] which will also be referred in sector 4.4 to define the process of human-machine interaction. Though WS-HumanTask focuses on software domain, it provides us with fairly fine basis to draw on.

In our opinion, the software of human component is just like a software client of the real people. Component user can use that as software to assign work to service executants just like what we do by cell phone, email or even verbal order. This software client includes three main parts: *tasklist*, log and interface.

In the view of the idea of WS-humanTask, a job assigned to the actor of a component is treated as a task and the *tasklist* is a task schedule for the component. It records when, where, to whom that the component need to provide service. It is a kind of continuously information.

Log is the utilizing history of the component. It includes serial number, date and time, source, type, event and user.

The two parts above are used frequently to record the call of components. All the functions of software service component are invoked through the interface. The outside world can call it just like calling general software interface. According to different uses, the interfaces are divided into common interface and service function interface.

Common interface corresponds to the general functions of all software service components, for example, query component's information, amend component's state, manage task list and log, etc. This kind of interface doesn't relate to specific service and is just used to query and manage its own information.

Service function interface represents a typical service function of component, such as transporting, packing, checking container, etc. This kind of function is reflected in the form of service task and service interface is the interface of a task.

The steps of making an XML file of service component to be software are as follows:

Step1. Building new software which is named same with the service component and adding an empty *tasklist* and log file.

Step2. Initializing the software and generating the common function of it: *tasklist* management and log management.

Step3. Adding the query function of software and linking with the XML file of service component.

Step4. Building the service function call interface and the parameters is the input of service behavior.

Software interfaces of a component are shown in Table 2. The left is the XML based service component description, and the right is the corresponding interfaces. For example, this component has an actor, and the provider of actor is named ZhangSan, so we can get this information by call the interface hasProvider (actor).

4.2 Virtualization Method for Software Component and Hardware Component

SaaS is a model of software deployment where an application is hosted as a service provided to customers across the Internet [9]. So customer doesn’t have to take care about software maintenance, ongoing operation, and can on-demand pricing. It is gaining a great deal of attractions today and more and more businesses are adopting SaaS for cost-effective software management solutions as well as business structure and process transformations [10]. These well-packaged software services can be called directly and don’t need any form of repackaging [11].

Of course, there are also many legacy systems that is developed using the traditional way. If we want to use the function of these old systems, we need to package some function interface as a web service. The research of web service packaging technology is fairly mature so we won’t go details in this paper.

In fact, there is no pure hardware component, because a behavior can hardly be implemented by hardware alone. This kind of component usually packages two types of behavior.

One can be carried out by big machine which is independent and has a high degree of automation, such as Computerized Numerical Control (CNC) machine. But it still needs a command from the outside world. For this type of component, we need to issue a task order to the operator, and the call mode as well as virtualized method is the same as the human component in Sebsection 4.1. The only difference is that the hardware takes place to be the main body of the actor.

The other is the behavior of the automation equipment which is controlled by its own software system. Packaging this kind of component involves developing a special interface for the system. A typical application is the Global Position System (GPS).

Table 2. The result of a component that is software virtualized

	Service component	*Function of software virtualized comonent*	

Schema of component	<tns:actor>
	<tns:provider>
	<tns:orgnizationName>ZhangSan</tns:orgnizationName>
	<tns:description>Self employed</tns:description>
	<tns:telephone>13936831568</tns:telephone>
	<tns:contactPerson>ZhangSan</tns:contactPerson>
	</tns:provider>
	<tns:availableTime>
	<tns:type>Everyday</tns:type>
	<tns:startTime>9:00</tns:startTime>
	<tns:endTime>19:00</tns:endTime>
	</tns:availableTime>
	</tns:actor>
	…
		Query function	hasActor(component)
			hasProvider(actor)
			hasAvailableTime(actor)
			…
			The return values of these functions are a string of the results.

Behavior of component	<tns:behavior>
	<tns:functionType>transormation_truck</tns:functionType>
	<tns:input>
	<tns:inputInformation>
	<tns:billObject>
	<tns:name>PaicheBill</tns:name>
	<tns:parameters>
	<tns:address/>
	<tns:arrivetime_plan/>
	</tns:parameters>
	</tns:billObject>
	</tns:inputInformation>
	</tns:input>
	</tns:behavior>
		Service function	callService(component, level, inputstring) // The value of inputstring is a string that is composed by the input of the behavior and separator. In this case, the inputstring is “Harbin Westgreat Street 207# @@@ 2008-11-21 11:00”. The reurn value of this function is a taskid.
			getResult(component, taskid) // Get the result of the service task. The return value is a result object.

		Task and log function	taskQuery(condition)
			logQuery(condition)
			…

4.3 Virtualization Method for Composite Component

The three types of service components mentioned above are basic components executed by single actor. But in the actual process of service, many acts of service process are scheduled following the relatively fixed mode, and some of the components are ordered with the regular emergence of high-frequency. In order to simplify the selection of components and service matching, reduce the computational complexity, and improve the efficiency of the service schedule, we combine these components a larger one for further reusability. It is similar with the establishment of the large size software component and web service in the software domain.

Because the three basic components are all packaged as software, we can directly draw on the existing standard used to build a large size web service to build composite service component. The common industry practice is using Web Services Business Process Execution Language (WS-BPEL) [12] and WS-BPEL Extension for People (BPEL4People) [8] to specify the information interactions between software service, and orchestrate them as a process. At the beginning of a BPEL process, <partnerLink> and <variable> is used to define the link of service and declare variables. The process is combined by <sequence> which is a group of activities that is called one by one, and <flow> which is a group of parallel activities. In the main body of process, <invoke> is used to call a service and <receive> is used to wait the service to callback. The control of the implementation logic includes <switch>, <while> and <pick>. The other details of BPEL and BPEL4People can be found in documents 8 and 12.

5. Computer-Aided Service Selection and Matching

Service component and virtualization is a kind of computer understandable way for service elements describing, saving and calling. So the computer can pre-select and match the service and provide us options, when customers make a complex service needs. And it can release us from the heavy work of comparison, calculation and communication, when there are more and more service options.

The needs of customers' are a lot of constraints on total or local quality indicators, such as function, time, cost, credit, etc. Service selection and matching is choosing appropriate service components and organizing them correctly to satisfy the customers. It equals to a multi-objective optimization problem and we use a genetic algorithm (GA) based service matching algorithm (SMA) to achieve it.

- Input

The input of SMA includes four parts: QoS parameter trees, service flow model, constraints sets, and service components sets.

QoS parameter trees are a set of key performance indicators of services and are used to evaluate the quality. The name and calculation method of these parameters are relatively stable and are divided by industry.

Service flow model is a XML file that is parsed from customers' voice. On one hand, it shows which simple parts the complex service includes; on the other hand, it describes the sequential logic of these simple parts and is used as the basis for some types of parameters such as time.

Constraints sets are the other part of customers' voice. They are divided into three types: local constraint is the constraint for one simple part; partial constraint is for two or more simple parts and global is for all the simple parts. All of the constraints are described according two formats: The first one is simple condition which is a triple form <qp, mo, value>; qp is quality parameter, mo is mathematical operator and value is the number and unit. For example, "time < 5 hour" is a simple condition. The other one is complex condition which is a logical expression of simple condition. For example, "if time < 5 hour then money = 100 RMB else money = 50 RMB". All the constraints have a weight to show the importance for customers.

Service components sets are the sets of service components classified according to service domain.

- Output

The output of SMA includes four parts: service components set, result of evaluation, and scheduling plan. Service components set are the components selected for each simple parts of the complex service. Result of evaluation is the sufficiency of the service components set to the customers' voice. Scheduling plan is a set of information for calling the service components, including time, place, and other useful information.

- Implementation logic

The implementation logic of SMA is the same with GA, and the principle of algorithm is shown as follows:

```
1. Initial population (first generation)
2. Determine the fitness of each chromosome/ individual
3. Repeat:
Perform selection
Perform crossover
Perform mutation
Determine the fitness of each individual
Until the stopping criterion applies
4. Return last population
```

The key operations of SMA are follows:

★ Chromosome coding

A chromosome can be seen as a result of service selection and matching. It is a set of gene with a certain order and each gene represent a service component. For example, in a travel services, a chromosome is a complete tourist routes and genes are attractions, transport, accommodation, catering, etc. A population is a set of the chromosomes that in the domain of this complex service.

★ Genetic operation

Generating: The population is generated following a random way and the population size is fifty. Each gene in one chromosome is generated by the service component that has the same function type. Here we use a pre-selection technique to limit the scope of the service components. We filter out those components which do not meet the local constraints before the SMA. So the service composition is under the condition of the components that meet the local constraints and if some of the gene can't find eligible components, the components that are closest to the demand are reserved as a substitute.

Selection/Reproduction: Here we use a mechanism for the survival of the fittest and the worst five chromosomes fall into disuse. Another mechanism in SMA selection is differential reproduction. The chromosomes are put into the breeding pond according to a [0,2] probability which is calculated by its fitness.

Crossover: New chromosomes are generated by exchanging the genes in the same location of two randomly selected chromosomes. Considering the characteristics of service composition, we crossover the chromosomes under niche protection. Niche presents a common feature of the same species in biology. In service composition, some simple parts of a complex service are also required to have some same features. For example, some parts should be provided by the same provider. So the niche protection is used to avoid the broken of these same features through the crossover.

Mutation: Mutation changes some genes of a chromosome and brings some new genes to the population. In SMA, the chromosome and the position of genes are selected randomly.

★ Fitness determination

In SMA, we use customer satisfaction to be the fitness of a complex service and use penalty function method to determine the satisfaction.

$$f(c) = P(c) - N(c)$$

The *f(c)* is fitness of a chromosome and it equals to the positive satisfaction *P(c)* minus the negative impact *N(c)* of those constraints that are not bound to meet.

The satisfaction of the chromosome *P(c)* is expressed as

$$P(c)=\sum_{i=1}^{n} w_i \prod_{j=1}^{m} S_j(i_j)^{wj} \sum \lambda_{as}$$

$$\left(\sum_{i=1}^{n} w_i = 1, \sum_{j=1}^{m} w_j = 1 \right)$$

There are *n* genes in a chromosome and w_i is the weight of gene i. For each gene, the number of QoS parameters is m, and w_i is the weight of parameter *j*. $S_j(i_j)$ is the satisfaction function of the QoS parameters *j* of gene *i*. λ_{as} is the additional contributions to the satisfaction of those genes that are intrinsic satisfaction associated.

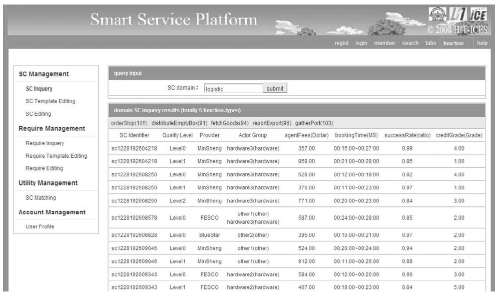
6. Case Study

The full container load (FCL) export business is one of the core businesses in ocean logistics service domain. It is a deep-division-business and the business chain can be broken down into a number of coupling parts of lower level. Because of the high degree of specialization, one participator is usually responsible for one single-part in the business chain. So in a FCL export business, it often involves many organizations and in most cases the number is greater than five. This situation leads extremely large cost for establishing business relations such as time, money and energy. At this stage in the industry, this problem is mainly solved through a long-term cooperative relationship. However, it makes the service to be a low-optional one and very likely to run another way counter to the trend of "On-demand". The contradiction is a typical problem exists in many service fields that have similar characteristics. So we choose it as case background and try to verify the virtualized service ecosystem's convenience and improvement on efficiency of system building.

A typical FCL export business mainly includes five parts: booking cabins, container yard choice and container allocation, loading goods, export declarations, and container gate-in. According to business needs, these parts can be further divided into different levels. For example, the part of loading goods includes container preparing, truck distributing, goods transformation etc. All of these can be seen as different sized behaviors and can have the providers. Firstly, they are packaged into service components according to the rules above. The main components for the five business links are shown in Figure 3. Figure 4 shows a matching result according to a customer's requirement. As soon as the customer confirms a solution from the result set, the platform will send the task message to the real actor through the virtualized software, and start the service

7. Conclusions

Nowadays, though service develops rapidly, the system for behavior service is still rudimentary and lacks coordination. The results of this paper try to make up for this. Although the virtualized system is still in the simulating and testing stage, it opens a brilliant prospect that we can orchestrate behavior service just like what we have done in web service domain.



query input

SC domain : logistic

domain SC inquery results (totally 5 function-types)

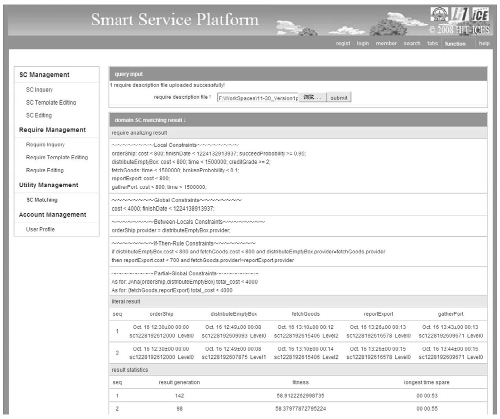
orderShip(105) distributeEmptyBox(91) fetchGoods(94) reportExport(96) gatherPort(103)

SC Identifier	Quality Level	Provider	Actor Group	agentFees(Dollar)	bookingTime(MS)	successRate(ratio)	creditGrade(Grade)
sc1228192604218	Level0	MinSheng	hardware3(hardware)	357.00	00:15:00~00:27:00	0.99	4.00
sc1228192604218	Level1	MinSheng	hardware3(hardware)	858.00	00:21:00~00:28:00	0.85	1.00
sc1228192608250	Level0	MinSheng	hardware3(hardware)	628.00	00:12:00~00:18:00	0.82	4.00
sc1228192608250	Level1	MinSheng	hardware3(hardware)	376.00	00:11:00~00:23:00	0.97	1.00
sc1228192608250	Level2	MinSheng	hardware3(hardware)	771.00	00:20:00~00:23:00	0.84	3.00

sc1228192608578	Level0	FESCO	other1(other)
			hardware3(hardware)
				587.00	00:24:00~00:28:00	0.85	2.00

sc1228192608828	Level0	BlueStar	other2(other)	395.00	00:10:00~00:21:00	0.97	2.00
sc1228192609046	Level0	MinSheng	other1(other)	524.00	00:20:00~00:24:00	0.94	2.00
sc1228192609046	Level1	MinSheng	other1(other)	912.00	00:11:00~00:26:00	0.88	2.00
sc1228192609343	Level0	FESCO	hardware2(hardware)	584.00	00:12:00~00:26:00	0.90	3.00
sc1228192609343	Level1	FESCO	hardware2(hardware)	467.00	00:19:00~00:23:00	0.84	5.00

Figure 3. Service components deployed on the platform



query input

1 require description file uploaded successfully!

require description file : F:\WorkSpaces\11-30_Version1p

domain SC matching result :

require analizing result

```
~~~~~~~~~~~Local Constraints~~~~~~~~~~~
orderShip: cost < 800; finishDate < 1224132913937; succeedProbobility >= 0.95;
distributeEmptyBox: cost < 800; time < 1500000; creditGrade >= 2;
fetchGoods: time < 1500000; brokenProbobility < 0.1;
reportExport: cost < 800;
gatherPort: cost < 800; time < 1500000;

~~~~~~~~~~Global Constraints~~~~~~~~~~
cost < 4000; finishDate < 1224138913937;

~~~~~~~~~~Between-Locals Constraints~~~~~~~~~~
orderShip.provider = distributeEmptyBox.provider;

~~~~~~~~~~If-Then-Rule Constraints~~~~~~~~~~
if distributeEmptyBox.cost < 800 and fetchGoods.cost < 800 and distributeEmptyBox.provider=fetchGoods.provider
then reportExport.cost < 700 and fetchGoods.provider!=reportExport.provider

~~~~~~~~~~~Partial-Global Constraints~~~~~~~~~~~
As for: JAha{orderShip,distributeEmptyBox} total_cost < 4000
As for: {fetchGoods,reportExport} total_cost < 4000
```

literal result

seq	orderShip	distributeEmptyBox	fetchGoods	reportExport	gatherPort
1	Oct. 16 12:30±00 00:00 sc1228192612000 Level0	Oct. 16 12:49±00 00:08 sc1228192609093 Level0	Oct. 16 13:10±00 00:12 sc1228192615406 Level2	Oct. 16 13:26±00 00:13 sc1228192616578 Level0	Oct. 16 13:43±00 00:13 sc1228192609671 Level0
2	Oct. 16 12:30±00 00:00 sc1228192612000 Level0	Oct. 16 12:49±00 00:08 sc1228192607875 Level1	Oct. 16 13:10±00 00:14 sc1228192615406 Level2	Oct. 16 13:26±00 00:15 sc1228192616578 Level0	Oct. 16 13:44±00 00:15 sc1228192609671 Level0

result statistics

seq	result generation	fitness	longest time spare
1	142	58.9122262998735	00 00:53
2	98	58.37977872795224	00 00:55

Figure 4. Matching result

From the case test we can find that, though the initialization of platform and add service component may cost a lot, for each using, it just needs less cost to edit customer's demand. A prominent advantage of virtualized system building is that the cost (time, money, and energy) is less impacted by the number of provider. In contrast to this, the cost in traditional system increases dramatically when the provider number grows. Though they use stabile service chain to improve the running condition, the custom degree is also limited. Sometimes, getting a full accordance with the wishes of customer solution from a large number of candidates by manpower can be an unpractical mission. Because the service matching is auto-calculated in virtualized system building, so it can release us from the complex computation and compare work, and we become able to pay attention to work more meaningful. At the micro level, we can use statistical methods and data mining to find which component and which combination is often used and then further explore the reasons and the law. At the macro level, based on statistics for a period of time, we may find the evolution of each part in a service business chain from an ecosystem angle.

Future work includes: establishing a sound interface system, applying this system into real service, and refining the component matching algorithm.

8. Acknowledgement

Research works in this paper are partial supported by the National High-Tech Research and Development Plan of China (2006AA01Z167, 2006AA04Z165) and the National Natural Science Foundation (NSF) of China (60673025).

REFERENCES

[1] J. Spohrer and P. Maglio, "Emergence of service science: Services sciences, management, engineering (SSME) as the next frontier in innovation," Nordic Service Innovation Workshop, Oslo, Norway, 2005.

[2] P. Maglio, S. Srinivasan, J. T. Kreulen, and J. Spohrer, "Service systems, service scientists, SSME, and innovation," Communications of the ACM, New York, Vol. 49(7): pp. 81−85, 2006.

[3] J. Spohrer and P. Maglio, "Emergence of service science: Services sciences, management, engineering (SSME) as the next frontier in innovation," Nordic Service Innovation Workshop, Oslo, Norway, 2005.

[4] http://www.Amazon.cn.

[5] http://www.Taobao.com.

[6] T. Mo, J. M. Xu, Z. J. Wang, Y. F. Ma, H. Y. Huang, Y.Wang, A. Liu, J. Zhu, and X. F. Xu, "SCS: A case study on service composition in automobile supply chain," Journal of Harbin Institute of Technology, Harbin, China 2008 Supplement 1. pp. 323−329, 2008.

[7] http://www.oasis-open.org/committees/download.php/27549/ws-humantask-1.1-spec-wd-02.doc.

[8] http://xml.coverpages.org/bpel4people.html.

[9] http://en.wikipedia.org/wiki/SaaS.

[10] SaaS 2.0: Software-as-a-Service as Next-Gen Business Platform, Saugatuck Technology.

[11] SaaS Integration Platforms: The Looming SaaS Deployment and Support Dilemma, Saugatuck Technology.

[12] http://www.oasis-open.org/committees/wsbpel.

(Edited by Vivian and Ann)

J. Serv. Sci. & Management, **2009, 2: 10-14**
Published Online March 2009 in SciRes (www.SciRP.org/journal/jssm)

Performance Evaluation Model of Engineering Project Management Based on Improved Wavelet Neural Network

Qinghua Zhang[1]**, Qiang Fu**[1,2]

[1]College of Water Conservancy and Civil Engineering, Northeast Agricultural University, Harbin China, [2]Visiting Scientist, Department of Renewable Resources, University of Alberta, Edmonton, T6G2E3, Canada.
Email: fuqiang@neau.edu.cn

Received October 8th, 2008; revised January 20th, 2009; accepted February 16th, 2009.

ABSTRACT

The scientific and reasonable performance evaluation is advantageous to promote the comprehensive management level of engineering projects. Benefited from constrictive and fluctuant of wavelet transform and self-study, self-adjustment and nonlinear mapping functions of wavelet neural network (*WNN*)*, and based on the existing assessment method and the index system, the performance evaluation model of engineering project management is established. One company is taken as the study object for this model. Compared with the conventional method, the influence of human factor is eliminated, thus the objectivity of the measure results is increased. A satisfactory result is concluded, thus a new approach is presented for engineering project management performance evaluation.*

Keywords: *wavelet neural network, entropy function, project management, performance evaluation*

1. Introduction

Project management is the systematic analysis and objective evaluation for the management of completed projects. It can put forward some suggestions for the future management and improving decision-making levels. Scientific and rational project management performance evaluation is conducive to improve the level of integrated management. At present the fields of academia and engineering had been achieved some results on this issue.

On the basis of fuzzy theory, a fuzzy integrative evaluation model of engineering management performance evaluation is developed [1,2]. Besides, main object method [3] is used for project management performance evaluation. However, the relationship between index systems of project management performance evaluation are non-linear, it is difficult to determine the model to express. And the subjectivity of the evaluation process is increased when specialists are required to determine the index weight. So there are some drawbacks in the traditional evaluation model.

Wavelet neural network (WNN) is constrictive and fluctuant of wavelet transform and has self-study, self-adjustment and nonlinear mapping functions of neural network which has made certain research achievements in the field of pattern recognition [4].Project management performance evaluation also belongs to pattern recognition, thus this paper tried to set a model using wavelet neural network model, with a view to produce good results.

2. Establishment of Index Systems in Engineering Project Management Performance Evaluation

According to the main object method, which means choosing one or two main objectives as the main objective of evaluation as long as other secondary objectives meet certain requirements. Therefore, take project investment, construction period, quality and safety as evaluation indexes as a basis for performance evaluation in the project management.

According to the existing documents, the following evaluation criteria are taken, as shown in Table 1.

3. Engineering Project Management Performance Evaluation Model Based on WNN

3.1 Structure Design of WNN Model

WNN [5] is a new type of function connected neural network based on the wavelet analysis. It is beneficial for the nonlinear function approximation, using non-linear wavelet replace the usual nonlinear neural activation function (such as Sigmoid function).

It is definition of the square accumulated function space:

$$L^2(R)=\{x(t):_R\int\left|x(t)\right|^2dt<\infty\} \tag{1}$$

In the function space, select a mother wavelet function (also known as wavelet basis function) $\psi(x)$ to meet the restrictive conditions:

$$C_{\phi}=\int_{R}\frac{\left|\psi(\omega)\right|^{2}}{\left|\omega\right|}d\omega<\infty \qquad (2)$$

where: $\psi(\omega)$ is the Fourier Transform of $\psi(x)$, then stretch and translational transform $\psi(x)$, wavelet basis function can be obtained.

$$\psi_{a,b}(x)=\frac{1}{\sqrt{|a|}}\psi\left[\frac{x-b}{a}\right]\ a,b\in R \qquad (3)$$

where: a is scaling factor and b is time translation factor. The signal can be approximated with a special constructed neural network. The transfer function is not Sigmoid nonlinear function but wavelet function. This paper uses Morlet wavelet function:

$$\psi(x)=\cos(1.75x)\exp(-x^{2}/2) \qquad (4)$$

A three-layer (one input layer, one hidden layer and one output layer) feed-forward network can approximate a nonlinear mapping with any degree of accuracy under normal circumstances. Aiming at the selected indexes of project management performance evaluation, the model will be expected to take construction period advance rates, cost saving rate, quality control scores, and security control as the input of a network, that means the number of input nodes is 4; take performance evaluation value as the output, that means the number of network output nodes is 1; the number of hidden nodes is 10 by texting, then a 4-10-1 three-layer network is established. Performance evaluation model structure of engineering project management based on WNN is shown in Figure 1.

3.2 Learning and Training of WNN

Through the study of network optimization indexes, set amendment of the network and wavelet function parameters by error back propagation algorithm, then reach the most optimal learning effects gradually. Learning algorithm steps as followed with application of Matlab7.0 programming:

Step 1: Set the input and output samples [6]. Produce evenly and randomly five numbers and their relative grades of experience given in the interval-level of Table 1 as the learning samples of network. Take 1, 2, 3, and 4 respectively as the four grades of excellent, good, qualified and poor in output layer. The data is shown in Table 2.

In order to solve the incommensurability between project investment, construction period, quality and security, in accordance with the project management performance evaluation and the actual situation of indexes, transform the original data of evaluation indexes into a range of [−1,1] as input of the network using nonlinear transformation function. Define the individual indexes utility function [7]:

$$\beta_{i}=\frac{1-e^{-ky}}{1+e^{-ky}} \qquad (5)$$

where: y is the relative value indicators of the actual value and the plan value; k is relative to the impact laws of the evaluation indexes on project management performance, and it can be the experience value.

Step 2: Initialization setting. The weights, threshold value of the network, as well as wavelet translation parameters and the scaling parameters are given evenly and randomly in the range [−1,1].

Table 1. Level partition of quantitative indexes

Level	Construction period advance rate	Cost saving rate	Quality control scores	Security control
Excellent	≥0.12	≥0.06	85	≤1‰
Good	0.06≤x<0.12	0.03≤x<0.06	75≤x<85	1‰≤x<2‰
Qualified	0≤x<0.06	0≤x<0.03	65≤x<75	2‰≤x<3‰
Poor	<0	<0	<65	≥3‰

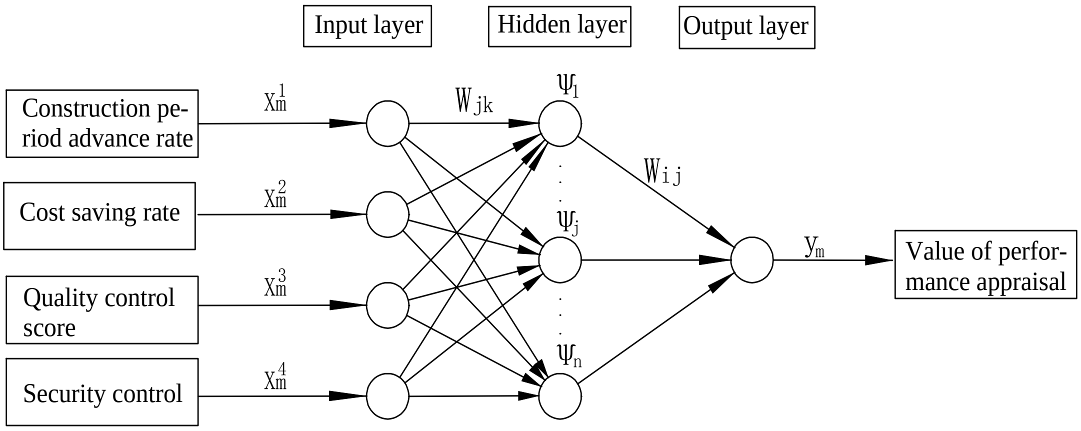


Figure 1. Engineering project management performance evaluation model

Table 2. The learning sample of project management performance evaluation model

Samples sequence number	Performance evaluation indexes				Performance evaluation
	Construction period advance rate	Cost saving rate	Quality control scores	Security control	
1	0.6555	0.7828	99.8274	0.0003	1
2	0.8906	0.7017	86.1415	0.0010	1
3	0.8207	0.5451	91.5090	0.0008	1
4	0.4573	0.3263	87.3103	0.0009	1
5	0.3033	0.9687	89.8753	0.0005	1
6	0.1064	0.0539	78.6508	0.0019	2
7	0.0799	0.0347	78.4761	0.0019	2
8	0.0836	0.0338	78.2145	0.0019	2
9	0.0848	0.0538	80.7094	0.0020	2
10	0.0763	0.0528	75.7117	0.0012	2
11	0.0335	0.0194	72.6387	0.0025	3
12	0.0012	0.0191	74.4602	0.0022	3
13	0.0456	0.0238	66.7535	0.0021	3
14	0.0350	0.0056	71.0947	0.0028	3
15	0.0195	0.0065	67.3856	0.0023	3
16	−0.1334	−0.0119	24.2475	0.1252	4
17	−0.9901	−0.1962	33.3788	0.4619	4
18	−0.2843	−0.8583	47.2422	0.6146	4
19	−0.9360	−0.2155	41.1326	0.8282	4
20	−0.7105	−0.9060	64.6641	0.4198	4

Step 3: Self learning process of WNN. Calculate the output value of wavelet neural network model according to the formula (6) using the current network parameter.

$$y = f\left(\sum_{j=1}^{n} w_{ij}\psi_{a,b}\left(w_{jk}x - b_j\right)/a_j\right) \qquad (6)$$

There are still many shortcomings of the present application of wavelet neural network. In order to solve the local minimum of network training, entropy function is used as cost function of neural network to accelerate the learning speed of the network.

Entropy function value is larger than the mean square error function value when a network error is large, the adjustment of the network parameters is larger than the use of the mean square error function, and network convergence speed is larger; entropy function value quickly becomes smaller when network errors becomes smaller, the adjustment of parameters correspondingly decrease to avoid oscillation, thereby the convergence rate of the network is improved, and meanwhile the network parameter adjustments around the local minimum is not zero, that is to say the network will not be at a local minimum. Therefore entropy function is taken as cost function of the network instead of the mean square error.

$$E = -\sum_{i=1}^{m}\left[d_i \ln y_i + \left(1-d_i\right)\ln\left(1-y_i\right)\right] \qquad (7)$$

where: d_i is the desired output of the network, y_i is actual output of the network.

Step 4: Repetitive adjustments of the network parameters. The memory and generalization ability can be rapidly realized, and convergence accelerated to attain forecast accuracy. The various parameters of WNN are modified using formula (8)-(11).

$$W_{ij} = W_{ij} - \eta\frac{\partial E}{\partial W_{ij}} + \alpha\Delta W_{ij} \qquad (8)$$

$$W_{jk} = W_{jk} - \eta\frac{\partial E}{\partial W_{jk}} + \alpha\Delta W_{jk} \qquad (9)$$

$$a_i = a_i - \eta\frac{\partial E}{\partial a_i} + \alpha\Delta a_i \qquad (10)$$

$$b_i = b_i - \eta\frac{\partial E}{\partial b_i} + \alpha\Delta b_i \qquad (11)$$

where: η is learning rate, α is momentum factor.

Step 5: When a network error is less than a pre-determined value or learning steps of maximum training value is reached, wavelet neural network learning is stopped, otherwise return to the third step to repeat training until the expected output of the network is generated.

3.3 Model Testing and Practical Application

The network is learned and trained repeatedly using Table 2. The test results show that the actual output and the expected output is very close, the error accuracy is as small as 10^{-4}, so it meets the requirement. Training results shows in Table 3.

Take out ten completed projects from a construction company in the last three years. Their project management performance evaluations are done. The basic data of projects are shown in Table 4.

Preprocess the data of evaluation indexes set by the main objectives method, and then applying the trained network, a project management performance evaluation model based on wavelet neural network is established. After the calculating of the model, project management performance evaluation results of the ten projects are obtained, which are shown in Table 5.

Table 3.The comparison between desired output of the network and actual output of the network of the learning samples

Sample number	1	2	3	4	5	6	7	8	9	10
Desired output	1	1	1	1	1	2	2	2	2	2
Actual output	0.9994	1.0004	1.0007	1.0001	0.9905	2.0074	1.9893	2.0424	1.9893	1.9980
Sample number	11	12	13	14	15	16	17	18	19	20
Desired output	3	3	3	3	3	4	4	4	4	4
Actual output	2.9858	2.9844	3.0324	2.9064	3.0705	3.9992	4.0000	3.9973	4.0005	3.9989

Table 4. Basic data of projects

Project number	Construction area/m^2	Schedule control/day		Cost control/million yuan		Quality control score		Security/‰	
		Plan time	Actual time	Contract price	Settlement price	Plan	Fact	Plan	Fact
1	109480	430	426	42632.00	42590.98	80	77	3	2
2	38962	485	487	5706.71	5610.00	80	80	3	2
3	212156	460	455	14808.53	14800.00	80	81	3	2
4	56766	365	334	9082.57	9078.13	90	85	3	0
5	130792	730	700	35655.12	35650.80	90	90	3	0
6	59004	550	548	21880.00	21850.56	87	87	3	0
7	72055	800	791	18542.00	18510.00	80	77	3	3
8	47797	355	335	7920.00	7856.00	75	69	3	1
9	98000	360	335	12000.00	11890.67	75	65	3	2
10	90009	560	547	1112.80	1112.02	75	72	3	1

Table 5. Results of performance evaluation

Project number	1	2	3	4	5	6	7	8	9	10
Output of network	2.9685	3.1004	2.9372	1.0166	2.1092	2.9166	3.1209	2.0630	2.9191	3.0247
Evaluate result	3	3	3	1	2	3	3	2	3	3

Seen from Table 5, all these ten projects have reached more than qualified rating. One of them is excellent and two of them are good. Evaluation results can be used as the basis of the plan implementation of future projects, progress control, cost assessment, quality control and security control, and it is helpful to enhance the level of integrated management.

4. Conclusions

Started from the purpose and requirements of project management performance evaluation, using the characteristics of wavelet neural networks which can describe the complex and nonlinear relationship, the relationship model between evaluation indexes and project management performance evaluation is established. The followed three conclusions are obtained:

1) There is no need to determine the weights by people using wavelet neural network model in the evaluation process, so the defects brought by experts when determining the weight is eliminated, therefore the accuracy and objectivity of the evaluation result is improved.

2) In order to avoid wavelet neural network training being at a local minimum, entropy function is taken as cost function of the network instead of the mean square error to accelerate the learning speed of the network .The testing proves the method feasible.

3) Performance evaluation model based on wavelet neural network explores a new way for project management performance evaluation, and enriches the performance evaluation method. The model has strong feasibility and accuracy which can be used as a scientific and rational basis for performance evaluation.

Table 5. Results of performance evaluation

Project number	1	2	3	4	5	6	7	8	9	10
Output of network	2.9685	3.1004	2.9372	1.0166	2.1092	2.9166	3.1209	2.0630	2.9191	3.0247
Evaluate result	3	3	3	1	2	3	3	2	3	3

REFERENCES

[1] L. X. Zhang, “Fuzzy comprehensive evaluation of the expressway project management performance evaluation,” Road traffic technology (applications) 5, pp. 169−171, 2007.

[2] S. H. Cai, M. Y. Zhou, and Z. C. Ye, “Fuzzy integrative evaluation method of management performance of engineering project,” Journal of YangZhou university (natural science), 5, pp. 57−60, 2002.

[3] V. Ireland, “The role of management actions in the cost, time and quality performance of high-rise commercial building projects,” Construction Management and Economics, 3, pp. 59−87, 1985.

[4] T. Brian and H. Szu, “Adaptive wavelet classification of acoustic backscatter and imagery,” Optical Engineering, 7, pp. 2192−2202, 1994.

[5] Q. H. Zhang and A. Benveniste, “Wavelet networks,” IEEE Transactions on Neural Networds, 6, 889−898, 1992.

[6] X. Y. Zhao., Q. Fu, and Z. X. Xing, “Application of projection pursuit grade evaluation model in comprehensive evaluation of changes in soil quality,” Acta Pedologica Sinica. 1, pp. 164−168, 2007.

(Edited by Vivian and Ann)

J. Serv. Sci. & Management, **2009, 2: 15-28**
Published Online March 2009 in SciRes (www.SciRP.org/journal/jssm)

CDS Evaluation Model with Neural Networks

Eliana Angelini[1], Alessandro Ludovici[1]

[1]University "G. d'Annunzio" of Pescara, University "G. d'Annunzio" of Pescara
Email: e.angelini@unich.it, a.ludovici1@tin.it

Received May 6th, 2008; revised December 27th, 2008; accepted February 5th, 2009.

ABSTRACT

This paper provides a methodology for valuing credit default swaps (*CDS*). *In these financial instruments a sequence of payments is promised in return for protection against the credit losses in the event of default. Given the widespread use of credit default swaps, one major concern is whether the credit risk has been priced accurately. Credit risk assessment of counterparty is an area of renewed interest due to the present financial crises.*

This article proposes a non parametric model for estimating pricing of the CDS, using learning networks, based on the structural approach pioneered by Merton [1] *as regards the independent variables; he proposed a model for assessing the credit risk of a company by characterizing the company's equity as a call option on its assets. The model that we are introducing turns out peculiar not only for the use of the neural network, but also for the use of the implied volatility of one-year options written on the shares of the analyzed companies, instead of historical volatility: this leads to a higher capability of getting the signals launched by the market about the future creditworthiness of the firm* (*historic volatility, being a medium value, brings in temporal lags in the evaluation*). *Besides, our analysis differs from the structural approach for the fact that it considers the* 30*-month mean-reverting historical series for CDS spreads, and this turns out to be one of the main advantages of our forward-looking model.*

Keywords: *credit derivatives, CDS, neural networks, pricing models, credit spreads, implied volatility*

1. Introduction

In recent years, the market for credit derivatives has expanded dramatically. Credit derivatives are flexible and efficient instruments that enable users to isolate and trade credit risk. Credit derivatives allow users to isolate credit risk from other quantitative and qualitative factors associated with owing an exposure. Hence, they can be used to transfer and hedge credit risk in an efficient and flexible manner, customized to a client's requirements. This transfer of credit risk may be complete or partial, and may be for the life of the asset or for a shorter period. Credit risk includes not just default or insolvency risk but also changes in credit spreads and thereby market values, changes in credit ratings and generic changes in credit quality. Credit derivatives can be used when a sale in the cash market is either not efficient or not possible. Even when cash market alternatives exist, credit derivatives may be preferred because they do not require funding. Furthermore, since derivatives are over-the-counter contracts, transactions are confidential. Finally, speed of settlement and liquidity are reasons why credit derivatives are a better alternative to the reinsurance market. Credit derivatives are swaps, forward and option contracts, particularly credit default swaps (CDS); they can be used to hedge against all these types of credit risk. For a simple credit default swap, over some time period, one counterparty (the protection seller) receives a predetermined fee payment from another counterparty (the protection buyer); in return, the protection seller agrees that in the case of a credit event of a reference entity, it will pay the seller the loss on a bond of the reference entity, that is the bond's par value less its recovery.

Nowadays, banks, corporate, hedge funds, insurance companies and pension funds are hugely exposed as buyers or sellers, or both. By transferring the risk, the CDS have acted as a kind of insurance and provided incentives for risk-taking. They are therefore at the heart of the present crisis.

Given the widespread use of credit default swaps, as an investment or a risk management tool, one major concern is whether the credit risk has been priced accurately. This article proposes a non parametric model for estimating pricing of these credit derivatives, using learning networks. The recent application of nonlinear methods, such as neural networks to credit risk analysis, shows promise of improving on traditional credit models. Neural networks differ from classical credit systems mainly in their black box nature and because they assume a non-linear relation among variables. The two main issues to be defined in a neural network application are the network typology and structure and the learning algorithm. The connections (links) among neurons have an associated weight which determines the type and intensity of the information exchanged. As regards the independent variables of the model, we start from the typical assumption of the structural approach based on the theoretical foundation of Merton's [1] option pricing model: the relevant information in order to evaluate credit risk can be obtained from the market data of the analyzed companies. The model developed by Merton views a firm's equity as an option on the firm (held by the shareholders) to either repay the debt of the firm when it is due, or abandon the firm without paying the obligations. What makes that model successful is its reliance on the equity market as an indicator, since it can be argued that the market capitalization of the firm (together with the firm's liabilities) reflect the solvency of the firm. Therefore, option pricing theory is used in order to create a link between the credit market and the securities market. The model that we are introducing turns out peculiar not only for the use of neural networks, but also for the use of the implied volatility of one-year options written on the shares of the companies, instead of historical volatility: this leads to a higher capability of getting the signals launched by the market about the creditworthiness of the firm (historical volatility, being a medium value, brings in temporal lags in the evaluation). Besides, our analysis differ from the structural approach for the fact that it consider the 30-month historical series for CDS spreads, and this turns out to be one of the main advantage of our forward-looking model.

The paper is organized as follows. The paper begins, in Section 1, by stating the implications of credit derivatives in portfolio credit risk management. In Section 2, we first briefly overview the main principles and characteristics of neural networks, focusing the attention above all on the concepts that are most useful for the application to financial instruments; then we describe the pricing model we developed and tested for credit derivatives. Section 3 develops the theory underlying our implementation of Merton's model. Section 4 describes the data and we present our results: the effectiveness of neural network in approximating the evaluation of credit default swap is illustrated. As regards the sample, it includes 18 American firms, relative to various fields, including financial institutions which, operating typically with a high leverage due both to the activity carried out and to the laws concerning the capital of banks, usually introduces remarkable factors of distortion in parametric models. We shall show that neural networks are not affected by this problem. The temporal range embraces the period September 2002-March 2006: we have considered the five-year CDS spread relative to each firm, for a total of 180 observations on a quarterly basis obtained through the Fitch™ database. As already pointed out, implied volatility has a determining role among the variables; in fact we have obtained a positive correlation with CDS spreads equal to 0.6338. Leverage is another key variable, obtained dividing the face value of the debt of the firm by the total of its liabilities (including the market capitalization), getting the data from the Bloomberg™ database. We have considered the risk free rate equal to one-year constant maturity Treasury Bills yield, taken from the Federal Reserve System database. We then discuss in detail the experimental settings and the results we obtained, leading to considerable accuracy in prediction. The architecture of the neural network is feed-forward, trained for 17000 learning epochs using the back-propagation algorithm, with two hidden layers of 9 and 10 neurons each: by the study carried out it turns out obvious that neural networks are able to totally capture the variability relative to the market dynamics of credit default swap. The paper ends evidencing that, as far as this field of the financial markets is concerned, neural networks constitute a highly valid instrument of calculation: in fact there still does not exist in literature a formula of evaluation for the CDS, able to tie the quoted spreads to the specific underlying variables of each examined firm, and the neural network can, as will be shown, satisfy this lack with high effectiveness, facing the problem of determination of the functional form from a statistical point of view. As we will show, it is easy to calculate the sensitivity of the CDS spread to each independent variable, in order to determine a statistical pricing formula for CDS.

The paper concludes with a discussion of advantages and limitations of the solution achieved.

2. Credit Derivatives: Innovative Financial Instruments

Credit derivatives are financial instruments used to transfer credit risk of loans and other assets. They are bilateral financial contracts with payoffs linked to a credit related event such as a default, credit downgrade or bankruptcy. There are various types, but the basic structures of all credit derivatives are swaps, options and forwards. Due to their high flexibility credit derivatives can be structured according to the end-users' needs. For instance, the transfer of credit risk can be effected to the whole life of the underlying asset or for a shorter time, and the transfer can be a complete or a partial one. Delivery can take place in the form of over the counter contracts or embedded in notes. Moreover, the underlying can consist of a single credit-sensitive asset or a pool of credit-sensitive assets [2].

2.1 Credit Derivatives: Products and Structures

The most important and widely used credit derivative is a credit default swap[1]. It is an agreement in which the one counterparty (the protection buyer) pays a periodic fee, typically expressed in fixed basis points on the notional amount, in return for a contingent payment to the other counterparty (the protection seller) in the event that a third-party reference credit defaults. A default is strictly defined in the contract to include, for example, bankruptcy, insolvency, and/or payment default. The definition of a credit event, the relevant obligations and the settlement mechanism used to determine the contingent payment are flexible and determined by negotiation between the

[1] The credit default swap is also known as credit default put, credit swap default swap, credit put or default put.

counterparties at the inception of the transaction. Since 1991, the International swap and Derivatives association (ISDA) has made available a standardized letter confirmation allowing dealers to transact credit swaps under the umbrella of an ISDA Master Agreement. The evolution of increasingly standardized terms in the credit derivatives market has been a major growth because it has reduced legal uncertainty that hampered the market's growth.

The contingent payment in the event of default can be identified as either:

- a payment of par by the protection seller in exchange for physical delivery of the defaulted underlying;
- a payment of par less the recovery value of the underlying as obtained from dealers;
- a payment of a binary, i.e. fixed, amount.

Credit default swaps can be viewed as an insurance against the default of the underlying or a put option on the underlying. Figure 1 exhibits the basic structure of a credit default swap.

Moreover, there is the total return swap, in which one counterparty (total return payer) pays the other counterparty (total return receiver) the total return of an asset (the reference obligation) for receiving a regular floating rate payment, such as Libor plus a spread. "Total return" comprises the sum of interest, fees and any change-in value payments (any appreciation or depreciation) with respect to the reference obligation.

In contrast to the credit default swap, the total return swap does not only transfer the credit risk but also the market risk of the underlying; it effectively creates a synthetic credit-sensitive instrument. A total return swap allows an investor to enjoy all of the cash flow benefits of a security without actually owing the security.

Credit spread option is an option on a reference credit's spread in the loan or bond market. In a spread put option one party pays a premium for the right to sell a bond to a counterparty at a certain spread at a definite time in the future. A credit spread option gives the buyer protection in the event of any unfavourable credit migration. In a default option, the asset can be put only on default. The credit spread is the differential yield between the reference credit and a pre-determined benchmark rate. Thus, in credit spread derivatives, payment is based on the movement of the value of one reference credit against another.

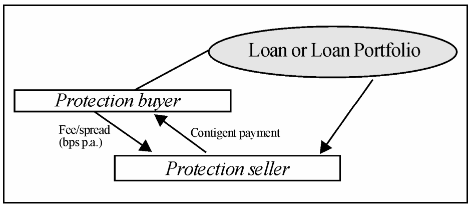


Figure 1. Credit default swap

that pays out if a specified company's rating is downgraded. This kind of option is sometimes embedded in bond structures.

Finally, credit linked notes are created by embedding credit derivatives in notes. Credit derivatives have the advantage that funding is not necessary; whereas credit linked notes have the benefit of avoiding counterparty risk. Credit linked notes are frequently issued by special purpose vehicles (corporations or trusts) that hold some form of collateral securities financed through the issuance of notes or certificates to the investor. The investor receives a coupon and par redemption, provided there has been no credit event of the reference entity. The vehicle enters into a credit swap with a third party in which it sells default protection in return for a premium that subsidizes the coupon to compensate the investor for the reference entity default risk.

2.2 Fundamental Attractions of Using Credit Derivatives

In theory, credit derivatives are tools that enable financial operators to manage their portfolio of credit risks more efficiently; they enable market participants to devise flexible personal approaches to the management of credit risk associated with a variety of underlying financial assets. The promise of these important instruments has not escaped regulators and policymakers. "Credit derivatives and other complex financial instruments have contributed to the development of a far more flexible, efficient, and hence resilient financial system than existed just a quarter-century ago" [3].

The credit derivatives market offers its users a range of tools which enable the transfer of credit risk. A brief review of the available products reveals that in most cases one party to a transaction receives a fee and commits to provide the other party with a payment should the credit quality of a third party deteriorate. Whilst the mechanism contained in these products are easy to understand, the broad range of applications is not immediately obvious.

The users of the risk-management benefits of credit derivatives tend to be quite diverse. An increasingly important user group includes financial institutions, corporate and fund managers. Financial institutions have embraced the full range of benefits; the use of credit derivatives by banks has been motivated by the desire to improve portfolio diversification and to improve the management of credit portfolios. Corporate is also looking to reduce the credit exposure to key trading partners and specifically they are interested in using credit derivatives to isolate credit risks in project financing. For fund managers, although the asset benefits of credit derivatives still suffer from lack of liquidity, the use of structures that hedge out spread risk has some appeal.

This paragraph focuses on a range of uses for credit derivatives and divides them between credit risk management and asset opportunities[2] [4,5].

[2] For more detailed information on the characteristics of credit derivatives see DAS, S., (1998); TAVAKOLI, J.M., (1998).

2.2.1 Using Credit Derivatives for Managing Credit Risk

The principal feature of these instruments is that they separate and isolate credit risk facilitating the trading of credit risk with the purpose of:

- replicating credit risk;
- transferring credit risk;
- hedging credit risk.

In practice, the rationale behind a transaction may relate to the management of credit lines, to regulatory capital offsets, to balance sheet optimization, portfolio hedging and diversification or pure risk reduction itself. Credit derivatives can be used as a risk management tool by portfolio managers to:

- Achieve portfolio diversification: credit derivatives can be used to achieve portfolio diversification by allowing access to previously unavailable credits. They can also be used to diversify across a range of borrowers and to gain exposure to an asset without owing it.
- Reduce concentration risk: investors can reduce portfolio credit risk concentrations using derivatives structures; they can thus manage country and industry risks. Reducing credit concentration in loan portfolios is commonly viewed as the main use of credit derivatives. However, to date credit derivatives are generally referenced to assets which are widely traded, i.e. for which market prices are readily available, or for which a rating by an international agency is at hand.
- Manage exposures while maintaining client relationships. Changes to credit risk management in the banking sector are an additional factor contributing to greater use of credit derivatives. Investors can use credit derivatives to reduce exposures without selling them. This effectively frees up credit lines, allowing more business to be done with a customer. Furthermore, a bank that is concerned about credit loss on a particular loan can protect itself by transferring the risk to someone else while keeping the loan on its books. As part of their credit risk management, banks are viewing credit derivatives more and more often as tradable products, which can be transferred to third parties before the maturity date [6,7,8].
- Manage regulatory capital: the new supervisory rules provided for by Basel II are also increasing the incentives for banks to use credit derivatives. Where guarantees or credit derivatives are direct, explicit, irrevocable and unconditional, and supervisors are satisfied that banks fulfil certain minimum operational conditions relating to risk management processes, they may allow banks to take account of such credit protection in calculating capital requirements. A guarantee or credit derivative must represent a direct claim on the protection provider and must be explicitly referenced to specific exposures or a pool of exposures, so that the extent of the cover is clearly defined and incontrovertible. Other than non-payment by a protection purchaser of money due in respect of the credit protection contract it must be irrevocable; there must be no clause in the contract that would allow the protection provider unilaterally to cancel the credit cover or that would increase the effective cost of cover as a result of deteriorating credit quality in the hedged exposure. It must also be unconditional; there should be no clause in the protection contract outside the direct control of the bank that could prevent the protection provider from being obliged to pay out in a timely manner in the event that the original counterparty fails to make the payment due. There are cases where a bank obtains credit protection for a basket of reference names and where the first default among the reference names triggers the credit protection and the credit event also terminates the contract. In this case, the bank may recognise regulatory capital relief for the asset within the basket with the lowest risk-weighted amount, but only if the notional amount is less than or equal to the notional amount of the credit derivative. In the case where the second default among the assets within the basket triggers the credit protection, the bank obtaining credit protection through such a product will only be able to recognise any capital relief if first-default-protection has also be obtained or when one of the assets within the basket has already defaulted [9].

2.2.2 Asset Opportunities

Credit derivatives have evolved to become an important financial asset class. As already argued, credit derivatives enable credit risk to be separated from the funding component of its underlying instrument; as it is often the form of the underlying instrument that creates obstacles for the investor, this separation of the credit risk creates important opportunities. The decision to use the asset opportunities of credit derivatives tends to be based on one of the following needs:

- Access to new markets: investors can create new assets with a specific maturity not currently available in the market;
- Obtain tailored investments: credit derivatives can be used to create instruments with exact risk- return profile sought. Maintaining diversity in credit portfolios can be challenging. This is particularly true when the portfolio manager has to submit with constraints such as currency denominations, listing considerations or maximum or minimum portfolio duration. Credit derivatives are being used to address this problem by providing tailored exposure to credits that are not otherwise available in the wished form or not available at all in the cash market.
- Improve the risk-return profile of portfolios: credit derivatives offer new possibilities of turning a given market opinion into an investment strategy. This particularly entails assumption of specific types of

credit risk without the acquisition of the asset itself. Instead of purchasing a specific bond, a market participant who considers some credit risks to be overvalued can earn an attractive premium as a protection seller in the credit default swap market. Premiums are generated without having to tie up any capital for the purchase of a bond issue (at least as long as no credit event occurs). On the other hand, market participants who consider risks to be underestimated can purchase protection by paying a premium. Owing to the limited possibilities for short sales in the bond market, hedge funds are increasingly entering into positions in credit derivative market to implement their financial strategies. In particular:

- to hedge dynamic risks: exposures that change with market movements can be hedged using credit derivatives;
- to manage illiquid credits: credit derivatives can be utilized to actively manage risk in large illiquid loans portfolios;
- to execute short credit positions: credit derivatives can be employed to execute short credit positions without the risk of a short squeeze or high financing costs. Hence, investors can use them to hedge or take advantage of deteriorating credit qualities;
- to hedge declining credit quality: default and spread options and swaps can be used to hedge failing credit qualities. Credit spread options and swaps can be used to hedge fluctuations in credit spreads without having to wait for default to get a payout.

3. The Neural Network Model

The general structure of a neural network model consists of simple processing units called nodes that interact with each other using weighted connections. Each unit (node) receives and processes inputs, and delivers a single output. The input can be raw or output of other processing units. The output can be the final product or an input to another unit. In processing the inputs, the model assigns a weight to each input, where weights represent the relative strength or importance of inputs. A neural net essentially represents a nonlinear discriminant function as a pattern of connections between its processing units.

Neural networks have been used in different fields of study, such as engineering, medicine, physics and others. Although the relative structures differ remarkably with one another, it is possible to point out some fundamental principles regarding essentially the functioning of such operative instruments. Moreover, it is important to start the treatment emphasizing that, in order to analyze the financial dynamics, relatively little complex networks are effective, at least compared to those of other fields[3] [10,11,12].

Neural networks offer several advantages over the traditional statistical methods. First, neural networks do not require the restrictive assumptions imposed by conventional methodologies. Second, neural networks can develop input-output map boundaries that are highly non linear[4] [13,14]. Third, they have greater fault tolerance and adaptability. Neural network examines all information available and it can incorporate the new information into the analysis promptly through its memorization of previous learning; it updates its weighting scheme so that it continually “learns” from experience. Thus, neural networks are flexible, adaptable systems that can in corporate changing conditions.

3.1 Architecture of Neural Networks

A neural network relates a set of input variables $\{x_i\}$, i=1,2,..k to a set of one or more output variables $\{y_j\}$, j=1,2,..h. An essential characteristic of a neural network, differently from other methods of approximation, is that it uses one or more hidden layers, in which the input variables are transformed by a logistic or logsigmoid function: this characteristic, as shown later, gives to these instruments a particular efficiency in modeling nonlinear statistical processes.

In the feed-forward neural network parallel elaboration is associated to the typical sequential elaboration of the linear methods of approximation. In fact while in the sequential elaboration particular weights are given to the input variables through the neurons of the input layer, in the parallel one the neurons of the hidden layer operate further transformations in order to improve the predictions. The connectors (between the input neurons and the neurons in the hidden layers, and between these and the output neurons) are called synapses. The feed-forward neural network with a single hidden layer is the simplest and at the same time the most used network in the economic and financial field.

Therefore the neurons process the input variables in two ways: firstly forming linear combinations and lastly transforming these combinations through a particular function, typically the logsigmoid function, illustrated in

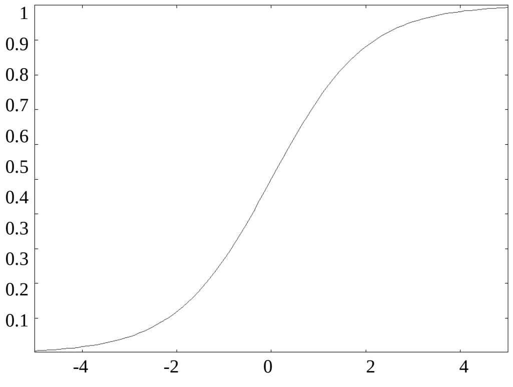


Figure 2. Logsigmoid function

[3] DOLCINO, F., GIANNINI, C., ROSSI, E., (1998). For a useful description of the phenomenon in general terms, see FLOREANO, D., NOLFI, S., (1993) and GORI, M., (2003).

[4] Such feature is important for financial analysis because several studies have shown that the relation between default risk and financial factor (variables) are often non linear. See WU and YU (1996); WU (1991).

Figure 2. An essential characteristic of this function is the threshold behavior near values 0 and 1, which turns out to be particularly suitable to economic problems, which usually, for very high (or very low) values of the independent variables, show little changes in response to small changes of the variables. At the analytical level, the neural network can be described by the following equations [15]:

$$n_{k,t} = \omega_{k,0} + \sum_{i=1}^{m} \omega_{k,i} x_{i,t} \tag{1}$$

$$N_{k,t} = L(n_{k,t}) = \frac{1}{1+e^{-n_{k,t}}} \tag{2}$$

$$y_t = \gamma_0 + \sum_{k=1}^{q} \gamma_k N_{k,t} \tag{3}$$

where L($n_{k,t}$) represents the logsigmoid activation function. It is a system with *m* input variables x_i and *q* neurons. A linear combination of these input variables, observed at time *t*, with the weights of the input neurons $\omega_{k,i}$ and the constant term (*bias*) $\omega_{k,0}$ forms the variable $n_{k,t}$. Then this variable is transformed by the logistic function and becomes the neuron $N_{k,t}$ at time or observation *t*. The set of *q* neurons at time or observation *t* is therefore linearly combined with the coefficient vector *k* and added to the constant term $\omega_{k,0}$ in order to obtain the output y_t concerning time or observation *t*, representing the prediction of the neural network for the analyzed variable. The feed forward neural network used with the logsigmoid activation function is often called multi-layer preceptor or MLP network. A highly complex problem could be treated widening this structure, and therefore using two (respectively N and P) or more hidden layers [15]:

$$n_{k,t} = \omega_{k,0} + \sum_{i=1}^{m} \omega_{k,i} x_{i,t} \tag{4}$$

$$N_{k,t} = L(n_{k,t}) = \frac{1}{1+e^{-n_{k,t}}} \tag{5}$$

$$\rho_{l,t} = \rho_{l,0} + \sum_{k=1}^{s} \rho_{l,k} N_{k,t} \tag{6}$$

$$\rho_{l,t} = \frac{1}{1+e^{-p_{l,t}}} \tag{7}$$

$$y_t = \gamma_0 + \sum_{l=1}^{q} \gamma_l P_{l,t} \tag{8}$$

Adding another hidden layer increases the number of parameters (weights) to be estimated by the factor (*s*+1) (*q*-1)+(*q*+1), since the net with a single hidden layer, with *m* input variables and *s* neurons has (*m*+1)*s*+(*s*+1) parameters, while the same net with two hidden layers and *q* neurons in the second hidden layer has (*m*+1)*s*+(*s*+1)*q*+ (*q*+1) parameters. However the disadvantage of these models for complexity does not consist of the number of parameters, which in any case use up degrees of freedom if the sample size is limited and requires a longer training time, but of the greater probability that the net converges to a local rather than global optimum. Anyway it has been demonstrated that a neural network with two layers is able to approximate any nonlinear function [16]. A further quality of this instrument consists exactly of the fact that it does not just approximate a phenomenon on the basis of a presumed functional form to be adapted, but at the same time it determines the functional form and proceeds to the evaluation of the weights.

In Figure 3 a net with a multiple number of output variables is illustrated. A neural network with a hidden layer and two output variables is described by the following equations:

$$n_{k,t} = \omega_{k,0} + \sum_{i=1}^{m} \omega_{k,i} x_{i,t} \tag{9}$$

$$N_{k,t} = L(n_{k,t}) = \frac{1}{1+e^{-n_{k,t}}} \tag{10}$$

$$y_{1,t} = \gamma_{1,0} + \sum_{k=1}^{q} \gamma_{1,k} N_{k,t} \tag{11}$$

$$y_{2,t} = \gamma_{2,0} + \sum_{k=1}^{q} \gamma_{2,k} N_{k,t} \tag{12}$$

It is possible to observe that adding an output variable implies the evaluation of (*q*+1) parameters more, equal to the number of neurons of the hidden layer increased of one unit. Therefore adding an output variable implies an increasing number of parameters to be estimated, equal to the number of the neurons of the hidden layer, not to the input variables. Using a neural network with multiple outputs makes sense only if these are closely correlated to the same set of input variables: as an example we could mention the temporal structure of the rates of inflation or of the rates of interest. One of the most common criticisms made to these instruments is that they are substantially black boxes: questions regarding the nature of the parameters, the reasons of the choice of their number, of

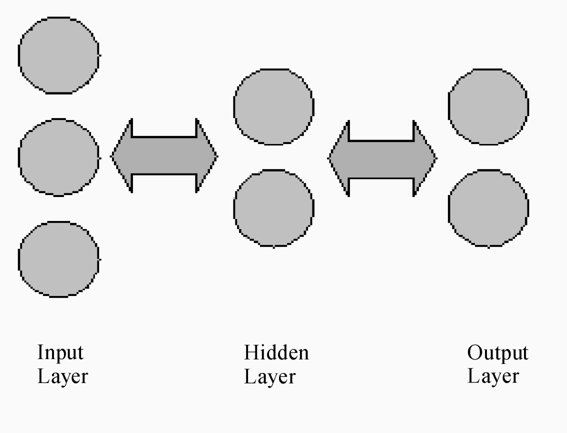


Figure 3. Neural network with one hidden layer and two output neurons

the number of the neurons, of the number of the hidden layers, the reasons that relate the architecture of the net to the structure of the underlying problem to be explained do not find an answer.

The risk, when models are based on a high number of parameters, is that their extreme flexibility [17], being able to explain anything and its opposite, ends up in not carrying any knowledge contribution. However, we must underline that the same criticism can be made to any statistical approximation method: therefore not only to neural networks, but also to linear models, univariate and multivariate regression and so on. Neural networks, in particular, are able to explain very irregular processes, on which it is therefore difficult to identify a precise relation of cause-effect. Therefore the black box criticism constitutes, paradoxically, also one of the greatest qualities of neural networks. In any case, the simplicity with which it is possible to increase the number of the parameters of the net must never make forget the importance, in any model, of the clarity of the assumptions.

3.2 Data Scaling

A neural network is not able to analyze data or to give solutions in absolute value: especially if there are data of an unusually elevated or reduced value, problems of overflow or underflow could happen. When instead sigmoid functions are used, it becomes indispensable to preprocess data: this family of functions in fact has a codominy of type [0,1] (or [−1,1] in the case of the log-sigmoid function), for which the values must be scaled to these intervals otherwise the output of the net would become useless, being equal to the superior or inferior threshold in correspondence of all the different values higher or lower than a determined limit. In other words, for a great amount of data not standardize to the interval the neurons would simply transmit the threshold value, so a wide part of the information would be lost. As far as the methods, the linear reduction transforms the series of values x_k in the series $\hat{x}_k$ x_k, using the following formulas:

$$\hat{x}_{k,t} = \frac{x_{k,t} - \min(x_k)}{\max(x_k) - \min(x_k)} \tag{13}$$

if the range is between 0 and 1, and

$$\hat{x}_{k,t} = 2\frac{x_{k,t} - \min(x_k)}{\max(x_k) - \min(x_k)} - 1 \tag{14}$$

if the desired range is between -1 and 1, while the logarithmic reduction uses the formula:

$$\hat{x}_{k,t} = \frac{\log(1 + x_{k,t})}{\log(\max(x_k))} \tag{15}$$

3.3 Learning Process

After the data have been scaled, we have to deal with the problem of the evaluation of the parameters (weights) through the process known as learning (training) of the neural network. Certainly it is a much more complex problem than the evaluation of the parameters of a linear model, as for the nature of high nonlinear complexity of neural networks. For these reasons numerous optimal solutions can exist, but they do not minimize the difference between the predictions of the net and the effective values to be evaluated. In short, in any non linear model it is necessary to begin the evaluation of the parameters on the basis of conditions which represent a guess of the value of the same. However, as it will be shown, the capability of the process of evaluation of the parameters to converge to a global optimum depends on the goodness of these initial hypothesis: in fact if it is situated near a local optimum instead of the global one [10], it is likely that the first one will be reached.

This is illustrated in Figure 3: the initial guess of the parameters (or weights of the neurons) could accidentally be situated wherever on the x-axis: if it is near a local minimum, the training process of the net would lead towards this. Later on, it will be observed that the training process of the network is completed when a point is reached in which the derivative of the loss function is null: we must remember that this condition, beyond the global optimum, identifies also the local ones and the saddle points. So it can be anticipated that if the learning coefficient, which indicates the sensibility of the net to the training process, is too low, this would lead to the impossibility of the network to escape from local optimums; while if it is too high, it could carry the training process to oscillate continuously far away from the optimum point, and therefore the network would diverge. In analytical terms, it is possible to illustrate the learning process of a net with two hidden layers, for which it is therefore necessary to determine the set of parameters $\Omega=\{\omega_{k,i}, \rho_{l,k},\gamma_l\}$.

The problem consists of [18] the minimizing of the loss function, defined as the sum of the squares of the differences between the observed data sample y and the prediction of the net $\hat{y}$:

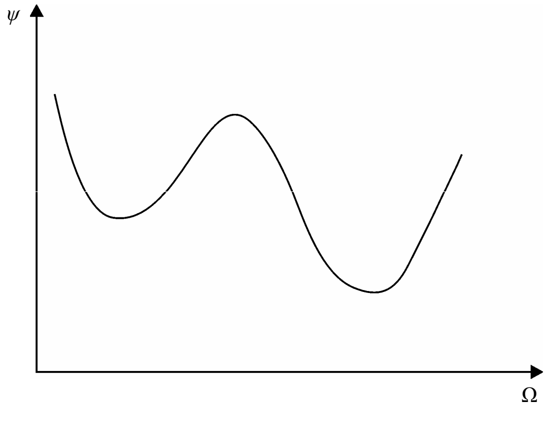


Figure 4. Example of succession of local and global minimums

$$\min_{(\Omega)} \psi(\Omega) = \sum_{t=1}^{T} (y_t - \hat{y}_t)^2 \quad (16)$$

$$\hat{y}_t = f(x_t;\Omega) \quad (17)$$

in which T is the number of the observations of the output vector *y*, and *f* (x_t;Ω) represents the neural network. Ψ is a nonlinear function of Ω. All nonlinear optimizations begin with an initial guess about the solution and try further, better solutions until finding the best possible within a reasonable number of iterations. Different methodologies have been proposed in order to lead this search: some make reference to complex results of logical- numerical analysis, e.g. genetic algorithms, in alternative to the classic method of the reduction of the gradient or Newton-Raphson method. In any case the chosen algorithm continues until the last iteration *n*, or in alternative a tolerance criterion can be set up, stopping the iterations when the reduction of the error function comes down a predefined tolerance value. In order to avoid local optimums, a solution could be to determine a first convergence of the process, and then to repeat it with a set of different initial parameters in order to verify whether the solution changes. Alternatively, numerous processes could be carried out to determine the best solution. However, there are the most important problems when the number of the parameters increases or the architecture of the network becomes particularly complex. Paul John Werbos proposed in the beginning of 1970's an alternative to the gradient method called back-propagation method. It is a very flexible method to avoid the problems caused by the evaluation of the Hessian matrix in the reduction of the gradient, and surely it is the most used method. In the passage from an iteration to the successive one in the process of evaluation of the parameters, the inverse Hessian matrix is in fact replaced by an identity matrix having dimension equal to the number *k* of the parameters, multiplied by the learning coefficient ρ:

$$(\Omega_1 - \Omega_0) = -H_0^{-1} Z_0 = -\rho Z_0 \quad (18)$$

In order to avoid oscillations this coefficient is chosen in the range [0.05,0.5] and it can also be endogenous, that is it can assume various values when the gradient comes down and the process seems to converge; or finally different coefficients for the various parameters can be adopted. However, the problem of the choice of this coefficient remains, together with the existence of local minimums. Moreover, low values of the learning coefficient, although as anticipated are able to avoid oscillations, can extend uselessly the convergence of the minimizing process. This can however be accelerated adding a 'momentum' for which at iteration *n* we will have:

$$(\Omega_n - \Omega_{n-1}) = -\rho Z_{n-1} + \mu(\Omega_{n-1} - \Omega_{n-2}) \quad (19)$$

Therefore, with μ generally equal to 0.9, the calculation of the parameters moves more fast outside a plateau in the error surface. Now we will briefly discuss the methods used to estimate the effectiveness of the output of the net. Relatively to the evaluation of the goodness of the predictions of the net, the most common index is R-squared (goodness of fit) especially as far as the capability of the net to predict the data with which it has been trained is concerned, and the root mean squared error (*Rmse*) as for the capability to generalize the predictions outside the data sample used for the training; in other words, divided the sample into two parts, the first (in sample) will be used in order to train the net, and the other (out of sample), in general equal to about 25% of total data, will be used to estimate the capability of the net to predict data coming from the same population but not used for the training.

However, as to the total amount of necessary data[5] [10], undoubtedly a neural network requires the evaluation of many more coefficients than, for example, a linear model, and this leads to the necessity of a wide sample. Surely the availability of wide samples improves the predictive abilities of the net, but it also implies longer training times. Moreover, the availability of a wide sample not always is a positive aspect, especially in the financial field where using very old data brings distortions in the models, because they tend to vary with extreme rapidity and therefore very remote data are no more in any relations with the present ones.

4. Credit Risk Approach: Our Assumptions

The recent history of financial markets shows how, to the impetuous development of the financial innovation process, which has invested all the structural components of the same, has been associated the constant engagement of the operators in finding more efficient computational methodologies, able to be an effective dynamic support of the analysis. Growing concerns about credit risk have created the need for sophisticated credit risk analysis and management tools. Credit risk measurement models and credit risk management tools are both of significant importance in the credit market.

The valuation of credit default swap depends on the credit quality of the reference entity. The default prediction has long been an important and widely studied topic. There are two main types of models that attempt to describe default processes in the credit risk literature: structural and reduced form models. The first approach is based on modeling the underlying dynamics of interest rates and firm characteristics and deriving the default probability based on these dynamics[6] [1,19,20,21]. So they use the evolution of firms' structural variables, such as asset and debt values, to determine the time of default. Merton's Model was the first modern model of default and is considered the first structural model. In Merton's model, a firm defaults if, at the time of servicing the debt, its assets are below its outstanding debt. In the second

[5] F. Dolcino, C. Giannini, and E. Rossi, where the concepts of "evaluation error" and "approximation error" are analyzed, 1998.

[6] R. C. Merton, 1974; F. Black and J. COX, 1976; F. A. Longstaff and E. Schwartz, 1995; H. E. Lelan and K. B. Toft, 1996; C. Dufresne and R Goldstein, 2001.

approach, instead of modeling the relationship of default with the features of a firm, this relationship is learned from the data. Reduced form models do not consider the relation between default and firm value in an explicit manner [22,23,24]. The time of default in intensity models is the first jump of an exogenously given jump process. The parameters governing the default hazard rate are inferred from market data. Structural default models provide a link between the credit quality of a firm and the firm's economic and financial conditions. Thus, defaults are endogenously generated within the model instead of exogenously given as in the reduced approach.

The focus of our model is on the structural approach, pioneered by Merton, with some important integration.

4.1 A Brief Review of the Structural Approach: Merton's Model

Merton proposes a simple model of the firm that provides a way of relating credit risk to the capital structure of the firm. The firm has issued two classes of securities: equity and debt. The equity receives no dividends. The debt is a pure discount bond. The value of the firm's assets is assumed to obey a lognormal diffusion process with a constant volatility. Merton adopts are the inexistence of transaction costs, bankruptcy costs, taxes or problems with indivisibilities of assets; continuous time trading; unrestricted borrowing and lending at a constant interest rate r; no restrictions on the short selling of the assets; the value of the firm is invariant under changes in its capital structure (Modigliani-Miller Theorem) and that the firm's asset value follows a diffusion process.

Merton models equity in this levered firm as a call option on the firm's assets with a strike price equal to the debt repayment amount (D). If at expiration (coinciding to the maturity of the firm's short-term liabilities, assumed to be composed of pure discount debt instruments) the market value of the firm's assets (V) exceeds the value of its debt, the firm's shareholders will exercise the option to "repurchase" the company's assets by repaying the debt. However, if the market value of the firm's assets falls below the value of its debt (V<D), the option will expire unexercised and the firm's shareholders will default. The probability of default (PD) until expiration is set equal to the maturity date of the firm's pure discount debt, typically assumed to be one year. Thus, the Pd until expiration is equal to the likelihood that the option will expire out of the money. To determine the PD, the call option can be valued using an iterative method to estimate the unobserved variables that determine the value of the equity call option, in particular, V (the market value of assets) and σ_V (the volatility of assets). These values for V and σ_V are then combined with the amount of debt liabilities D that have to be repaid at a given credit horizon in order to calculate the firm's distance to default, defined to be: (V-D)/ σ_V or the number of standard deviations between current asset values and the debt repayment amount. The higher the distance to default (denoted DD), the lower the PD. To convert the DD into a PD estimate, Merton assumes that asset values are log-normally distributed.

Define E as the value of the firm's equity and V as the value of its assets. Let E_0 and V_0 be the values of E and V today; in the Merton framework we have:

$$E_0 = V_0 N(d_1) - De^{-rt} N(d_2)$$

$$d_1 = \frac{\ln(V_0 / D) + (r + \sigma_V^2 / 2)T}{\sigma_V \sqrt{T}}$$

$$d_2 = d_1 - \sigma_V \sqrt{T}$$

where σ_V is the volatility of the asset value and r is the risk free rate of interest, both of which are assumed to be constant. Define D* = De^{-rt} as the present value of the promised debt payment and let L=D* /V_0 be a measure of leverage. Because the equity value is a function of the asset value we can use Ito's lemma to determine the instantaneous volatility of the equity from the asset volatility:

$$\sigma_E E_0 = \frac{\partial E}{\partial V} \partial_V V_0$$

$$\frac{\partial E}{\partial V} = N(d_1)$$

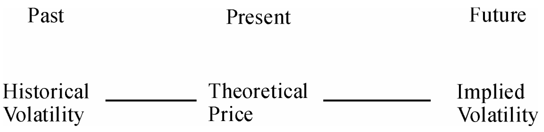
where σ_E is the instantaneous volatility of the company's equity at time zero. These equations allow V_0 and σ_V to be obtained from E_0, σ_E, L and T. The risk neutral probability, P, that the company will default by time T is the probability that shareholders will not exercise their call option to buy the assets of the company for D at the time T. This depends only on the leverage, L, the asset volatility, σ, and the time of repayment T.

4.2 CDS Valuation

In our analysis, we present some extensions because the model needs to make the necessary assumptions to adapt the dynamics of the firm's asset value process.

We suggest a new way of implementing Merton's model using implied volatility, instead of historical volatility: this leads to a higher capability of getting the signals launched by the market about the creditworthiness of the firm. The historical volatility is the realized volatility of a financial instrument over a given time period. Generally, this measure is calculated by determining the average deviation from the average price of a financial instrument in the given time period. Standard deviation is the most common but not the only way to calculate historical volatility. By definition, historical volatility will always be backward looking and lag the real-time volatility environment. In the current market environment, however, where both stocks and implied volatility measures are rising, many measures of historical volatility begin to seem no more useful.

The implied volatility of an option contract is the volatility implied by the market price of the option based on an option pricing model. Implied volatility is a forward-looking measure, and differs from historical volatility that is calculated from known past prices of a security.



Historical volatility tells us how volatile as asset has been in the past. Implied volatility is the markets view on how volatile an asset will be in the future. To determine an option's implied volatility, we have to use a pricing model. We can tell how high/low implied volatility is by comparing the market price of an option to the options theoretical fair value. This is why we need to use an option pricing model - to determine the fair value of an option and hence know if the market price for the option is over/under valued.

In our analysis, equity implied volatilities observed in the equity options market has received much exploration. Our neural network model is based on using the implied volatility of one-year options written on the shares issued by the company. It is an attractive alternative to the traditional structural approach; this implementation allows to use a forward-looking model. Otherwise, our model differs from the structural approach for the fact that it consider the 30-month historical series for CDS spreads: we show that the use of these credit spreads in addition to other inputs, provides a significant improvement in the accuracy of the model.

We use a model that takes these inputs:

- *Leverage* of the firm: the level of indebtedness is a significant enterprise-specific determinant of risk.
- *Implied volatility:* theoretical value designed to represent the volatility of the security underlying an option as determined by the price of the option. The factors that affect implied volatility are the exercise price, the risk-free rate, the maturity date and the price of the option.
- *Historical CDS spreads serie*: a CDS is a derivative that protects the buyer against default by a particular company. The CDS spread is the amount paid for protection and is a direct market-based measure of the company's credit risk. CDS spreads contain information which is significant for estimating the probabilities of the occurrence of credit events.
- *Recovery rate*: percentage of notional of the reference asset repays in the event of default.
- *Risk free rate*: is the interest rate that it is assumed can be obtained by investing in financial instruments with no default risk.

5. Data and Empirical Results

In this section the potentialities of neural networks in the approximation of the pricing of credit derivatives will be shown using real market data, collected from Fitch™ and Bloomberg™ data bases.

Starting from September 2002, we have collected on a quarterly basis data regarding 5-year maturity CDS spreads of 18 companies from various economic sectors, together with data concerning the leverage of the firms, the implied volatility of 1-year maturity call options written on the equities of the firms, and the risk free rate assumed to be equal to the 1-year constant maturity Treasury Bill yield. As regards the recovery rate, we have used the most commonly values adopted by the operators to price CDS, depending on the economic sector to which the reference entity belongs to. In the following diagrams we show the sample collected until March 2006, therefore covering 14 quarters.

As regards the risk free rate, we must consider that a portfolio made up of a risky bond with yield equal to *i* and a CDS written on it with a spread equal to *sp* is virtually free of any credit risk, so its yield must be equal to the risk free rate; therefore we have the following approximation:

Table 1. Details of the companies included in the sample

Sample description

N	Ticker	Name	Market Cap. (bln $)
1	AA	ALCOA Inc.	30,18
2	BA	Boeing Company (The)	71,91
3	CCL	Carnival Corporation	30,13
4	COX	Cox Communications Inc. *	5,9
5	CTX	Centex Corporation	6,15
6	CVS	CVS Corporation	26,96
7	CZN	Citizens Communications Corporation	4,81
8	FD	Federated Department Stores Inc.	23,16
9	GPS	Gap, Inc. (The)	16,23
10	IBM	International Business Machines Corporation	149,11
11	JPM	JPMorgan Chase & Co.	177,41
12	JWN	Nordstrom Incorporated	15,03
13	LEH	Lehman Brothers Holdings Inc.	43,46
14	LEN	Lennar Corporation	6,74
15	MAR	Marriott International, Inc.	19,51
16	MCD	McDonald's Corporation	56,05
17	SBC	AT&T Inc.	233,83
18	TXT	Textron Financial Corporation	12,21

* Company was delisted on December, 9th 2004. This fact does not affect in any way our results.

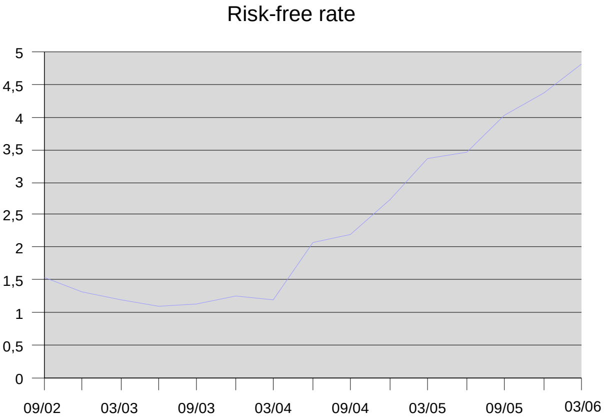


Figure 5. Risk free rate during our study (*Source: Federal Reserve System*)

Table 2. Recovery rates (*Source: Altman and Kishore* (1996))

Economic sector	Recovery rate
Hotel chains	0,26
Department stores	0,33
Finance	0,36
Telecommunications	0,37
Constructions	0,39
Metal and mechanic	0,42
Food	0,45

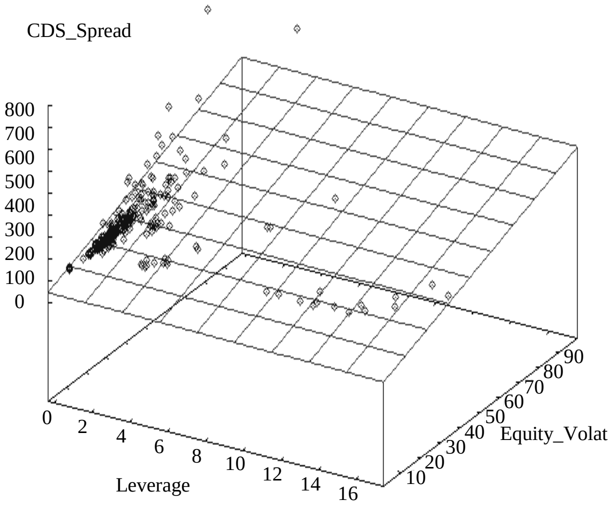


Figure 6. Relationship between CDS Spread, Leverage and Equity volatility in our sample (*Source: our elaborations)*

$$r_f = i - sp$$

showing an inverse relationship between *sp* and *rf*, confirmed by market data. We have the following correlation values:

Source: our elaborations

Variable	Correlation with CDS Spread
Risk-free (Rf)	-0,2187
Recovery rate (R)	-0,1475
Leverage (L)	-0,0485
Equity volatility (V)	0,6338

Of course we can notice a negative correlation with *R* (the recovery rate) and a strong positive correlation with *V* (the implied volatility which in our study proves to be very effective in predicting creditworthiness deterioration). The absence of a correlation with the leverage should not seem strange: our sample in fact includes financial companies too, which typically have a very high gearing ratio and a low CDS spread due to prudential regulation: in any case the neural network can solve this problem very well because of its nonparametric capabilities. Without considering the financial firms, the correlation of leverage and credit spreads would rise to 0.317.

The sample is made up of companies coming from different economic sectors, as it is easy to catch reading the recovery rates applied: of course we consider only big (or at least medium)-caps, the only ones for which a liquid market for CDS exists. In Figure 6 we show the relationship between CDS spread, Leverage and Equity volatility. It is evident that there is no linear relation between them. Moreover, only a few data are characterized by a leverage of more than 2: of course these can only be banks, which for prudential regulation can have a high gearing ratio. In the following part we will show how neural networks are able to price both industrial and financial firms at the same time, even if they show a strongly different leverage.

We have used a feed forward neural network, with the back propagation algorithm; it is a 4-layer network, with two hidden layers and therefore an output layer of only one node (the CDS spread).

The input layer consists of 18 nodes: in the first four nodes we have the risk free rate, the recovery rate, the leverage and the implied volatility of the firm; in the remaining 14 nodes we have the series of quarterly CDS spreads of the firm. If there is a lack in the data, we just use the value of the preceding quarter. This approach merges data coming from the firm with data (the CDS spreads) coming from the market, giving great effectiveness to the predictions of the network. Moreover the power of this approach can be appreciated observing that in this way the network is able to price CDS with reference entities coming both from the industrial field (which usually have low leverages and high CDS spreads) and from the financial field (which have an extremely high gearing ratio but are characterized by a history of low CDS spreads because of the prudential regulation, using this detail to discriminate between them). Figure 7 shows the structure of the network. The sample has of course been shuffled; the learning parameter has been settled to 0.5 and the initial parameters of the neurons have been chosen in the range [−2,2]. Our study shows that a logarithmic reduction is more efficient, because our sample consists of extremely variable data, so a simple linear reduction would enhance the distortions brought by the so- called outliers, that is data very different from the rest of the sample.

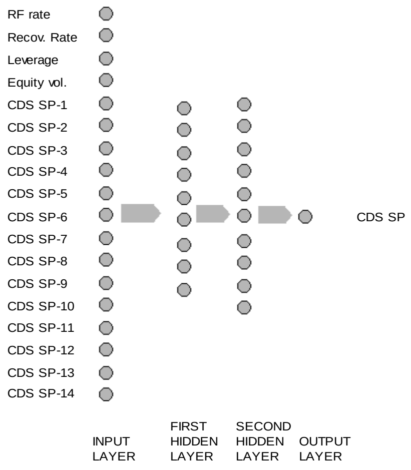


Figure 7. Structure of the neural network (*Source: our elaborations*)

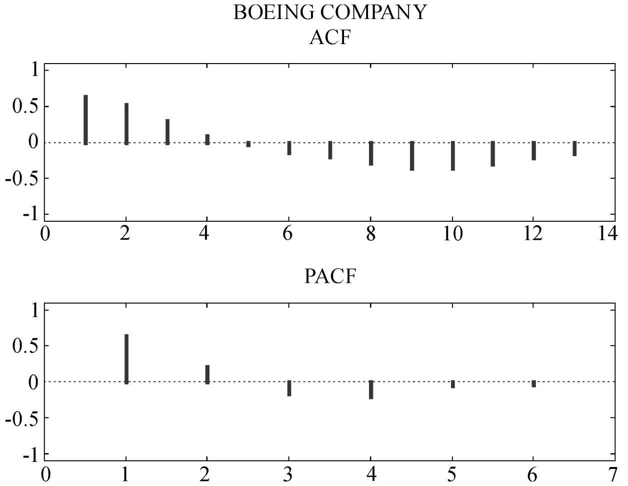


Figure 8. Typical correlogram of a CDS spread time serie (*Source: our elaborations*)

In Figure 8 we show as an example the correlogram for the CDS spread time series of The Boeing Company only, for the sake of simplicity, but we obtained the same structure for all the companies included in our sample: in the first part we can see the correlation between each value and a delayed value (the delay being expressed on the x-axis); the second part shows the correlation between each value and *p* preceding values, with *p* on the x-axis. It is therefore evident that the correlation between values, even if decreasing, is strong, so the series is autoregressive; we can then express each value in terms of the preceding ones. In this sense a CDS spread is more similar to an interest rate than to an equity price, so that it shows a mean reversion process which tends to pull spreads higher (lower) than some long-run average level back to this value over time. Obviously we shall have a negative (positive) drift. The sinusoidal cycle observable in the correlogram explains this phenomenon: moreover, it is a consequence of the strict relationship between CDS spreads and risk-free interest rates already discussed [25].

Figure 9 showing in red the neural network predictions and in yellow the real market data, confirms the effectiveness of the neural network in predicting CDS spreads.

In Table 3 and 4 the values of *R-squared* and *Rmse* are shown: as it is easy to observe, the results are highly coherent. We compare the results from or implementation with another model: Creditgrades™. We must stress the point that using traditional models such as Creditgrades™ we would obtain predictions almost useless, even excluding banks from the sample; neural networks surely are a great pricing instrument in order to evaluate credit spreads. The architecture of the neural network is feed forward, trained for 17000 learning epochs using the back propagation algorithm. Therefore it turns out obvious that neural networks are able to totally capture the variability relative to the market dynamics of credit derivatives: because of the fact that in literature there is no unanimity on the determination of the form of the CDS spread evaluation function, neural networks can therefore be seen as effective instruments of elaboration able to satisfy this lack from a statistical point of view.

Figure 10 shows a “delta” for a CDS contract: in fact we find on the x-axis the leverage, and on the y-axis the values calculated with the finite differences method, that is:

$$\Delta = \lim_{h \to 0} \frac{SP(lev+h) - SP(lev)}{h}$$

In a similar manner we can calculate for a CDS all the “greek” letters typical of derivative contracts using the outputs of the neural network with h-10^{-6}. It is evident in

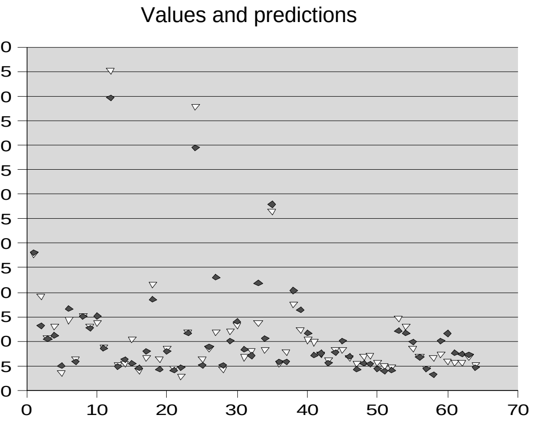


Figure 9. Market data (in yellow) and predictions of the neural network (in red) (*Source: our elaborations*)

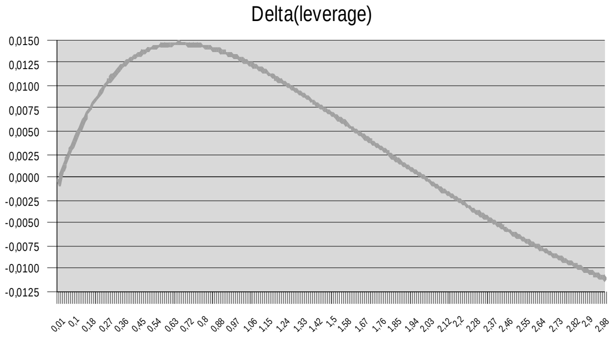


Figure 10. Relationship between delta and leverage (*Source: our elaborations*)

Table 3. Approximation of the neural network (*Source: our elaborations*)

Error	Value
R-squared	0,9082
Root mean squared error	14,3988

Table 4. Comparing statistical results (*Source: our elaborations*)

	NN	Credit Grades	Linear regression
Correlation	0,9636	−0,02	0,9309
Rmse	14,3988	>100	30,86
R-square	0,9086	>1	0,8566

Figure 10 shows a “delta” for a CDS contract: in fact we find on the x-axis the leverage, and on the y-axis the values calculated with the finite differences method, that is:

$$\Delta = \lim_{h \to 0} \frac{SP(lev+h) - SP(lev)}{h} \tag{20}$$

In a similar manner we can calculate for a CDS all the “greek” letters typical of derivative contracts using the outputs of the neural network with h-10^{-6}. It is evident in the diagram that for high leverages “delta” becomes negative: in fact we must remember that highly leveraged companies belong usually to the financial sector, so that they are less risky because of the prudential regulation. This effect is explained very well by the network, in fact for low leverages (typical of the industrial field) we see a direct relationship between leverage and CDS spreads. In other words, the neural network is able to recognize the risk of the activity carried out by the company using the time series of its CDS spread: in the part of our study covering the correlation, we obtained an average value for each observation and the preceding one of 0.90, as it is evident from the correlogram shown above. This correlation, along with the part regarding the independent variables, typical of the structural approach, explains the major part of the variability of CDS spreads.

6. Conclusions and Future Work

In this paper we have discussed an innovative approach to the study of CDS valuation, using neural networks. Our analysis is based on modeling the underlying dynamics of interest rates and firm characteristics and deriving the default probability based on these dynamics (the structural approach).

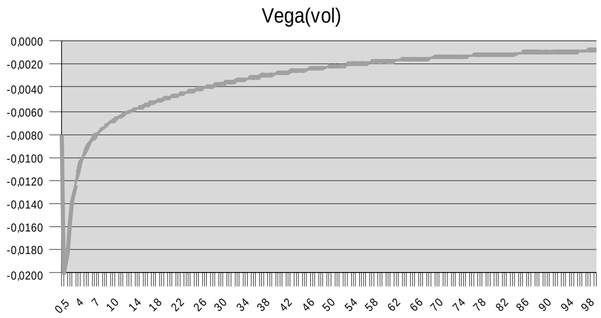


Figure 11. Relationship between vega and equity volatility *(Source: our elaborations)*

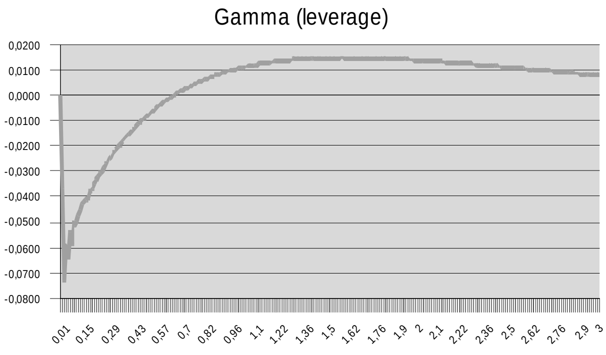


Figure 12. Relationship between gamma and leverage (*Source: our elaborations*)

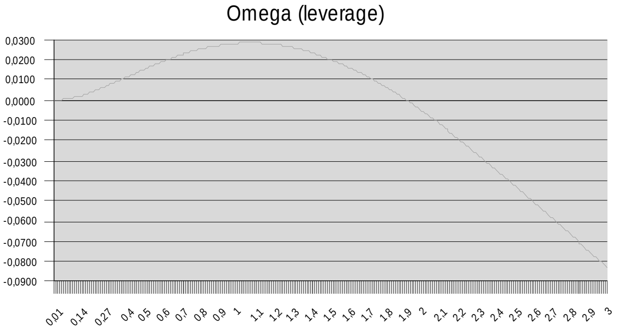


Figure 13. Relationship between omega and leverage (*Source: our elaborations*)

The model that we propose is peculiar for the use of the implied volatility of one-year options written on the shares of the analyzed companies, instead of historical volatility. Besides, the model differs from the structural approach for the fact that it considers the 30-month historical series for CDS spreads, including additional market variables. This implementation allows to use a forward-looking model and to capture the dynamic behavior of CDS spreads and equity volatility. This approach merges data coming from the firm with data (the CDS spreads) coming from the market, giving great effectiveness to the predictions of the neural network. Moreover, the power of this model can be appreciated observing that in this way the network is able to price CDS with reference entities coming both from the industrial field (which usually have low leverages and high CDS spreads) and from the financial field (which have an extremely high gearing ratio but are characterized by a history of low CDS spreads because of the prudential regulation, using this detail to discriminate between them).

We find that the neural network technique is useful for analyzing the pricing of a credit default swap. Our model produces a much lower forecasting error than those traditional models, such as Creditgrades^{TM}, indicating a relatively high precision in the neural network prediction. In particular, in the last part, starting from the high correlation observed between each CDS spread value and the preceding one in the time series of each company, we have trained a neural network based both on these time series and on the structural details of the firms, that is leverage, option-implied equity volatility and recovery rates. Our results in terms of *R-squared* and *Rmse* are highly coherent and are confirmed by the empirical data.

Our analysis presents the results that we have achieved and shows that the neural network model offers an alternative to traditional methodologies to deal with complicated issues related to CDS valuation.

Anyway, in this period, the CDS market is particularly volatile. The impact on the economy of the deflating housing bubble, the credit crisis in general, have stoked fear about increasing corporate defaults. This crisis is about credit risk. A credit bubble has ballooned for years, being enhanced by the existence of CDS. As credit originators can pass their risk to other agents, they have been less careful about the quality of their loans. In that sense, CDS have given an incentive for distributing more credit to more risky borrowers. As banks and all financial institutions and companies have committed themselves in the CDS market, they are now highly dependent on market continuity and on its smooth functioning. The failure of a major participant (bankruptcies of Bear Sterns, then those of AIG and Lehman Brothers) can put at stake all the others; the faith in the reliability of the market has been deeply shaken by these events.

In any case, some aspects of the proposed evaluation methodology require additional research: the possible next step for the research community is to improve the models in the case of catastrophic circumstances (the so-called LFHI (low frequency-high impact) events); another interesting case of study would regard the analysis of the recent financial crisis when more reliable information regarding financial companies will be available.

REFERENCES

[1] R. C. Merton, “On the pricing of corporate debt: The risk structure of interest rate,” The Journal of Finance, 29 1974.

[2] S Henke, H. P. Burghof, and B. Rudolph, “Credit securitization and credit derivatives: Financial instruments and the credit risk management of middle market com-mercial loan portfolios”, CFS Working paper Nr, July 1998.

[3] A. Greenspan, “Economic flexibility,” Speech to HM Treasury Enterprise Conference, London, UK, 2004.

[4] S. DAS, “Credit derivatives: Trading & Management of Credit & Default Risk,” John Wiley & Sons, Chicago, 1998.

[5] J. M. Tavakoli, “Credit derivatives: A guide to instruments and applications,” John Wiley & Sons, Chicago, 1998.

[6] G. R. Duffee and C. Zhou, “Credit derivatives in banking: useful tools for managing risk?” Journal of Monetary Economics, No. 48, 2001.

[7] R. Stultz, “Risk management and derivatives,” South-Western Publishing, 2003.

[8] B. A. Minton, R. Stultz, and R.Williamson, “How much do bank use credit derivatives to reduce risk?” Working Papers, 2005.

[9] Bank for international settlement, “International convergence of capital measurement and capital standards,” Basel Committee on Banking Supervision, A Revised Framework, Update November 2005.

[10] F. Dolcino, C. Giannini, and Rossi, E, “Reti neurali artificiali per l’analisi e la previsione di serie finanziarie,” Collana studi del Credito Italiano, 1998.

[11] D. Floreano and S. Nolfi, “Reti neurali: algoritmi di apprendimento, ambiente di apprendimento, architettura,” in Giornale Italiano di Psicologia, a. XX, pp. 15−50, febbraio 1993.

[12] M. Gori, “Introduzione alle reti neurali artificiali,” in Mondo Digitale n. 4, AICA, settembre 2003.

[13] C. Wu and C. H.Yu, “Risk aversion and the yield of corporate debt,” in Journal of Banking and Finance, No. 20, 1996.

[14] C. Wu, “A certainty equivalent approach to municipal bond default risk estimation,” in Journal of Financial Research, 1991.

[15] P. D. Mcnelis, “Neural networks in finance,” Elsevier Academic Press, 2005.

[16] A. Beltratti, M. Serio, and P. Terna, “Neural networks for economic and financial modelling,” International Thomson Computer Press, 1996.

[17] S. Hykin, “Neural networks: A comprehensive foundation,” Prentice Hall International, 1999.

[18] P. Werbos, “Backpropagation, past and future,” in Proceedings of the IEEE International conference on neural networks, IEEE press, 1988.

[19] F. Black and J. Cox, “Valuing corporate securities: Some effects of bond indenture provisions,” Journal of Finance, pp. 31, 1976.

[20] H. E. Lelan and K. B. Toft, “Optimal capital structure, endogenous bankruptcy, and the term structure of credit spreads,” The Journal of Finance, pp. 51, 1996.

[21] Collin dufresne and P. R. Goldstein, “Do credit spreads reflect stationary leverage ratios,” Journal of Finance, pp. 52, 2001.

[22] R. A. Jarrow and S. M. Turnbull, “Pricing derivatives on financial securities subject to credit risk,” The Journal of Finance, pp. 50, 1995.

[23] R. Jarrow, D. Lando, and S. Turnbull, “A markov model for the term structure of credit spreads,” Review of Financial Studies, pp. 10, 1997.

[24] D. Duffie and K. J. Singleton, (1998), “Modelling term structures of defaultable bonds,” Review of Financial Studies, pp. 12, 1999.

[25] J. C. Hull, “Opzioni, futures e altri derivati,” Il Sole 24Ore S. p. A., 2003.

(Edited by Vivian and Ann)

J. Serv. Sci. & Management, **2009, 2: 29-35**
Published Online March 2009 in SciRes (www.SciRP.org/journal/jssm)

Analysis of Service Processes Characteristics across a Range of Enterprises

John Maleyeff[1]

[1]Rensselaer Polytechnic Institute, Hartford Campus.
Email: maleyj@rpi.edu

Received November 10th, 2008; revised January 15th, 2009; accepted February 13th, 2009.

ABSTRACT

The structure of services processes was explored using a database of 168 *service processes that existed within a wide range of enterprises. The results indicate that applications within service science are not limited to the service industry and that service processes have many similar characteristics. The similarities exist across industry sectors* (*i.e., manufacturing, service*), *customer types* (*i.e., internal, external*) *and enterprise size* (*large, SME*). *A few differences exist and their importance is discussed. It is suggested that an important field within the multidisciplinary umbrella of service science is organizational behavior.*

Keywords: *service science, internal services, service marketing*

1. Introduction

Service delivery dominates activities performed by the workforce in the United States and other developed countries. For example, in July 2008, 79% of the U.S. workforce was employed in a service enterprise that was classified as one of the following: retail, government, education, health services, professional, business services, hospitality, leisure, or other services. The remaining 21% of the U.S. workforce was employed in enterprises classified as farming, manufacturing, construction, or other goods producing [1]. But, of these “goods producing” workers, a significant number are also engaged in service delivery. For example, many workers within manufacturing enterprises provide support or aftermarket services to users of their products (e.g., training, troubleshooting, or maintenance). Further, and perhaps more significantly, “goods producing” workers include internal support providers that deliver value to customers inside the firm. These workers are positioned in various departments such as finance, marketing, engineering, human resources, or information technology.

“Service Science” is an emerging academic discipline created in response to the need for organizations (busi-++ness, industry, non-profits, government, healthcare, etc.) to better understand how to create, manage, and improve services for the benefit of consumers, internal entities, and external partners [2]. The research reported in this article was an effort to contribute to the field of service science by studying the structure of services from a process-oriented perspective derived from lean management principles. Using a database created by the analysis of 168 service processes, the goal was to determine the similarities and differences among service processes that: 1) offer various types of services, 2) deliver value to internal versus external customers, 3) exist within a large versus a small or medium enterprise (SME), 4) consist of transformations that are informational versus not informational, and 5) exist within manufacturing versus service industries.

The results of this research should be useful to managers in both manufacturing and service enterprises, and academic researchers who are concerned with service improvement and service innovation. In the remainder of this article, background is provided that includes results from prior research and recent publications that address a variety of topics within service science, including the classification of services, the management of services for internal customers, and the analysis of service process characteristics. After the research methodology is explained and the resulting data are tabulated, results are described. A discussion follows that places the results in a context appropriate for management decision makers. The paper concludes with recommendations on future research directions.

2. Background

Although this research does not set out to create a new classification of services, similar services are combined for purpose of analysis and therefore a review of the relevant literature is warranted. Perhaps the most popular classification scheme was offered by Schmenner [3]. This classification separates services into four types based on two characteristics: 1) the level of customer interaction and customization, and 2) the degree of labor intensity. The resulting classification includes the following four sets of service processes (the level of interaction & customization, and the level of labor intensity is indicated for each): the service factory (low, low), the service shop (high, low), the mass service (low, high), and the professional service (high, high). Fitzsimmons and Fitzsimmons [4] provide a set of challenges that would need to be addressed by managers within each class.

Examples of other classification schemes and related efforts to provide typologies for services have focused on the level of direct customer contact [5], the amount of customer involvement [6], customers' perceptions of services [7], and the amount of customization in the service output [8]. A thorough list of articles that address the classification of services is provided by Cunningham *et al.* [7].

Whether or not any of the various service classification schemes have enhanced the management of services is an open question. For example, Verma [9] shows that only 4 of 22 important management challenges are affected by the differences in Schmenner's classification scheme. Then again, Silvestro *et al.* [10] argued that service strategy, control, and performance measurement would differ for professional services, service shops, and mass services. The transition from tangible goods to intangible equivalents, such as maps, videotapes, and newspapers, has also impacted the usefulness of traditional classification schemes [11].

Because the majority of the service processes studied in this research would be classified as an internal service, a review of the relevant literature is warranted. Davis [12] defined internal service operations as "behind-the-scenes routines, procedures, and activities that provide the necessary support to the company's more visible functions." With effects that are often hidden from the view of senior managers, internal services are often the first to be affected by downsizing or outsourcing [13]. The fact that internal service departments sometimes display an attitude that suggests superiority or independence can make them unsympathetic victims within the corporate structure [14].

Research has shown that external customer satisfaction is enhanced by improved internal customer satisfaction [15]. It has been suggested, however, that many organizations are not equipped to understand how to deal with the challenges associated with internal service management [16]. Without a common understanding of how an internal service operates, mistakes are common. For example, technology is often implemented without an understanding of the associated implications [17]. Similarly, an accountant may create budgets that motivate suboptimal behavior due to arbitrary cost allocation schemes [18]. Johnston [15] argues that inadequate attention on internal services. He reported that, in the three major service journals between 1996 and 2006, only 8% of articles dealt with research into internal services.

The majority of the services analyzed in this research would be classified as a professional service based on Schmenner's scheme. But, it has been suggested that little agreement exists regarding the definition of a professional service [19]. A definition suggested by Harte and Dale [20], who define a professional service as consisting of "intangible outputs, with qualitative rather than quantitative criteria being the main measures for customer satisfaction, high buyer-interaction levels and lack of heterogeneity," would appear to characterize professional services studied in this research. Professional services are also commonly associated with characteristics such as "specialist knowledge, autonomy, altruism, self-regulation, and a high degree of participation and customization" [21].

Laing and Lian [22] suggest a classification of professional services based on the level of inter-organizational relationships, ranging from almost transactional to a fully integrated. Hausman [23] showed that customer relationships were more important than the professional competence of service providers. Similarly, Day and Barksdale [24] suggest that service providers' understanding of client needs and their communication skills are the main determinants of quality for clients of architectural and engineering firms. And, Ojasalo [25] provides a list of ten characteristics of a professional service based on an extensive literature review. Characteristics such as "a high degree of customer uncertainty" and "affected by characteristics of information", as well as "a problem-solving approach" are notably present in the list. Finally, various mechanisms that weaken customer relationships in professional services have been studied by Åkerlund [26].

In this research, the approach to organizing service processes to explore their underlying characteristics made use of lean management principles [27]. In particular, each transaction within the process is described as being value-added (a task that the customer cannot do or wishes not to do) or non-value-added (other tasks or activities such as inspecting work, moving documents from one department not to another, or various forms of delays). All non-value-added activities would be inherently wasteful, although some may be necessary in the short term due to the structure of the service process (e.g., a delay caused by moving documents from one department to another is necessary if the departments are not co-located).

The use of a lean management approach is motivated by a desire to organize service processes so that groups are created that are likely to make use of similar improvement or innovation approaches. To make an analogy to manufacturing, an effort to reduce the setup time for a drilling process may not be concerned with the overall volume of production. Similarly, in a service, an effort to reduce errors during an information handoff may not be concerned with the whether or not the service offering was standard or customized. An example that illustrates the benefits of this approach is a hospital trauma team that learned how to improve the treatment of emergency patients by studying pit crews at automobile races [28].

A qualitative study concluded that services delivered by organizations whose customers were other businesses were similar in structure to those services delivered by organizations whose customers were consumers [15]. But important differences have been reported between services for internal customers and those for external customers. These differences include the lack of choice provided to internal customers [29], limited empathy between service providers and internal customers [30], and inter-departmental dynamics that often lead to misunderstandings of priorities [31].

3. Methodology

A total of 168 service systems were included in this study. Each service system was analyzed by a professional employee of the organization who was very familiar with the activities associated with the delivery of the service and had access to customers of the service. Most of the services were primarily for internal customers, but many served primarily external customers, and some served both internal and external customers in about equal measure. No single analyst studied more than one service. All of the analysts were enrolled in a part-time graduate management program on the Hartford, Connecticut campus of Rensselaer Polytechnic Institute, in a three-credit course called Service Operations Management.

The 168 service systems did not constitute a random sample nor were they carefully selected in a controlled experiment. However, the range of firms represented and the services chosen was broad, albeit biased due to the disproportionate number of scientists and engineers in the student body. Sixty-three different enterprises were represented. One large corporation dominated the group, constituting 52 of the 168 services.

For each service system, the analyst was asked to perform a comprehensive study of its structure (by creating a process map or flowchart to illustrate how the various activities interact to provide the service), identify customers as either internal or external (or both), ask several customers to list strengths and weaknesses of the process, and list key performance dimensions important to customers. The resulting reports followed a standard template that allowed for easy tabulation of key results.

A database was created to capture important data related to each service process. It includes, for each service: the name of the enterprise within which the service took place (these data were not available for 8 services), the size of the enterprise (classified as a large enterprise or a SME), the type of enterprise-manufacturing or service (these data were not available for 10 services), a brief description of the process, the number of employees directly involved with service delivery (these data were not available for 104 services), the number of departments or functions directly involved with service delivery, the primary type of customer (classified as internal, external, or both), and an indication of whether or not information was the key service transformation. The data were placed into a MINITAB worksheet in preparation for tabulation and statistical analysis.

4. Service Process Types

While studying the 168 reports, it became apparent that a finite number of specific types of value-added activities took place, most involving informational transformations. While listing these activities, it was clear that many seemingly dissimilar service possesses consisted of similar sets of transformations (e.g., an audit to determine if a worker is following standard protocol and the testing of a material to determine if it meets specifications both involve evaluation of actual performance and comparison to a standard).

An exhaustive qualitative analysis of the 168 service processes resulted in the classification of six service process types. This set should not be considered comprehensive because the sample of services was not random. It would, however, serve to create an effective analysis structure for this research. Table 1 shows the six service process types, along with examples of each type and the number of occurrences of each type in the database.

Most of the service process types would be classified as a professional service using Schmenner's classification scheme, because they have high levels of both customization and customer contact. In most cases, however, employees delivering a service did not hold strong allegiance to their professions as would, for example, lawyers or physicians. One type that would not always operate as a professional service would be "gathering," the collection and reporting of information that is often disseminated to a wide variety of customers and not always customized for each customer's use. In these cases, the service would be classified as a mass service. A few other examples of services that would not be classified as a professional service would be found in each type.

Table 1. Service process types

Type	Description	Examples	No.
Trouble-shooting	Solves a customer's problem	IT help desk Parts return Root cause investigation Complains handling	26
Gathering (and subsequent documenting)	Provides instructions or summarizes information for use by others	Installation instructions Maintenance guidelines Accounting statements Accident reporting Environmental Compliance	21
Evaluation	Determine whether or not a specification or a standard is met	Auditing Design change Laboratory testing Part inspection Bill payment	38
Analysis	Determine if resources should be allocated for a requested purpose	Proposal writing Sales quoting Data analysis New business analysis	20
Planning	Planning, tracking, and controlling projects and other activities	Software integration Project management Metric tracking Employee orientation Recruitment	40
Consultation	Provide specific expertise to assist customers	Tool design Forecasting Software development Supplier selection Logistic support	23

Table 2. Summary of results by service type

Service Type (No.)	Internal Only	External Only	Internal & External	Average # Functions	Median # Employees	Information Delivery
Troubleshooting (26)	38%	46%	16%	4.6	10.0	85%
Gathering (21)	86%	10%	4%	4.5	10.0	100%
Evaluation (38)	79%	13%	8%	4.8	16.0	100%
Analysis (20)	70%	20%	10%	5.9	14.0	100%
Planning (40)	65%	20%	15%	5.3	20.0	83%
Consultation (23)	56%	30%	14%	5.0	8.0	83%
Overall (168)	**66%**	**23%**	**11%**	**5.0**	**12.0**	**91%**

5. Analysis & Results

Table 2 lists, by type, the percentage of services with primarily internal customers, the percentage of services with primarily external customers, the percentage of services for both internal and external customers, the average number of organizational functions (e.g., internal departments) directly involved with the delivery of the service, the median number of employees directly involved in the delivery of the service, and the percentage of services whose transformations were informational. For the results reported in this section, details on the statistical routines are included in the Appendix.

5.1 Most Services Served Internal Customers

Table 2 shows that about two-thirds of services had only internal customers and less than one-fourth of the services had only external customers. The prevalence of services for internal customers was relatively high for all service types, but there was a significant difference in their prevalence across service types (p=0.005). Specifically, the prevalence of services for internal customers was lower for troubleshooting services. With this category removed, no difference was evident across the service types in the prevalence of services for internal customers (p=0.169).

5.2 Services Consist of Inter-Departmental Process Flows

Table 2 shows that the number of functions (i.e., departments) involved directly with delivering the service averaged 5.0 functions. And, there was no significant difference in the number of functions across service types (p=0.684). Similarly, the median number of employees directly involved with delivering the service was 12.0, and there was no significantly difference in the number of employees across service types (p=0.745). Figure 1 provides the distribution of the number of functions that participate in delivering each service. It appears to be very likely that a service process will cross more than a few departmental lines within an enterprise.

5.3 Information Transformations were Dominant

Table 2 shows that a predominance of informational transformations took place, although some variation existed across service types (p=0.012). The gathering, evaluation, and analysis service types all consisted exclusively of services that provide information. But, well over 80% of services classified as troubleshooting, planning, and consultation also consisted of informational transformations. Examples of cases where information was not the main transformation included the coordination of part's receipt from vendors, the repair of a mechanical device, and the dispensing of drugs by a pharmacy. In all of these services, however, information was an important secondary output that needs to be managed effectively.

5.4 Services for Internal Customers are Similar to Services for External Customers

Table 3 shows that, when comparing services meant for internal customers with those meant for external customers, no differences were found in three key characteristics. First, there was no difference in the number of functions involved in service delivery (p=0.470). Second, there was no difference in the number of employees involved in service delivery (p=0.653). And third, there was no difference in the prevalence of information related services (p=0.522).

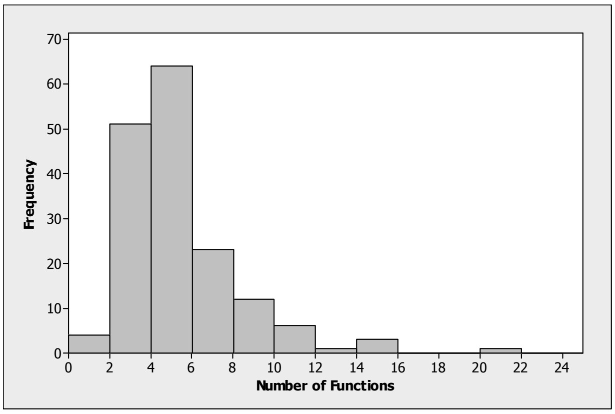


Figure 1. Distribution for number of functions delivering a service

Table 3. Summary of results by customer

Customer (No.)	Average # Functions	Median # Employees	Information Delivery
Internal (111)	5.2	14.0	93%
External (38)	4.6	10.0	87%
Both (19)	4.7	12.0	89%
Overall (168)	**5.0**	**12.0**	**91%**

5.5 Manufacturing and Service Enterprises Provide Similar Services

When comparing services found in manufacturing enterprises with those found in service enterprises, the mix of service types was similar (p=0.663). Table 4 shows that no differences were evident in four key characteristics. First, there was no difference in the mix of customers (p=0.258). Second, there was no difference in the number of functions involved in service delivery (p=0.124). Third, there was no difference in the number of employees involved in service delivery (p=0.344). And fourth, there was no difference in the prevalence of information related services (p=0.692).

5.6 Large Enterprises and SME's Provide Services with Some Differences

When comparing services found in large enterprises with those found in a SME, the mix of service types within the enterprises was similar (p=0.167). Table 5 shows that services found within large enterprises were more likely to have primarily internal customers (p=0.003). But, when comparing services in a large enterprise to services in a SME, no differences were evident in three other key characteristics. First, there was no difference in the number of functions involved in service delivery (p=0.329). Second, there was no difference in the number of employees involved in service delivery (p=0.228). And third, there was no difference in the prevalence of information related services (p=0.756).

The results of the analysis of enterprise size were repeated when analyzing data for the large corporation that was disproportionately represented in the sample of services, with one exception. This exception was that, within this corporation, more employees were involved with the delivery of a service (p=0.001). Specifically, the median number of employees delivering the service was 40 versus a median of 10 for other organizations. This result may be of interest, because the large corporation operates with a rigorous "standard work" policy that could result in tasks that were easily performed by more than a select few individuals.

5.7 Services with Information Transformations May be Similar to Other Services

Table 6 shows that, when comparing the many services that consisted of an informational transformation with the few services that consisted of another type of transformation, no differences were evident in the number of functions involved in service delivery or in the number of employees involved in service delivery. These results should not be considered conclusive, because only 15 of the services involving deliverables other than information.

6. Discussion

The study of services, and in particular the field of service science, may have greater relevance than conventional wisdom would dictate. For example, the results detailed above have implications for managers of both manufacturing and services enterprises because few critical differences exist in services found within manufacturing and service enterprises. Perhaps Albrecht [30] was correct in suggesting that the manufacturing-service distinction is becoming blurred and that "the only real distinction anymore is the relative proportion of tangible and intangible value sold and delivered." In addition, services delivered to either internal or external customers, as well as those found in any size organization, possess more similarities than differences.

Table 4. Summary of results by enterprise focus

Enterprise Focus (No.)	Internal Only	External Only	Internal & External	Average # Functions	Median # Employees	Information Delivery
Manufacturing (87)	71%	17%	12%	5.3	14.5	92%
Services (71)	62%	28%	10%	4.7	10.0	90%
Overall (158)	**66%**	**23%**	**11%**	**5.0**	**12.0**	**91%**

Table 5. Summary of results by enterprise size

Enterprise Size (No.)	Internal Only	External Only	Internal & External	Average # Functions	Median # Employees	Information Delivery
Large (118)	74%	16%	10%	5.2	15.0	92%
Small/Medium (40)	48%	40%	12%	4.8	8.0	90%
Overall (158)	**66%**	**23%**	**11%**	**5.0**	**12.0**	**91%**

Table 6. Summary of results by transformation

Transformation (No.)	Internal Only	External Only	Internal & External	Average # Functions	Median # Employees
Informational (153)	67%	22%	11%	5.0	12.0
Other (15)	53%	33%	14%	4.8	5.0
Overall (168)	**66%**	**23%**	**11%**	**5.0**	**12.0**

The results also provide some insight into special organizationally-based challenges in service improvement and service innovation. Given the average of 5 functions per service process, it is likely that change efforts would be hampered by ownership confusion, lack of commitment, competing reward systems, and other organizational barriers. Further, an individual manager's motivation to improve a service may be affected by the relatively few employees within each department that take part in the delivery of each service that flows through that department (roughly 2 employees per department). Strong leadership is necessary to overcome organizational barriers and bring cross-functional teams together for improving processes.

The predominance of information transformations in all service types is an important aspect of service improvement and innovation. The importance of information transformations in internal services has been noted previously by Maleyeff [32]. He also offered suggestions on the types of actions that managers should take, including a recommendation to focus improvement efforts on controlling the important information rather than the physical manifestations of information (documents, blueprints, and other tangible forms of service output). The ability to understand and control information flow would appear to be an important skill for managers of any service.

7. Conclusions & Future Work

It would be a mistake to consider the applications within service science to be limited to the service industry. Service processes have similar characteristics, regardless of whether they exist within manufacturing enterprises or service enterprises, and regardless of whether or not the customer is internal or external. Further, with the confirmation that service processes can be expected to flow through more than a few departments within an enterprise, perhaps the most important field within the multidisciplinary umbrella of service science is organizational behavior. It appears that service processes share a number of common characteristics that should interest researchers and practitioners in this field.

Many suggestions may be offered for extending this research. An improved service classification scheme, specifically designed to compliment service improvement and innovation efforts, may be useful. A more thorough and far reaching analysis of the specific value-added tasks that make up service processes could lead to a better understanding of how to modularize efforts at improvement and innovation. That is, perhaps researchers can help find approaches to solve certain problems that have universal rather than local application. It would also be interesting to determine if the results found here would be repeated within a more robust sample of services. Finally, studies of how customer satisfaction is affected by service process characteristics would be helpful. For example, for the large corporation that disproportionately represented the sample studied in this research, does their "standard work" policy translate to higher levels of satisfaction compared with similar enterprises that allow for more flexibility in service delivery?

REFERENCES

[1] Bureau of Labor Statistics, U. S. department of Labor Newsletter, USDL 08−1049, 2008.

[2] R. C. Larson, "Service science: At the intersection of management, social, and engineering sciences," IBM Systems Journal, 47(1), pp. 41−51, 2008.

[3] Schmenner and W. Roger, "How can service businesses survive and prosper," Sloan Management Review, 27(3), pp. 21−32, 1986.

[4] J. A. Fitzsimmons and M. J. Fitzsimmons, "Service management," 5th Edition, McGraw-Hill, New York, USA, 2006.

[5] R. B. Chase, "The customer contact approach to services: Theoretical bases and practical extensions," Operations Research, 29(4), pp. 698−706, 1981.

[6] U. Wemmerlöv, "A taxonomy for service processes and its implications for system design," International Journal of Service Industry Management, 1(3), pp. 20−40, 1989.

[7] L. F. Cunningham, C. E. Young, W. Ulaga, and M. Lee, "Consumer views of service classification in the USA and France," Journal of Services Marketing, 18(6), pp. 421−432, 2004

[8] D. L. Kellogg and W. Nie, "A framework for strategic service management," Journal of Operations Management, 13, pp. 323−337, 1995.

[9] R. Verma, "An empirical analysis of management challenges in service factories, service shops, mass services, and professional services," International Journal of Service Industry Management, 11(1), pp. 8−25, 2000.

[10] R. Silvestro, L. Fitzgerald, and R. Johnston, "Towards a classification of service processes," International Journal of Service Industry Management, 3(3), pp. 62−75, 1992.

[11] P. Hill, "Tangibles, intangibles and services: A new taxonomy for the classification of output," Canadian Journal of Economics, 32(2), pp. 426−446, 1999.

[12] T. R. V. Davis, "Internal service operations: Strategies for increasing their effectiveness and controlling their cost," Organizational Dynamics, 20(2), pp. 5−22, 1991.

[13] C. R. Jones, "Customer satisfaction assessment for 'internal' suppliers," Managing Service Quality, 6(1), pp. 45−48, 1996.

[14] A. Wilson, "The internal service department-justifying your existence," Logistics Information Management, 11(1), pp. 58−61, 1998.

[15] D. D. Gremler, K. R. Evans, and M. J. Bitner, "The internals service encounter," International Journal of Service Industry Management, 5(2), pp. 34−56, 1994.

[16] R. Johnston, "Internal service-barriers, flows, and assessment," International Journal of Service Industry Management, 19(2), pp. 210−231, 2008.

[17] P. A. Smart, R. S. Maull, Z. J. Radnor, and T. J. Housel, “An approach for identifying value in business processes,” Journal of Knowledge Management, **7**(4), pp. 49−61, 2003.

[18] L. Kren, “Planning internal service department resources to avoid suboptimal behavior,” The CPA Journal, 78(1), pp. 54−57, 2008.

[19] M. V. Thakor and A. Kumar, “What is a professional service? A conceptual review and bi-national investigation,” The Journal of Services Marketing, 14(1), pp. 63−82, 2000.

[20] H. G. Harte and B. G. Dale, “Improving quality in professional service organizations: A review of the key issues,” Managing Service Quality, 5(3), pp. 34−44, 1995.

[21] E. Jaakkola and A. Halinen, “Problem solving within professional services: Evidence from the medical field,” International Journal of Service Industry Management, 17(5), pp. 409−429, 2006.

[22] A. W. Laing and P. C. S. Lian, “Inter-organizational relationships in professional services: Towards a typology of service relationships,” The Journal of Services Marketing, 19(2), pp. 114−127, 2005.

[23] A. V. Hausman, “Professional service relationships: A multi-context study of factors impacting satisfaction, re-patronization, and recommendations,” Journal of Services Marketing, 17(3), pp. 226−242, 2003.

[24] E. Day and H. C. Barksdale, “How firms select professional services,” Industrial Marketing Management, 21(2), pp. 85−91, 1992.

[25] J. Ojasalo, “Characteristics of professional services and managerial approaches for achieving quality excellence,” The Business Review, **7**(2), pp. 61−68, 2007.

[26] H. Åkerlund, “Fading customer relationships in professional services,” Managing Service Quality, 15(2), pp. 156−171. 2005.

[27] J. P. Womack and D. T. Jones, “Lean Thinking,” 2nd Edition. Free Press, New York, USA, 2003.

[28] G. Naik, “New formula: A hospital races to learn lessons of Ferrari pit stop,” Wall Street Journal, A1, November 14, 2006.

[29] P. J. A. Nagel and W. W. Cilliers, “Customer satisfaction: A comprehensive approach,” International Journal of Physical Distribution & Logistics Management, 20(6), pp. 2−46, 1990.

[30] K. Albrecht, “The service within,” McGraw-Hill, New York, NY, USA, 1990.

[31] S. Auty and G. Long, “‘Tribal warfare’ and gaps affecting internal service quality,” International Journal of Service Industry Management, 10(1), pp. 7−22, 1999.

[32] J. Maleyeff, “Exploration of internal service systems using lean principles,” Management Decision, 44(5), pp. 674−689, 2006.

Appendix

Basic statistical tools were incorporated using MINITAB statistical software and the resulting p-value is included in the discussion of results. A p-value represents the probability that random chance alone would have produced the effects found in the data. Traditionally, when a p-value is less than 0.05 (5%) the effect is said to be statistically significant.

For analyses to determine if a certain characteristic (e.g., enterprise size, service type) affected the number of functions involved directly in delivering the service, a one-way ANOVA was used. In all of the cases analyzed and reported in this article, homogeneity was confirmed and the resulting residuals were found to be normally distributed with a common variance. A transformation to the natural log of the number of functions was necessary to ensure normality of residuals.

Mood’s median test was used for analyses to determine if a certain characteristic (e.g., enterprise size, service type) affected the number of employees involved directly in delivering the service (a one-way ANOVA was not used because the distribution of the number of employees was highly skewed and a few outliers existed). For analyses to determine if a certain characteristic (e.g., enterprise size, service type) affected a binary variable (e.g., internal or external customer, informational or not informational), a two-sample hypothesis test for proportions was used. Chi-square hypothesis tests were used for analyses to determine if differences in the service types affected a certain binary variable (e.g., prevalence of internal customers, prevalence of informational transformations).

(Edited by Vivian and Ann)

J. Serv. Sci. & Management, **2009, 2: 36-42**
Published Online March 2009 in SciRes (www.SciRP.org/journal/jssm)

A Nonmonotone Line Search Method for Regression Analysis*

Gonglin Yuan[1], Zengxin Wei[1]

[1]College of Mathematics and Information Science, Guangxi University, Nanning, Guangxi, 530004, P. R. China.
Email: glyuan@gxu.edu.cn

Received January 7th, 2009; revised February 9th, 2009; accepted February 28th, 2009.

ABSTRACT

In this paper, we propose a nonmonotone line search combining with the search direction (*G. L. Yuan and Z. X.Wei, New Line Search Methods for Unconstrained Optimization, Journal of the Korean Statistical Society,* 38(2009), *pp.* 29-39.) *for regression problems. The global convergence of the given method will be established under suitable conditions. Numerical results show that the presented algorithm is more competitive than the normal methods.*

Keywords: *regression analysis, fitting method, optimization, nonmonotone, global convergence*

1. Introduction

It is well known that the regression analysis often arises in economies, finance, trade, law, meteorology, medicine, biology, chemistry, engineering, physics, education, history, sociology, psychology, and so on [1,2,3,4,5,6,7]. The classical regression model is defined by

$$Y=h(X_1, X_2, \ldots, X_p)+\varepsilon$$

where Y is the response variable, Xi is predictor variable, i=1,2, …, p, $p>0$ is an integer constant, and ε is the error. The function $h(X_1, X_2, \ldots, X_p)$ describes the relation between Y and $X=(X_1, X_2, \ldots, X_p)$. If h is linear function, then we can get the following linear regression model

$$Y=\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p+\varepsilon \qquad (1)$$

which is the most simple regression model, where β_0, β_1, …, β_p are regression parameters. On the other hand, the regression model is called nonlinear regression. We all know that there are many nonlinear regression could be linearization [8,9,10,11,12,13]. Then many authors are devoted to the linear model [14,15,16,17,18,19]. Now we will concentrate on the linear model to discuss the following problems. One of the most important work of the regress analysis is to estimate the parameters $\beta=(\beta_0,\beta_1,\cdots,\beta_p)$.

The least squares method is an important fitting method to determined the parameters $\beta=(\beta_0,\beta_1,\cdots,\beta_p)$, which is defined by

$$\min_{\beta\in\Re p+1} S(\beta)=\sum_{i=1}^{m}(h_i-(\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\beta_p X_{ip}))^2 \qquad (2)$$

where h_i is the data valuation of the ith response variable, X_{i1}, X_{i2}, …, X_{ip} are p data valuation of the ith predictor variable, and m is the number of the data. If the dimensionp and the number m is small, then we can obtain the parameters $\beta=(\beta_0,\beta_1,\cdots,\beta_p)$ from extreme value of calculus. From the definition (2), it is not difficult to see that this problem (2) is the same as the following unconstrained optimization problem

$$\min_{x\in\Re^n} f(x) \qquad (3)$$

In this paper, we will concentrate on this problem (3) where $f:\ \Re^n\to\ \Re$ is continuously differentiable (linear or nonlinear). For regression problem (3), if the dimension n is large and the function f is complex, then the method of extreme value of calculus will fail. In order to solve this problem, numerical methods are often used, such as steepest descent method, Newton method, and Guass-Newton method [5,6,7]. Numerical method, i.e., the iterative method is to generates a sequence of points $\{x_k\}$ which will terminate or converge to a point x^* in some sense. The line search method is one of the most effective numerical method, which is defined by

$$x_{k+1}=x_k+\alpha_k d_k, k=0,1,2,\cdots \qquad (4)$$

where α_k is determined by a line search is the steplength, and d_k which determines different line search methods [20,21,22,23,24,25,26,27] is a descent direction of f at x_k.

Due to its simplicity and its very low memory requirement, the conjugate gradient method is a powerful line search method for solving the large scale optimization problems. This method can avoid, like steepest de-

*This work is supported by China NSF grands 10761001 and the Scientific Research Foundation of Guangxi University (Grant No. X081082).

scent method, the computation and storage of some matrices associated with the Hessian of objective functions. Conjugate gradient method has the form

$$d_{k+1} = \begin{cases} -g_{k+1} + \beta_k d_{k,} & if \;\; k \ge 1 \\ -g_{k+1}, & if \;\; k = 0 \end{cases} \tag{5}$$

where $g_k = \nabla f(x_k)$ is the gradient of f(x) at x_k, $\beta_k \in \Re$ is a scalar which determines the different conjugate gradient method [28,29,30,31,32,33,34,35,36,37]. Throughout this paper, we denote $f(x_k)$ by f_k, $\nabla f(x_k)$ by g_k, and $\nabla f(x_{k+1})$ by g_{k+1}, respectively. $\|.\|$ denotes the Euclidian norm of vectors. However, the following sufficiently des cent condition which is very important to insure the global convergence of the optimization problems

$$g_k^T dk \le -c\|gk\|^2 \text{, } for \; all \; k \ge 0 \text{ and some constant } c > 0 \tag{6}$$

is difficult to be satisfied by nonlinear conjugate gradient method, and this condition may be crucial for conjugate gradient methods [38]. At present, the global convergence of the PRP conjugate gradient method is still open when the weak Wolfe-Powell line search rule is used. Considering this case, Yuan and Wei [27] proposed a new direction defined by

$$d_{k+1} = \begin{cases} -g_{k+1} + \dfrac{\| g_{k+1} \|^2}{-g_{k+1}^T d_k} d_k, if \; g_{k+1}^T d_k \ne 0 \\ -g_{k+1} \qquad\qquad otherwise \end{cases} \tag{6}$$

where $d_0 = -\nabla f_0 = -g_0$. If $d_k^T g_{k+1} \ne 0$, it is easy to see that the search direction d_k is the vector sum of the gradient $-g_k$ and the former search direction d_{k-1}, which is similar to conjugate gradient method. Otherwise, the steepest descent method is used as restart condition. Computational features should be effective. It is easy to see that the sufficiently descent condition (6) is true without carrying out any line search technique by this way. The global convergence has been established. Moreover, numerical results of the problems [39] and two regression analysis show that the given method is more competitive than the other similar methods [27].

Normally the steplength α_k is generated by the following weak Wolfe-Powell (WWP): Find a steplength α_k such that

$$f(x_k + \alpha_k d_k) \le f(x_k) + \sigma_1 \alpha_k g_k^T d_k \tag{8}$$

$$g_{k+1}^T d_k \ge \sigma_2 g_k^T d_k \tag{9}$$

where $0 < \sigma_1 < \sigma_2 < 1$. The monotone line search technique is often used to get the stepsize α_k, however monotonicity may cause a series of very small steps if the contours of objective function are a family of curves with large curvature [40]. More recently, the nonmonotonic line search for solving unconstrained optimization is proposed by Grippo *et al.* in [40,41,42] and further studied by [43,44] etc. Grippo, Lamparillo, and Lucidi [40] proposed the following nonmonotone line search that they call it GLL line search. GLL line search: Select steplength α_k satisfying

$$f_{k+1} \leqslant \max_{0 \le j \le n(k)} f_{k-j} + \varepsilon_1 \alpha_k d_k^T g_k \tag{10}$$

$$g_{k+1}^T d_k \ge \max\{\varepsilon_2, 1 - (\alpha_k \| d_k \|)^p\} g_k^T d_k \tag{11}$$

where $p \in (-\infty, 1)$, k = 0, 1, 2, …, $\varepsilon_1 \in (0,1), \varepsilon_2 \in (0, \frac{1}{2})$, $n(k) = \min\{H,k\}$, $H \geqslant 0$ is an integer constant. Combinng this line search and the normal BFGS formula, Han and Liu [45] established the global convergence of the convex objective function. Numerical results show that this method is more competitive to the normal BFGS method with WWP line search. Yuan and Wei [46] proved the superlinear convergence of the new nonmonotone BFGS algorithm.

Motivated by the above observations, we propose a nonmonotone method on the basic of Yuan and Wei [27] and Grippo, Lamparillo, and Lucidi [40]. The major contribution of this paper is an extension of the new direction in [27] to the nonmonotone line search scheme, and to concentrate on the regression analysis problems. Under suitable conditions, we establish the global convergence of the method. The numerical experiments of the proposed method on a set of problems indicate that it is interesting.

This paper is organized as follows. In the next section, the proposed algorithm is given. Under some reasonable conditions, the global convergence of the given method is established in Section 3. Numerical results and a conclusion are presented in Section 4 and in Section 5, respectively.

2. Algorithms

The proposed algorithm is given as follows.

Nonmonotone line search Algorithm (NLSA).

Step 0: Choose an initial point $x_0 \in \Re^n$, $0 < \varepsilon < 1$, $0 < \varepsilon_1 < \varepsilon_2 < 1$, $p \in (-\infty, 1)$. an integer constant H>0. Set $d_0 = -\nabla f_0 = -g_0$, k :=0;

Step 1: If $\| g_k \|_2 \leqslant \varepsilon$, then stop; Otherwise go to step 2;

Step 2: Compute steplength α_k by Wolfe line search (10) and (11), let $x_{k+1} = x_k + \alpha_k d_k$.

Step 3: Calculate the search direction d_{k+1} by (7).

Step 4: Set k: $=k$+1 and go to step 1.

Yuan and Wei [27] also presented two algorithms; here we stated them as follows. First another line search is given [47]: find a steplength α_k satisfying

$$f(x_k + \alpha_k d_k) \le C_k + \sigma_1 \alpha_k g_k^T d_k \tag{12}$$

$$g_{k+1}^T d_k \ge \sigma_2 g_k^T d_k \tag{13}$$

where $0 < \sigma_1 < \sigma_2 < 1$,

$$C_{k+1} = \frac{\mu_k Q_k C_k + f_{k+1}}{Q_{k+1}}, Q_{k+1} = \mu_k Q_k + 1,$$

$$C_0 = f_0, Q_0 = 1, \mu_k \in [\mu_{\min}, \mu_{\max}], 0 \le \mu_{\min} \le \mu_{\max} \le 1$$

Algorithm 1 [27].

Step 0: Choose an initial point $x_0 \in \Re^n$, $0 < \varepsilon < 1$, $0 < \sigma_1 < \sigma_2 < 1$. Set $d_0 = -\nabla f_0 = -g_0$, $k := 0$;

Step 1: If $\| g_k \|_2 \le \varepsilon$, then stop; Otherwise go to step 2;

Step 2: Compute steplength α_k by Wolfe line search (8) and (9), let $x_{k+1} = x_k + \alpha_k d_k$.

Step 3: Calculate the search direction d_{k+1} by (7).

Step 4: Set $k := k+1$ and go to step 1.

Algorithm 2 [27].

Step 0: Choose an initial point $x_0 \in \Re^n$, $0 < \varepsilon < 1$, $0<\mu<1$, $0 < \sigma_1 < \sigma_2 < 1$. Set $C_0 = f_0, Q_0 = 1$, $d_0 = -\nabla f_0 = -g_0$, $k := 0$;

Step 1: If $\| g_k \|_2 \le \varepsilon$, then stop; Otherwise go to step 2;

Step 2: Compute steplength α_k by the nonmonotone Wolfe line search (12) and (13), let $x_{k+1} = x_k + \alpha_k d_k$

Step 3: Calculate the search direction d_{k+1} by (7).

Step 4: Let

$$Q_{k+1} = \mu Q_k + 1, C_{k+1} = \frac{\mu Q_k C_k + f_{k+1}}{Q_{k+1}} \tag{14}$$

Step 5: Set $k := k+1$ and go to step 1.

We will concentrate on the convergent results of NLSA in the following section.

3. Convergence Analysis

In order to establish the convergence of NLSA, the following assumptions are often needed [27,29,31,34,35,48].

Assumption 3.1: 1) f is bounded below on the bounded level set $\phi = \{x \in \Re^n : f(x) \leqslant f(x_0)\}$; 2) In ϕ, f is differentiable and its gradient is Lipschitz continuous, namely, there exists a constants $L>0$ such that $\|g(x)-g(y)\| \leqslant L\|x-y\|$, for all $x, y \in \phi$.

In the following, we assume that $\|g_k\| \ne 0$ for all k, for otherwise a stationary point has been found. The following lemma shows that the search direction dk satisfies the sufficiently descent condition without any line search technique.

Lemma 3.1 (Lemma 3.1 in [27]) Consider (7). Then we have (6).

Based on Lemma 3.1 and Assumption 3.1, let us prove the global convergence theorem of NLSA.

Theorem 3.1 Let $\{\alpha_k, d_k, x_{k+1}, g_{k+1}\}$ be generated by the NLSA, and Assumption 3.1 holds. Then we have

$$\sum_{k=0}^{\infty} \left(\frac{g_k^T d_k}{\|d_k\|} \right)^2 < +\infty \tag{15}$$

and thus

$$\lim_{k\to\infty} \left(\frac{g_k^T d_k}{\| d_k \|} \right)^2 = 0 \tag{16}$$

Proof. Denote that

$$f(x_{h(k)}) = \max_{0\le j\le n(k)} f(x_{k-j}), n(k) = \min\{H, k\}.$$

Using Lemma 3.1 and (10), we have

$$f_{k+1} \le \max_{0\le j\le n(k)} f_{k-j} + \varepsilon_1 \alpha_k d_k^T g_k \le \max_{0\le j\le n(k)} f_{k-j} = f(x_{h(k)})$$

Thus, we get

$$\begin{aligned} f(x_{h(k)}) &= \max_{0\le j\le n(k)} f(x_{k-j}) \\ &\le \max\left\{ f(x_{h(k)}) = \max_{0\le j\le n(k)} f(x_{k-1-j}), f_k \right\} \\ &= \max\left\{ f(x_{h(k-1)}), f(x_k) \right\} \\ &= f(x_{h(k-1)}), k = 1,2,..., \end{aligned} \tag{17}$$

i.e., the sequence $\{f(x_{h(k)}\}$ monotonically decreases. Since $f(x_{h(0)}) = f(x_0)$, we deduce that

$$f(x_k) \le f(x_{h(k-1)}) \le ... \le f(x_{h(0)}) = f_0$$

then $x_k \in \phi$. By Assumption 3.1: 1), we know that there exists a positive constant M such that

$$\| x \| \le M$$

Therefore,

$$\| \alpha_k d_k \| = \| x_{k+1} - x_k \| \le \| x_{k+1} \| + \| x_k \| \le 2M.$$

By (11), we have

$$\max\left\{\varepsilon_2, 1 - (\alpha_k \| d_k \|^p)\right\} \ge \max\left\{\varepsilon_2, 1 - (2M)\right\}^P$$

Let $\varepsilon_3 = \max\left\{\varepsilon_2, 1 - (2M)^P\right\} \in (0.1)$. Using (11) and Assumption 3.1: 2), we have

$$(\varepsilon_3 - 1) g_k^T d_k \le (g_{k+1} - g_k)^T d_k \le \| g_{k+1} - g_k \| \| d_k \| \le \alpha_k L \| d_k \|^2$$

Then we get

$$\alpha_k \ge \frac{(\varepsilon_3 - 1) g_k^T d_k}{L \| d_k \|^2} \tag{18}$$

By (10) and Lemma 3.1, we obtain

$$f_{k+1} \le f(x_{h(k)}) + \varepsilon_1 \alpha_k d_k^T g_k \le f(x_{h(k)}) - \frac{\varepsilon_1 (1 - \varepsilon_2)}{L} \left(\frac{d_k^T g_k}{\| d_k \|} \right)^2 \tag{19}$$

By Lemma 2.5 in [45], we conclude that from (19)

$$\sum_{k=0}^{\infty} \left(\frac{g_k^T d_k}{\| d_k \|} \right)^2 < +\infty \tag{20}$$

Therefore, (15) holds. (15) implies (16). The proof is complete.

Remark. If there exists a constant c_0>0 such that $\| d_k \| \le c_0 \| g_k \|$ for all sufficiently large k. By (6) and (16), it is easy to obtain $\|g_k\| \to 0$ as $k \to \infty$.

4. Numerical Results

In this section, we report some numerical results with NLST, Algorithm 1, and Algorithm 2. All codes were written in MATLAB and run on PC with 2.60GHz CPU processor and 256MB memory and Windows XP operation system. The parameters and the rules are the same to those of [27], we state it as follows: $\sigma_1 = 0.1, \sigma_2 = 0.9, \mu = 10^{-4}, \varepsilon = 10^{-5}$. Since the line search cannot always ensure the descent condition $d_k^T g_k < 0$, uphill search direction may occur in the numerical experiments. In this case, the line search rule maybe fails. In order to avoid this case, the stepsize _k will be accepted if the searching number is more than twenty five in the line search. We will stop the program if the condition $\| \nabla f(\beta) \| 1e-5$ is satisfied. We also stop the program if the iteration number is more than one thousand, and the corresponding method is considered to be failed. In this experiment, the direction is defined by:

$$d_{k+1} = \begin{cases} -g_{k+1} + \dfrac{\| g_{k+1} \|^2}{-g_{k+1}^T d_k} d_{k,} \ if \ g_{k+1}^T d_k < 1e-10 \\ -g_{k+1}, \qquad\qquad otherwise \end{cases} \tag{21}$$

The parameters of the presented algorithm is chosen as: $\varepsilon_1 = 0.01, \varepsilon_2 = 0.1$, p=5, H=8.

In this section, we will test three practical problems to show the efficiency of the proposed algorithm, where Problem 1 and 2 can be seen from [27]. In Table 1 and 2, the initial points are the same to those of paper [27] and the results of Algorithm 1 and Algorithm 2 can also be seen from [27]. In order to show the efficiency of these algorithms, the residuals of sum of squares is defined by

$$SSE_p(\hat{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_1)^2,$$

where $\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + ... + \hat{\beta}_p X_{ip}$, i = 1, 2, …, n, and $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$ are the parameters when the program is stopped or the solution is obtained from one way. Let

$$RMS_p(\hat{\beta}) = \frac{SSE_p(\hat{\beta})}{n-p}$$

where n is the number of terms in problems, and p is the number of parameters, if RMS_p is smaller, then the corresponding method is better [49].

The columns of the tables 4−6 have the following meaning:

β^*: the approximate solution from the method of extreme value of calculus or some software. $\grave{\beta}$: the solution as the program is terminated. $\breve{\beta}$: the initial point. ε_*: the relative error between RMS_p (β^*) and RMSp ($\grave{\beta}$) defined by $\varepsilon_* = \dfrac{RMS_p(\beta^*) - RMS_p(\beta)}{RMS_p(\beta^*)}$.

Problem 1. In the following table, there is data of some kind of commodity between year demand and price:

The statistical results indicate that the demand will possibly change though the price is inconvenient, and the demand will be possible invariably though the price changes. Overall, the demand will decrease with the increase of the price. Our objective is to find out the approximate function between the demand and the price, namely, we need to find the regression equation of d to the *p*.

It is not difficult to see that the price p and the demand d are linear relations. Denote the regression function by $\grave{d} = \beta_0 + \beta_1 p$, where β_0 and β_1 are the regression parameters.

Our work is to get β_0 and β_1. By least squares method, we need to solve the following problem

$$\min \sum_{i=0}^{n} (d_i - (\beta_0 + \beta_1 p_i))^2$$

and obtain β_0 and β_1, where n=10. Then the corresponding unconstrained optimization problem is defined by

$$\min_{\beta \in R^2} f(\beta) = \sum_{i=1}^{n} (d_i - \beta(1, p_i))^2 \tag{22}$$

Problem 2. In the following table, there is data of the age x and the average height H of a pine tree:

Similar to problem 1, it is easy to see that the age x and the average height H are parabola relations. Denote the regression function by $\hat{h} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where β_0, β_1 and β_2 are the regression parameters. Using least squares method, we need to solve the following problem

$$\min \sum_{i=0}^{n} (h_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2))^2$$

and obtain β_0, β_1 and β_2, where n=10. Then the corresponding unconstrained optimization problem is defined by

$$\min_{\beta \in R^3} f(\beta) = \sum_{i=1}^{n} (h_i - \beta(1, x_{i,} x_i^2))^2 \tag{23}$$

It is well known that the above problems (22) and (24) can be solved by extreme value of calculus. Here we will solve these two problems by our methods and other two methods, respectively.

Problem 3. Supervisor Performance (Chapter 3 in [49]).

Table 1. Demand and price

Price p_i($)	1	2	2	2.3	2.5	2.6	2.8	3	3.3	3.5
Demand d_i (500g)	5	3.5	3	2.7	2.4	2.5	2	1.5	1.2	1.2

Table 2. Data of the age x and the average height H of a pine tree

x_i	2	3	4	5	6	7	8	9	10	11
h_i	5.6	8	10.4	12.8	15.3	17.8	19.9	21.4	22.4	23.2

Table 3. The data of appraisal to supervisor

line	Y	X1	X2	X3	X4	X5	X6
1	43	51	30	39	61	92	45
2	63	64	51	54	63	73	47
3	71	70	68	69	76	86	48
4	61	63	45	47	54	84	35
5	81	78	56	66	71	83	47
6	43	55	49	44	54	49	34
7	58	67	42	56	66	68	35
8	71	75	50	55	70	66	41
9	72	82	72	67	71	83	31
10	67	61	45	47	62	80	41
11	64	53	53	58	58	67	34
12	67	60	47	39	59	74	41
13	69	62	57	42	55	63	25
14	68	83	83	45	59	77	35
15	77	77	54	72	79	77	46
16	81	90	50	72	60	54	36
17	74	85	64	69	79	79	63
18	65	60	65	75	55	80	60
19	65	70	46	57	75	85	46
20	50	58	68	54	64	78	52
21	50	40	33	34	43	64	33
22	64	61	52	62	66	80	41
23	53	66	52	50	63	80	37
24	40	37	42	58	50	57	49
25	63	54	42	48	66	75	33
26	66	77	66	63	88	76	72
27	78	75	58	74	80	78	49
28	48	57	44	45	51	83	38
29	85	85	71	71	77	74	55
30	82	82	39	59	64	78	39

where Y is overall appraisal to supervisor, X_1 denotes to processes employee's complaining, X_2 refer to do not permit the privilege, X_3 is the opportunity about study, X_4 is promoted based on the work achievement, X_5 refer to too nitpick to the bad performance, and X_6 is the speed of promoting to the better work. The above data can also be found at: http://www.ilr.cornell.edu/%7Ehadi/RABE3/Data/P054. txt.

Assume that the relation between Y and Xi (i=1, 2, …, 6) is linear [49], similar to Problem 1 and 2, the corresponding unconstrained optimization problem is defined by

$$\min_{\beta\in R^7} f(\beta)=\sum_{i=1}^{n}(h_i-\beta(1,x_{i1},x_{i2},\ldots,x_{i6}))^2 \qquad (24)$$

where n = 30. The regression equation from one fitting way (see Chapter 3.8 in [49]) is given by

$$\hat{Y}=10.787+0.613X_1-0.073X_2+0.320X_3+0.081X_4$$
$$+0.038\,X_5-0.217X_6$$

which means that β^* =(10.787,0.613,−0.073,0.320, 0.081,0.038,−0.217). For Problem 3, the initial points are chosen as follows:

$\breve{\beta}_1$=(10, 0.1, −0.05, 1, 0.1, 2, −0.1); $\breve{\beta}_2$=(10, −0.1, 0.05, −1, −0.1, −2, 0.1);

$\breve{\beta}_3$=(10.1, −0.01, 0.5, −0.2, −0.01, −0.2, 4); $\breve{\beta}_4$=(10.8, −100, 20, −70, −50, −40, 60);

$\breve{\beta}_5$= (9, 0.01, −0.5, 1, 0.01, 2, −0.01); $\breve{\beta}_6$= (11, 0.01, −0.5, 1, 0.01, 2, −0.01).

These numerical results of Table 4-6 indicate that proposed algorithm is more competitive than those of Algorithm 1 and 2, and the initial points do not influence the results obviously about these three methods. Moreover, the numerical results of NLSA, Algorithm 1, and Algorithm 2 are better than those of these methods from extreme value of calculus or some software. Then we can conclude that the numerical method will outperform the method of extreme value of calculus in some sense, and some software for regression analysis could be further improved in the future. Overall, the direction defined by (7) is notable.

Table 4. Test results for Problem 1

β^*=(6.5–1.6)	$\breve{\beta}$	$\grave{\beta}$	RMSp ($\grave{\beta}$)	RMSp(β^*)	ε_*
Algorithm 1	(1, −0.01)	(6.438301, −1.575289)	0.039736	0.040100	0.908%
	(−10,0.04)	(6.438280, −1.575313)	0.039736	0.040100	0.908%
	(−2, −1.0)	(6.438285, −1.575314)	0.039736	0.040100	0.908%
	(15,15)	(6.438287, −1.575316)	0.039736	0.040100	0.908%
Algorithm 2	(1, −0.01)	(6.438301, −1.575289)	0.039736	0.040100	0.908%
	(−10,0.04)	(6.438280, −1.575313)	0.039736	0.040100	0.908%
	(−2,-1.0)	(6.438285, −1.575314)	0.039736	0.040100	0.908%
	(15,15)	(6.438287, −1.575316)	0.039736	0.040100	0.908%
NLSA	(1, −0.01)	(6.438280, −1.575312)	0.039736	0.040100	0.908%
	−10,0.04)	(6.438292, −1.575317)	0.039736	0.040100	0.908%
	(−2,-1.0)	(6.438291, −1.575316)	0.039736	0.040100	0.908%
	(15,15)	(6.438280, −1.575312)	0.039736	0.040100	0.908%

Table 5. Test results for Problem 2

β^*=(−1.33, 3.46, −0.11)	β	β	RMSp (β)	RMSp (β^*)	ε_*
Algorithm 1	(−1.1,3.0, −0.5)	(−1.296574, 3.450247, −0.107896)	0.171774	0.183900	6.5938%
	(−1.2,3.2, −0.3)	(−1.328742, 3.460876, −0.108650)	0.171712	0.183900	6.6273%
	(−0.003,7.0, −0.8)	(−1.328504, 3.460798, −0.108646)	0.171713	0.183900	6.6272%
	(−0.001,7.0, −0.5)	(−1.321726, 3.458558, −0.108483)	0.171717	0.183900	6.6248%
Algorithm 2	(−1.1,3.0, −0.5)	(−1.296574, 3.450247, −0.107896)	0.171774	0.183900	6.5938%
	(−1.2,3.2, −0.3)	(−1.328742, 3.460876, −0.108650)	0.171712	0.183900	6.6273%
	(−0.003,7.0, −0.8)	(−1.328504, 3.460798, −0.108646)	0.171713	0.183900	6.6272%
	(−0.001,7.0, −0.5)	(−1.321726, 3.458558, −0.108483)	0.171717	0.183900	6.6248%
NLSA	(−1.1,3.0, −0.5)	(−1.331296, 3.461720, −0.108711)	0.171712	0.183900	6.6274%
	(−1.2,3.2, −0.3)	(−1.331232, 3.461699, −0.108709)	0.171712	0.183900	6.6274%
	(−0.003,7.0, −0.8)	(−1.331140, 3.461669, −0.108707)	0.171712	0.183900	6.6274%
	(−0.001,7.0, −0.5)	(−1.202673, 3.422106, −0.106011)	0.172583	0.183900	6.1539%

Table 6. Test results for Problem 2

β^*	β	β	RMSp(β)	RMSp(β^*)	ε_*
Algorithm 1	$\breve{\beta}_1$	(10.011713, 0.502264, -0.002329, 0.361596, 0.061871, 0.152295, -0.353686)	85.261440	89.584291	4.8255%
	$\breve{\beta}_2$	(10.124457, 0.502394, -0.002598, 0.361313, 0.061446, 0.151381, -0.353527)	85.235105	89.584291	4.8549%
	$\breve{\beta}_3$	(10.294617, 0.502056, -0.002462, 0.360523, 0.062746, 0.149161, -0.354270)	85.196215	89.584291	4.8983%
	$\breve{\beta}_4$	(11.404702, 0.501820, -0.004943, 0.357265, 0.060921, 0.140326, -0.354036)	84.963796	89.584291	5.1577%
	$\breve{\beta}_5$	(9.542516, 0.503279, -0.001805, 0.362715, 0.061217, 0.156318, -0.352638)	85.375457	89.584291	4.6982%
	$\breve{\beta}_6$	(11.071364, 0.501290, -0.004085, 0.358312, 0.062185, 0.143081, -0.354614)	85.029566	89.584291	5.0843%
Algorithm 2	$\breve{\beta}_1$	(10.011713, 0.502264, -0.002329, 0.361596, 0.061871, 0.152295, -0.353686)	85.261440	89.584291	4.8255%
	$\breve{\beta}_2$	(10.166214, 0.502293, -0.002549, 0.360902, 0.062002, 0.151044, -0.354147)	85.225461	89.584291	4.8656%
	$\breve{\beta}_3$	(10.639778, 0.502423, -0.003742, 0.360018, 0.060167, 0.147253, -0.353327)	85.119812	89.584291	4.9836%
	$\breve{\beta}_4$	(11.404239, 0.501827, -0.004935, 0.357227, 0.060988, 0.140322, -0.354037)	84.963893	89.584291	5.1576%
	$\breve{\beta}_5$	(11.404239, 0.501827, -0.004935, 0.357227, 0.060988, 0.140322, -0.354037)	85.506424	89.584291	4.5520%
	$\breve{\beta}_6$	(11.032035, 0.501940, -0.004251, 0.358407, 0.061171, 0.143518, -0.353940)	85.037491	89.584291	4.5520%
NLSA	$\breve{\beta}_1$	(10.326165, 0.502177, -0.002900, 0.360625, 0.061701, 0.149611, -0.353760)	85.189017	89.584291	4.9063%
	$\breve{\beta}_2$	(10.042910, 0.501267, -0.001983, 0.359836, 0.065677, 0.151241, -0.354909)	85.254692	89.584291	4.8330%
	$\breve{\beta}_3$	(10.525637, 0.502094, -0.003292, 0.359987, 0.061542, 0.147873, -0.353823)	85.144572	89.584291	4.9559%
	$\breve{\beta}_4$	(11.431772, 0.501805, -0.005001, 0.357160, 0.060909, 0.140080, -0.354047)	84.958622	89.584291	5.1635%
	$\breve{\beta}_5$	(9.653770, 0.502364, -0.001653, 0.362701, 0.062144, 0.155364, -0.353611)	85.347711	89.584291	4.7292%
	$\breve{\beta}_6$	(11.504977, 0.501791, -0.005132, 0.356938, 0.060866, 0.139459, -0.354060)	84.944709	89.584291	5.1790%

5. Conclusions

The major contribution of this paper is an extension of the direction (7) to a nonmonotone line search technique (GLL line search). The presented method possess global convergence and the numerical results show that the given algorithm is successful for the test problems. These test numerical results further show that the direction defined by (7) is notable. We hope the method can be a further topic for the regression analysis.

For further research, we should study other line search methods for regression analysis.

Moreover, more numerical experiments for large practical problems about regression analysis should be done in the future.

REFERENCES

[1] D. M. Bates and D. G. Watts, “Nonlinear regression analysis and its applications,” New York: John Wiley & Sons, 1988.

[2] S. Chatterjee and M. Machler, “Robust regression: A weighted least squares approach, communications in statistics,” Theorey and Methods, 26, pp. 1381−1394, 1997.

[3] R. Christensen, “Analysis of variance, design and regression: Applied statistical methods,” New York: Chapman and Hall, 1996.

[4] N. R. Draper and H. Smith, “Applied regression analysis,” 3rd ed., New York: John Wiley & Sons, 1998.

[5] F. A. Graybill and H. K. Iyer, “Regression analysis: Concepts and applications, Belmont,” CA: Duxbury Press, 1994.

[6] R. F. Gunst and R. L. Mason, “Regression analysis and its application: A data-Oriented approach,” New York: Marcel Dekker, 1980.

[7] R. H. Myers, “Classical and modern regression with applications,” 2nd edition, Boston: PWS-KENT Publishing Company, 1990.

[8] R. C. Rao, “Linear statistical inference and its applications,”New York: John Wiley & Sons, 1973.

[9] D. A. Ratkowsky, “Nonlinear regression modeling: A unified practical approach,” New York: Marcel Dekker, 1983.

[10] D. A. Ratkowsky, “Handbook of nonlinear regression modeling,” New York: Marcel Dekker, 1990.

[11] A. C. Rencher, “Methods of multivariate analysis,” New York: John Wiley & Sons, 1995.

[12] G. A. F. Seber and C. J. Wild, “Nonlinear regression,” New York: John Wiley & Sons, 1989.

[13] A. Sen and M. Srivastava, “Regression analysis: Theory, methods, and applications,” New York: Springer-Verlag, 1990.

[14] J. Fox, “Linear statistical models and related methods,” New York: John Wiley & Sons, 1984.

[15] S. Haberman and A. E. Renshaw, “Generalized linear models and actuarial science,” The Statistician, 45, pp. 407−436, 1996.

[16] S. Haberman and A. E. Renshaw, “Generalized linear models and excess mortality from peptic ulcers,” Insurance: Mathematics and Economics, 9, pp. 147−154, 1990.

[17] R. R. Hocking, “The analysis and selection of variables in linear regression,” Biometrics, 32, pp. 1−49, 1976.

[18] P. McCullagh and J. A. Nelder, “Generalized linear models,” London: Chapman and Hall, 1989.

[19] J. A. Nelder and R. J. Verral, “Credibility theory and generalized linear models,” ASTIN Bulletin, 27, pp. 71−82, 1997.

[20] M. Raydan, “The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem,”SIAM Journal of Optimization, 7, pp. 26−33, 1997.

[21] J. Schropp, “A note on minimization problems and multistep methods,” Numerical Mathematics, 78, pp. 87−101, 1997.

[22] J. Schropp, “One-step and multistep procedures for constrained minimization problems,” IMA Journal of Numerical Analysis, 20, pp. 135−152, 2000.

[23] D. J. Van. Wyk, “Differential optimization techniques,” Appl. Math. Model, 8, pp. 419−424, 1984.

[24] M. N. Vrahatis, G. S. Androulakis, J. N. Lambrinos, and G. D. Magolas, “A class of gradient unconstrained minimization algorithms with adaptive stepsize,” Journal of Computational and Applied Mathematics, 114, pp. 367−386, 2000.

[25] G. L. Yuan and X. W. Lu, “A new line search method with trust region for unconstrained optimization,” Communications on Applied Nonlinear Analysis, Vol. 15, No. 1, pp. 35−49, 2008.

[26] G. Yuan, X. Lu, and Z. Wei, “New two-point stepsize gradient methods for solving unconstrained optimization problems,” Natural Science Journal of Xiangtan University, (1)29, pp. 13−15, 2007.

[27] G. L. Yuan and Z. X. Wei, “New line search methods for unconstrained optimization,” Journal of the Korean Statistical Society, 38, pp. 29−39, 2009.

[28] Y. Dai, “A nonmonotone conjugate gradient algorithm for unconstrained optimization,” Journal of Systems Science and Complexity, 15, pp. 139−145, 2002.

[29] Y. Dai and Y. Yuan, “A nonlinear conjugate gradient with a strong global convergence properties,” SIAM Journal of Optimization, 10, pp. 177−182, 2000.

[30] R. Fletcher, “Practical Method of Optimization,” Vol 1: Unconstrained Optimization, 2nd edition, Wiley, New York, 1997.

[31] R. Fletcher and C. Reeves, “Function minimization by conjugate gradients,” The Computer Journal, 7, pp, 149−154, 1964.

[32] Y. Liu and C. Storey, “Effcient generalized conjugate gradient algorithms, part 1: theory,” Journal of Optimization Theory and Application, 69, pp. 17−41, 1992.

[33] E. Polak and G. Ribiere, “Note sur la convergence de directions conjugees,” Rev. Francaise informat Recherche Operatinelle, 3e Annee, 16, pp. 35−43, 1969.

[34] Z. Wei, G. Li, and L. Qi, “New nonlinear conjugate gradient formulas for large-scale unconstrained optimization problems,” Applied Mathematics and Computation, 179, pp. 407−430, 2006.

[35] Z. Wei, S. Yao, and L. Liu, “The convergence properties of some new conjugate gradient methods,” Applied Mathematics and Computation, 183, pp. 1341−1350, 2006.

[36] G. L. Yuan, “Modified nonlinear conjugate gradient methods with sufficient descent property for large-scale optimization problems,” Optimization Letters, DOI: 10. 1007/s11590−008−0086−5, 2008.

[37] G. L. Yuan and X. W. Lu, “A modified PRP conjugate gradient method,” Annals of Operations Research, 166, pp. 73−90, 2009.

[38] J. C. Gibert and J. Nocedal, “Global convergence properties of conjugate gradient methods for optimization,” SIAM Journal of Optimization, 2, pp. 21−42, 1992.

[39] J. J. Mor´e, B. S. Garbow, and K. E. Hillstrome, “Testing unconstrained optimization software,” ACM Transactions Math. Software, 7, pp. 17−41, 1981.

[40] L. Grippo, F. Lamparillo, and S. Lucidi, “A nonmonotone line search technique for Newton’s method,” SIAM Journal of Numerical Analysis, 23, pp. 707−716, 1986.

[41] L. Grippo, F. Lamparillo, and S. Lucidi, “A truncate Newton method with nonmonotone line search for unconstrained optimization,” Journal of Optimization Theory and Applications, 60, pp. 401−419, 1989.

[42] L. Grippo, F. Lamparillo, and S. Lucidi, “A class of nonmonotone stabilization methods in unconstrained optimization,” Numerical Mathematics, 59, pp. 779−805, 1991.

[43] G. H. Liu, J. Y. Han, and D. F. Sun, “Global convergence analysis of the BFGS algorithm with nonmonotone linesearch,” Optimization, Vol. 34, pp. 147−159, 1995.

[44] G. H. Liu, J. M. Peng, The convergence properties of a nonmonotonic algorithm,” Journal of Computational Mathematics, 1, pp. 65−71, 1992.

[45] J. Y. Han and G. H. Liu, “Global convergence analysis of a new nonmonotone BFGS algorithm on convex objective functions,” Computational Optimization and Applications 7, pp. 277−289, 1997.

[46] G. L. Yuan and Z. X. Wei, “The superlinear convergence analysis of a nonmonotone BFGS algorithm on convex objective functions,” Acta Mathematica Sinica, English Series, Vol. 24, No. 1, pp. 35−42, 2008.

[47] H. C. Zhang and W. W. Hager, “A nonmonotone line search technique and its application to unconstrained optimization,” SIAM Journal of Optimization, Vol. 14, No. 4, pp. 1043−1056, 2004.

[48] M. R. Hestenes and E. Stiefel, “Method of conjugate gradient for solving linear equations,” J, Res. Nat. Bur. Stand., 49, pp. 409−436, 1952.

[49] S. Chatterjee, A. S. Hadi, and B. Price, “Regression analysis by example,” 3rd Edition, John Wiley & Sons, 2000.

(Edited by Vivian and Ann)

J. Serv. Sci. & Management, **2009, 2: 43-46**
Published Online March 2009 in SciRes (www.SciRP.org/journal/jssm)

Study on Measuring Methods of Real Estate Speculative Bubble

Yifei Lai[1], Huawei Xu[1], Junping Jia[1]

[1]School of Economics and Management, WuhanUniversity, Wuhan, Hubei, China.
Email: lyf37319@163.Com

Received December 18th, 2008; revised January 14th, 2009; accepted February 5th, 2009.

ABSTRACT

The paper analyses the causes of the bubble of the real estate, then elaborates real estate bubble theory based on speculation. This paper establishes a regression model of real estate's price with relevant economic variables, and builds the econometric model to measure real estate's speculative bubble. In application of the model for empirical research on real estate's speculative bubble of Chongqing, the paper concludes that globally there is no bubble in Chongqing (but on the edge of bubble). Finally, by analyzing the common points about the existence of bubble, the paper indicates that real estate's investment and macro-economic indexes are disjointed, and thus the investment is excessive, which can in turn corroborate the conclusions obtained by the measuring model.

Keywords: *real estate, speculative bubble, measuring model of speculative bubble, overinvestment*

1. Introduction

At present, China is under the pressure of high inflation, and real estate price soared quickly. From 2005 to 2007, the government adopted a series of macro-control policies, for example, in March 2005, "country's 8 items" came out, so that the regulation of real estate was boosted to a high degree of polity; in April 2006, mortgage interest rate rose again; in May, "country's 6 items" came out, then waged a new round of large-scale control; at the same period, China begun to impose tax on second-hand estate's sales; in July, China begun to impose personal income tax of transferring second-hand estate; in September, down payment rose. However, the real estate industry continues to show strong-run tendency, real estate price remains high. The real estate industry is highly relevant to many other industries, and its positive run can promote the development of other industries. Otherwise, if the real estate industry goes against the Law of value of market, its price separates from the market base but keeps irrational growth, the bubble is inevitable, and when the bubble has expanded to a certain degree to leak, then the financial system will bear the brunt, and even the national economy will experience turbulence. In 1997, the breakdown of real estate speculative bubble plunged Japan into stagnant economic downturn. Thus, it is of great significance to measure the speculative bubble of the current real estate market and identify the over-investment. Chongqing's real estate was selected as an example to measure its speculative bubble applying the proposed methods.

2. The Theory and Measuring Model

2.1 Real Estate Bubble Based on Speculative Theory

According to the reasons of real estate bubble, under the effect of consumer expectation, there are many positive feedback effects, namely, investors dealing according to the tendency of past asset price, not to the real price. Thus, we can consider, positive feedback deal determines the change of future demand in real estate market, and then the expectation of future real estate price in market is determined by the expectation of future change of demand. Firstly, when the increasing rate of real estate price exceeds credit loan rate, the real estate price speculation comes out. At this time, investors achieve speculative aim by changing hand to get the price difference; secondly, there is time interval between buying and selling, which provides speculative possibility. At last, because of the imperfection of market mechanism and information asymmetries, the price arbitrage action of speculators will result in the achievement of expectation.

2.2 The Measuring Model of Real Estate Speculative Bubble

In terms of positive feedback mechanism, real estate price at current period will be affected by past several real estate price. Considering that speculations are general short-term price arbitrage action by selling real estate, because speculators are not aiming to achieve the steady long-term profit in the future, e.g., earning rents after buying estate.

Let h_t figures real increasing rate of real estate price at period t, which is achieved after eliminating the growth part of real estate price due to the increase of income Y_t. According to the analysis above, the real increase of future real estate price is only determined by the price expectation of economic subjects in terms of the price at current period, so we establish the econometric model below:

$$h_t = \theta_1 h_{t-1} + \theta_2 h_{t-2} + \varepsilon_t \qquad (1)$$

h_{t-1}, h_{t-2} separately figures the real rate of growth lagging one period and two periods, here at most two lag periods are discussed. θ_1 figures the effect of the increasing rate of real estate price lagging one period to the real estate price at current period, reflecting one-year economic subject's expectation of future real estate price tendency, so we define the coefficient θ_1 to mainly reflect real estate speculative bubble. Because we mainly consider speculators' short-term (one year) speculative action, then when we consider economic subjects' expectation at short-term but over one year, θ_2 is considered as the ancillary index to reflect how economic subjects' action affects the growth of real estate price after one year, ε_t figures the unexpected shock at current period.

3. Emprical Analysis

3.1 The Choice of Parameters

According to the factors affecting real estate price, we choose one-year credit loan rate of commercial banks as mortgaged lending rate of real estate, disposable income of urban residents, real estate price. We adopt weighted processing the one-year credit loan due to its change in a year. The influence of inflation on disposable income is eliminated along with the years. The real estate price is available from the literature data. Credit loan rate of banks reflects the support degree of finance institutions to the development of real estate industry; it also reflects the attitude of government toward the development of real estate industry. Because there are mainly urban residents buying estate (especially speculating) in cities, so we choose the disposable income of urban residents to reflect residents' demand (or consumption ability) for real estate.

3.2 The Establishment of Regression Equations

In order to establish regression model between real estate price and other variables, and ensure that, under confidence level, other variables prominently affect real estate price, and variables in the equation do not have correlation each other, we choose Stepwise regression to establish equation. In the equation, P_t (Price) means real estate price at current period, I_t (Rate) means rate, P_{t-1} (Lag price) figures real estate price lagging one period; Y_t (Income) figures the disposable income of urban residents. Data is mainly from "Statistical Yearbook 2007 of Chongqing" and correlative years' statistical yearbooks of Chongqing. The model is the foundation of establishing real estate price speculative bubble measuring model, so the accuracy of equation's establishment is essential.

According to the analysis above, we establish regression equation below:

$$\ln P_t = a_0 + a_1 \ln I_t + a_2 \ln Y_t + a_3 \ln P_{t-1} + \varepsilon \qquad (2)$$

The value of p that variables stay and kick out in the regression equation are separately set at 0.1 and 0.15,. In application of SAS 9.0 to do regression analysis, partial results of regression are shown below.

1) Regression model

$$\ln P_t = 1.131 + 0.126 \ln I_t + 0.37 \ln Y_t + 0.479 \ln P_{t-1} \qquad (3)$$

Through the analysis, we know that, under the confidence level of 10%, variables all stay in the equation, and can be considered to produce significant affects on the change of real estate price

2) The model fitting effect

From analytical results, we know that, the value of fitting degree equals 0.9892, and adjusted R^2 equals 0.9867, so the model is available wholly.

3) D-W test

The value of D-W equals 1.649, which is between 1.54 and 2.46, and indicates, under the confidence level of 5%, we can refuse the assumption of the sequence correlation.

3.3 The Measuring Model of Real Estate Speculation Bubble

From (3), the formula below can be obtained,

$$P_{t-1} = e^{1.131} I_{t-1}^{\ 0.26}\, Y_{t-1}^{\ 0.37}\, P_{t-2}^{\ 0.479}$$

$$P'_t = \frac{P_t}{I_t^{\ 0.131} Y_t^{\ 0.37}} = e^{1.131}\, P_{t-1}^{\ 0.479} \qquad (5)$$

or

$$\frac{P'_t}{P'_{t-1}} = \left(\frac{P_{t-1}}{P_{t-2}}\right)^{0.479}$$

New P'_t sequence eliminates the rate and income affect of real estate price, so we can better investigate the bubble due to price speculation.

$$h_t = \frac{P'_t}{P'_{t-1}} - 1$$

From (2),

$$h_t = \left(\frac{P_{t-1}}{P_{t-2}}\right)^{0.479} - 1 \qquad (6)$$

According to the analysis above, h_t means the real rate of real estate price at current period, so in terms of the formula $h_t = \theta_1 h_{t-1} + \theta_2 h_{t-2} + \varepsilon_t$ obtained from positive feedback mechanism, we can employ coefficient θ_1 to measure (one year) the degree of speculative bubble.

3.4 Econometric Model Measuring the Coefficient of Price Bubble

The tendency of real increasing rate of real estate price h_t (HPRATE) can be obtained easily, in application of

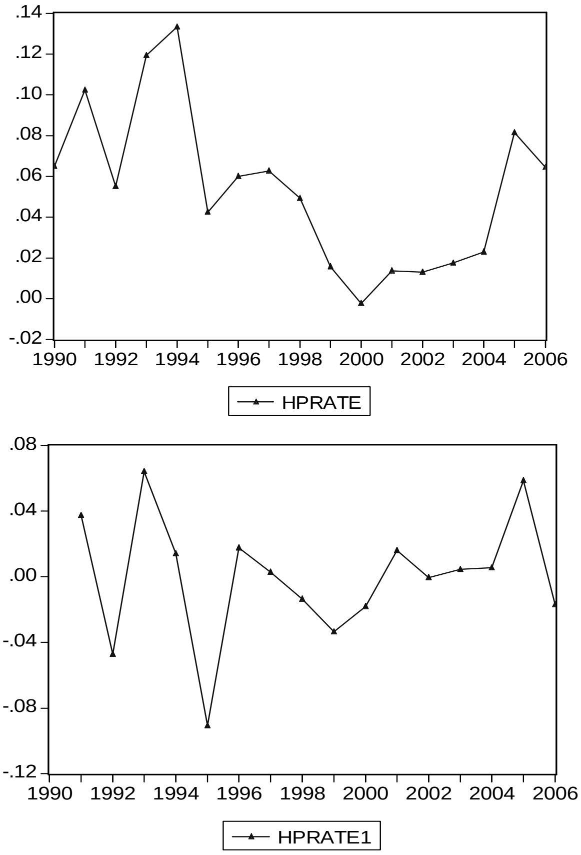


Figure 1. Trend of real increasing rate of real estate price h_t (HPRATE) and first difference (HPRATE1) time series

Eviews5.0 to conduct time series analysis, co-integration test with HPRATE, we find that it is not integrative under the level, after first difference, this time series (HPRATE1) become a unit root series, and the trend chat of the two are shown as following.

From the trend chart of HPRATE1, we know that, real increasing rate of real estate price took on large fluctuation in 1995, and then it fell to normal level in 1996, this illustrated that there existed affective factors producing large shock to real estate price, we consider that there exists structure change. In fact, the investment of real estate in our country between1992 and 1993 was exceeded, then in 1994, the government carried out macro-control of our economy, “City Real Estate Management Law” coming out, leading economy to practice soft landing. Considering the lag effect of macroeconomics policy, we set dummy variable PL to reflect the shock of policy.

Establish the following model:

$$h_t = \theta_1 h_{t-1} + \theta_2 h_{t-2} + \beta PL + \varepsilon_t \quad (7)$$

From the table above, we get the model below:
HPRATE1=0.007−0.12 * PL+[AR(1)=−0.63,
MA(1)=0.91, BACKCAST=1992] (8)
(0.78) (−3.49) (−3.21)

Table 1. Eviews analysis results

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.006593	0.008439	0.781300	0.4511
PL	−0.116048	0.033162	−3.499400	0.0050
AR(1)	−0.631575	0.196634	−3.211928	0.0083
MA(1)	0.910039	0.049313	18.45421	0.0000
R-squared	0.723529	Durbin-Watson stat		1.96375

Obviously, the original model is:

$$h_t = 0.007 - 0.12\, PL + 0.37\, h_{t-1} + 0.63\, h_{t-2} \quad (9)$$

3.5 Results Analysis

From the results, we know that coefficient of real estate speculation θ_1 equals 0.37 which is close to the international alertness line 0.4, but it is not at bubble level; ancillary index θ_2 equals 0.63, which also reflects the expectation (over one year) of economic subject in certain extent images bubble degree. Coefficient of policy shock variable PL β equals −0.12, which reflects that the macroeconomics policy in 1994 well restrained real estate speculation. So we can have following conclusion:

The real estate industry in Chongqing is on the edge of bubble, but does not take on bubble, the macroeconomics policies in 1994 worked. This conclusion basically accords with the reality of Chongqing.

3.6 Further Analysis of the Model Conclusion

From the conclusion of the model, we know that our Chongqing is on the edge of bubble, but not has bubble, but why there generally exist “bubble intimidation theory”, “bubble perdition theory”. We support the conclusion of speculation measuring model analysis by analyzing the relation between real estate investment and macroeconomics variable. We choose real estate investment (REINV) and GDP of Chongqing, and also analyze the relation of real estate investment and disposable income (INCOME) of urban residents.

1) Co-integration test. Through ADF test, we find the real estate investment (REINV2), GDP (GDP2, urban residents’ income (INCOME2) are all 2-order integration variables. Co-integration test uses M3 model and AIC principle to choose lag order. The lag order is 3 periods in co-integration test of REINV2 and GDP2, the value of AIC is 13.11743; the lag order is 3 periods in the co-integration of REINV2 and GDP2, the value of AIC is 15.84799. The results of Co-integration test are shown in Table 2.

2) Granger causality test. Because the three variables are all 2-order integration variables, we can directly use VAR model to conduct Granger causality test. We try to choose different orders to study the sensitivity of the test results, and the results are shown in Table 4.

Table 2. REINV2 and INCOME2 Johansen co-integration test results

Count of Co- integration Equations	Eiqenvalue	T-Statistic	5% Critical Value	prob.**
None	0.376550	9.292334	15.49471	0.3390
At most one	0.136708	2.205029	3.841466	0.1376
Count of Co-integration Equations	Eiqenvalue	Most Eiqenvalue Statistic	5% Critical Value	prob.**
none	0.376550	7.087305	14.26460	0.4789
At most one	0.136708	2.205029	3.841466	0.1376

Table 3. REINV2 and GDP2 Johansen co-integration test results

Count of Co-integration equations	Eiqenvalue	T-Statistic	5% Critical Value	prob.**
None	0.395752	10.20828	15.49471	0.2651
At most one	0.162037	2.651715	3.841466	0.1034
Count of Co-integration equations	Eiqenvalue	Most Eiqenvalue Statistic	5% Critical Value	prob.**
None	0.395752	7.556565	14.26460	0.4256
At most one	0.162037	2.651715	3.841466	0.1034

Table 4. Two groups of variables' Granger causality test results

Variables	H_0 (Null Hypothesis)	Probability				
		Lag=1	Lag=2	Lag=3	Lag=4	Lag=5
REINV2-GDP2	GDP2 does not Granger Cause INVEST2	0.726	0.643	0.32	0.790	0.884
	INVEST2 does not Granger Cause GDP2	0.074	0.109	0.016	0.137	0.413
REINV2-INCOME2	INCOME2 does not Granger Cause INVEST2	0.777	0.997	0.258	0.540	0.312
	INVEST2 does not Granger Cause INCOME2	0.0001	0.005	0.033	0.157	0.418

From Table 3 and Table 4, we can get that under the confidence level of 0.1, real estate investment, GDP and the income of urban residents do not have co-integration, which shows that real estate investment has been out of the track of macroeconomic development, and overinvestment appears.

Table 4 further shows that under the confidence level of 0.1, real estate investment and GDP, income only have one-way causality. Real estate investment can promote the growth of GDP and strongly promote income growth in the short term, which once again indicates the existence of speculation which is short-term.

4. Conclusions

This paper empirically analyzes real estate speculative bubble of Chongqing after establishing bubble measuring model. The results indicate that Chongqing's real estate has not reached the bubble level, but is close to critical value. The paper also analyzes the opinions of bubble theory in the society, and empirical analysis pointes out that there is excessive investment of real estate, which separates from the growth of GDP and residents' income in Chongqing. The conclusion accordingly supports the empirical analysis results of speculative bubble measuring model.

Empirical analysis concludes that real estate of Chongqing is on the verge of bubble, which offers consultation for policymaker to work out corresponding measures.

In addition, through macroeconomics regulation, the government should scale down the investment of real estate to the normal level, coordinate with macroeconomics indexes of the region, to avoid excessive investment then to further induce expansion of bubble.

REFERENCES

[1] P. Wang, "Market efficiency and rationality in property investment," Journal of Real Estate Finance and Economics, 21(2): pp.185−200, 2000.

[2] C. Lizieri and S. Satchell, "Interacion between property and equity markets: An investigation of linkages in the UK 1972−1992," Journal of Real Estate Finance and Economics, 1997(15): pp.11−25, 1997.

[3] J. Abraham and P. H. Hendershott, "Bubbles in metropolitan housing markets," Housing Res, 1995(6): pp. 191−207, 1995.

[4] P. Bacon, F. Mac Cabe, and A. Murphy, "An economics assessment of recent house price developments," [M] Government of Ireland Publication, Dublin, 1998.

[5] J. Eatwell, M. Milgate, and P. Newman, "The new palgrave: A dictionary of economics," London Mac. Millan, Vol. 1, pp. 28.

[6] K. H. Kim and H. S. Seoung, "Speculation and price bubbles in the korean and japanese real estate markets," Journal of Real Estate Finance and Economics [J], 1993(6): pp. 73−86.

[7] K. H. Kim and H. S. Lee, "Real estate price bubble and price forecasts in Korea," Proceedings of 5th AsRES conference in Beijing, 2000.

[8] J. K. Zhou, "Financially support excessively and the real estate bubble," Beijing University Press, 2005.

[9] H. Y. Liu and H. Zhang, "Real estate and socio-economic," Tsinghua University Press, 2006.

(Edited by Vivian and Ann)

J. Serv. Sci. & Management, **2009, 2: 47-55**
Published Online March 2009 in SciRes (www.SciRP.org/journal/jssm)

Polluting Productions and Sustainable Economic Growth: A Local Stability Analysis

Giovanni Bella[1]

[1]University of Cagliari, Italy.
Email: bella@unica.it

Received January 25th, 2009; revised February 24th, 2009; accepted March 10th, 2009.

ABSTRACT

The aim of this paper is to analyze the link between natural capital and economic growth, in a Romer-type economy characterized by dirty emissions in the production process, and to examine the conditions under which a sustainable growth, which implies a decreasing level of dirty emissions, might be both feasible and optimal. This work is close to Aghion-Howitt (1998) *with some more general specifications, in particular regarding the structure of preferences and the technological sector. We also deeply study the transitional dynamics of this economy towards the steady state, and conclude that a determinate saddle path sustainable equilibrium can be reached even in presence of a long run positive level of polluting emissions, thanks to a growing level of new home-made inventories, without whom some indeterminacy problems are likely to emerge.*

Keywords: *local stability, sustainable growth, aghion-howitt model*

1. Introduction

It is commonly believed that economic development might lead to overexploitation of natural resources and intensification of environmental damages, as for example the augment of carbon dioxide concentrations in the atmosphere due to an increase in transportation services. On the contrary, empirical evidence suggests that rich societies seek a less polluted environment to live in, so they are more willing to invest in abatement technologies and enforce environmental regulations. The logical consequence must be that economic activity will then also lower the dirtiness of any existing production technique, which leaves the door open, for example, to those supporting the so-called Environmental Kuznets Curve hypothesis [1].

During the last three decades, this counterposition encouraged many economists to develop models in which economic growth depends on the extractive use of the environment. Inspired by the work of the Club of Rome and its pessimistic view on the possibility to attain long-run growth under environmental constraints, these models tried to depict the conflict between growth and the environment. Over time the variety of models grew rapidly, differing not only with respect to the basic framework adopted but also with respect to the type of environmental resource being considered and the problem analyzed, mainly because each model has specific properties that become useful for the analysis of either specific economic concerns.

More recently, research on endogenous growth and the environment turned more and more attention from one- to multi-sector models, where knowledge accumulation might have the potential of lowering environmental damages through an increase in technological progress [2,3,4,5]. Likewise, in the seminal paper of Aghion and Howitt [6], to whom we will be referring to as *AH* from now on, it is shown that an unlimited growth can indeed be sustained when account is taken of both environmental resource use and innovation in abatement activities [7,8].

Broadly speaking, the properties of endogenous technical change have been widely investigated in the existing economic literature, with some indeterminacy problems and Hopf bifurcating outcomes being of particular concern [9]. On the contrary, we want to show in this paper that the introduction of the environmental issue can drive the economy back to a unique, locally stable, equilibrium solution, where sustainability of consumption is finally reached. However, this occurs only if a specific sustainability rule, stating that consumption and natural capital grow at the same rate, is to be followed, which is also consistent with a forward looking individual behavior and no myopic statement of the adopted social policy. Therefore, the basic question we want to address is whether a sustainable path can be reached even if some dirty production processes, assumed here to be necessary for any economic activity are adopted.

To this bulk of literature this paper devotes particular attention, aimed at developing a model close to *AH*, that considers pollution as a choice variable entering the production function as a measure of *dirtiness*, whose external effects allow to increase the level of output [10]. It seems then to be interpreted as pollution is a necessary part of production and economic growth. Moving a step forward from *AH*, we also deeply concentrate on the study of the transitional dynamics of the model, and provide the whole necessary and sufficient conditions for the existence of a feasible steady- state equilibrium path associated with a positive long-run growth. Moreover, we conclude that a determinate saddle path sustainable equilibrium can be reached even in presence of a long run positive level of polluting emissions, thanks to a growing level of new home-made inventories, without whom some indeterminacy problems are likely to emerge [11,12].

The rest of the paper is organized as follows. In section 2, we derive the formal structure of the model, with particular attention to the set of preferences, the level of technology, and the introduction of pollution as a crucial variable for the system to grow. In Section 3, we concentrate on the solution of the optimization problem, and deeply investigate the stability properties of the associated steady state solution. The transitional dynamics of this economy will provide some interesting results and policy suggestions on the way to drive an economy along a *sustainable growth* path. A final section concludes, and a subsequent Appendix provides all the necessary proofs.

2. A Model with Dirtiness

Before we enter the algebraic version of the model, we ought to provide some detailed explanations related to the production function, the dynamics of the environment, and the set of preferences used to characterize the economy, whose properties we want to investigate in the rest of the paper.

Following Romer, 1990, let S_0 represent the fixed amount of skilled labour, which can be devoted to production of the final good, S_Y, or to improvement of technology, S_A. Henceforth, we will normalize the problem by assuming

$$S_0 = S_Y + S_A = 1 \tag{1}$$

In particular, technology (*A*) is not fixed. It can be created by engaging human capital in research, growing over time according to

$$\frac{\dot{A}}{A} = \varphi + \gamma S_A \tag{2}$$

γ indicates the research success parameter. Let us then assume that technology A be partly the result of endogenous (home-made) R&D efforts, γS_A , whilst the remaining part depends on some exogenous new inventories, whose spill-over effects can be synthesized through a constant *catch-up* parameter, φ [13]. We assume $\dot{A}/A = \varphi + \gamma S_A > 0$, as long as either φ or γ and S_A are set positive.[1] Thus, technology can grow without bound. We will show afterwards that, if we relax the positiveness assumption on γ, the economy will face the emergence of some undesired and indeterminate equilibrium problems.

Moreover, although technology is not directly linked to pollution here, we basically consider the discovery of new goods, or new (i.e. less polluting) production processes, as the implicit way societies follow to broadly reduce their dependence from environmental resources. Basically, we are saying that each new inventory due to technological advance is also assumed to be cleaner than the previous one. This is also consistent with the empirical evidence that developed societies seek a less polluted environment to live in.

Note also that research activity is assumed to be human-capital-intensive and technology-intensive, with no capital (K) and ordinary unskilled labour (L) engaged in that activity. To produce the final good Y, however, K does enter as an input along with human capital S_Y and technology A.[2] According to the assumptions made in *AH*, the main feature of this economy is that production is also affected by another variable indicating the intensity of pollution, $z(t) \in [0,1]$, such that higher values of z yield more of the good but also more pollution[3]

$$Y = A^{\alpha}(1 - S_A)^{\alpha} K^{1-\alpha} z \tag{3}$$

We may also consider z as a measure of dirtiness of the existing production technique [10]. For example, focus on cheese manufacturing. Only a fraction of the raw milk processed gives rise to white cheese (or other diary products), the remaining is called whey, a liquid by-product, only partially recyclable, which constitutes the greater part of the resulting pollution loads. In other words, we are assuming that production of output arises at the expenses of the environment, with some polluting emissions being necessarily needed.

Moreover, it is assumed that the flow of pollution P is proportional to the level of production, and that the use of cleaner technologies (which means low values of z) reduces the pollution/output ratio.[4] Formally,

$$P = Yz^{\gamma} \tag{4}$$

[1] It is assumed that technology does not depreciate

[2] Remember that in Romer, 1990, technology is assumed to be made up of an infinite set of designs for capital, which (for simplicity) enter the production function in an additively separable manner, given by

$$Y = \eta^{\alpha+\beta-1}(S_Y A)^{\alpha}(LA)^{\beta} K^{1-\alpha-\beta} \quad 0 < \alpha, \beta < 1$$

where η represents the units of capital goods to produce one unit of any type of design. Here we assume, for simplicity, that there is no unskilled labour; that is all workers are supposed to be specialized. Let us then consider $L = S_Y$ to derive Equation (3), and normalize the scale parameter to unity, for simplicity ($\eta^{\alpha+\beta-1} = 1$).

[3] This production function exhibits constant returns to scale at a disaggregate level because each firm takes *z* as given. On the contrary, a social planner can internalize this kind of externality, due to pollution intensity, thus obtaining increasing returns.

[4] The extractive use of the environment in production can either be modeled as an input to production or, like here, as a by-product of production; that is, pollution influences output indirectly.

To clarify the utility of using both variables, P and z, as two sides of the same coin (the damages to the environment), let us make another example that can be drawn from current industrially advanced economies. Basically, oil combustion is being needed either to feed the engine of our cars or to stoke the furnaces of our firms, with CO_2 emissions being an unavoidable consequence. Referring to our model it would imply that only a fraction of the oil burnt (z) serves to produce final output (Y), the rest being pushed into the atmosphere as a resulting emissions' burden. Nevertheless, it is indeed true that not all of these emissions are damaging, since carbon sequestration due to forests allows, for example, to reduce the total pollution loads.

What distinguishes this economy from the one defined in *AH* is that we let pollution, P, depend also on the parameter expressing research success in technological advances, γ; that is like assuming that the bigger γ-values the smaller the impact of dirty techniques on pollution, and then the cleaner the ecosystem.[5]

On the other hand, the level of investment in physical capital is given by the usual functional form $\dot{K} = Y - C$.

2.1 Dynamics of the Environment

Commonly, the environmental sector can be represented by the dynamics of the stock of natural capital available to the economy, E:

$$\dot{E} = N(E) - P \tag{5}$$

where $N(E)$ determines the speed at which nature regenerates, while P measures the negative effect due to polluting emission.[6] The former is constantly reduced not only by economic activities, but also by non-anthropogenic processes, such that ecosystems have to devote part of their regeneration capacity to the maintenance of their own structure.[7]

If the capacity for regeneration exceeds the requirements for maintenance, $N(E)$ becomes positive. $N(E)$ can therefore be interpreted as the difference between natural resource reproduction and resource use for maintenance [14] that determines nature's capacity to recover from pollution and resource extraction [4,5].

Some authors [15] propose a linear representation of the regeneration function

$$N(E) = \theta E \tag{6}$$

where θ denotes the constant rate of regeneration.[8]

Following this approach, if we substitute both Equations (4) and (6) into (5), we explicitly end up with

$$\dot{E} = \theta E - Yz^{\gamma} \tag{7}$$

which represents the environmental constraint to be used in the subsequent maximization problem.

2.2 The Set of Preferences

Let the preferences of the representative agent depend either on the level of consumption, C_t, or the stock of natural capital available to the economy, E_t.[9] The intertemporal utility function is then given by

$$\int_0^{\infty} U(C_t, E_t) e^{-\rho t} dt \tag{8}$$

where ρ is the social discount rate.[10]

$U(\cdot)$ is continuous, twice differentiable, and possesses the following properties: $U_C > 0$, $U_E > 0$, $U_{CC} \leq 0$. Also suppose that $U(\cdot)$ is concave with respect to its two arguments: $U_{CC} \cdot U_{EE} - (U_{CE})^2 \geq 0$.[11]

Theoretically, sustainable development usually comprises two conditions. Firstly, a non-decreasing level of consumption or utility levels, and secondly a constant or improving state of the environment. Whether sustainable development in this sense can be optimal, depends on the functional form of the utility function [4].[12]

A specific utility function is assumed here to have the following CES structure

$$U(\cdot) = \frac{(CE)^{1-\sigma} - 1}{1-\sigma}$$

where $\sigma > 0$ represents the inverse of the intertemporal elasticity of substitution.

This functional form guarantees that both C and E grow at the same rate, so that the C/E ratio is constant in equilibrium [16].[13] We show in the next

[5] Conversely, a negative value of γ reduces abatement programs, thus finally increasing the amount of pollution realized.

[6] We follow here the broad definition of natural capital given by Costanza and Daly, 1992.

[7] Conventional wisdom holds that plants will purify the air, helping to reduce concentration of harmful gases. But, recently, it has been shown that when temperatures exceed a threshold, trees and other plants emit chemicals that encourage toxic ozone production (Science, 2004).

[8] Although several criticisms have been raised against the algebraic simplicity of this specification (e.g., Rosendahl, 1996), it remains still widely used in the literature of the field, as for example in our reference model of Aghion and Howitt, 1998.

[9] One interpretation would be forests, which contribute to welfare both as sources of timber and also as stocks which provide many ecosystem services to society (for example, carbon's sequestration, preservation of bio-diversity)

[10] For simplicity, time subscripts will be omitted in the rest of the paper

[11] Constraints to the optimization problem could, for example, be introduced by defining critical minimum levels for natural capital (Barbier and Markandya, 1990) or by excluding decreasing utility paths (Pezzey, 1992). But as these restrictions usually involve inequality constraints, they may complicate the optimization problem considerably

[12] While *AH* deal (to simplify the analysis) with a logarithmic, thus separable, utility function, we prefer to introduce a non-separable function instead (as in Musu, 1995), that allows to compare consumption and environmental quality as two substitutes, according to agents' tastes towards them. Nonetheless, it will be shown that both assumptions can be finally reconciled

[13] We show that an improvement in natural capital is conductive to growth only if we assume that consumption and natural capital are substitutes, which implies $U_{CE} < 0$. Therefore, households will be willing to postpone part of their consumption opportunities only if the expected stock of natural capital is improved

section that this assumption is rich of powerful consequences. In particular, for growth to be balanced, it will allow us to both derive (in equilibrium) a constant lower-bound level of dirty emissions, and a constant level of the pollution/output ratio either.

3. The Social Planner Maximization Problem

We assume that the social planner has to maximize the following discounted CES utility function,

$$\int_0^\infty \frac{(CE)^{1-\sigma}-1}{1-\sigma}e^{-\rho t}dt$$

subject to the following constraints:

$$\dot{K}=A^\alpha(1-S_A)^\alpha K^{1-\alpha}z-C$$
$$\dot{A}=(\varphi+\gamma S_A)A$$
$$\dot{E}=\theta E-A^\alpha(1-S_A)^\alpha K^{1-\alpha}z^{1+\gamma}$$

and given initial positive values:

$$A(0)=A_0 \quad K(0)=K_0 \quad E(0)=E_0$$

The current value Hamiltonian is given by

$$H_c=\frac{(CE)^{1-\sigma}-1}{1-\sigma}+\lambda\left[A^\alpha(1-S_A)^\alpha K^{1-\alpha}z-C\right]+$$
$$+\mu\left[\theta E-A^\alpha(1-S_A)^\alpha K^{1-\alpha}z^{1+\gamma}\right]+\vartheta[(\varphi+\gamma S_A)A]$$

where λ, μ and ϑ denote the costate variables associated with the accumulation of physical capital, natural capital and knowledge capital, respectively.[14]

Solution to this optimal control problem implies the following necessary first order conditions[15]

$$C^{-\sigma}E^{1-\sigma}=\lambda$$

$$\lambda=\mu(1+\gamma)z^\gamma$$

$$\lambda\alpha A^\alpha(1-S_A)^{\alpha-1}K^{1-\alpha}z-\mu\alpha A^\alpha(1-S_A)^{\alpha-1}K^{1-\alpha}z^{1+\gamma}=\vartheta\gamma A$$

accompanied by the equation of motion for each costate variable, that can be obtained with a bit of mathematical manipulation:

$$\frac{\dot{\lambda}}{\lambda}=-\left(\frac{\gamma}{1+\gamma}\right)(1-\alpha)A^\alpha(1-S_A)^\alpha K^{-\alpha}z+\rho$$
$$\frac{\dot{\mu}}{\mu}=\rho-\theta-\frac{C}{E}(1+\gamma)z^\gamma$$
$$\frac{\dot{\vartheta}}{\vartheta}=\rho-(\varphi+\gamma)$$

and the transversality conditions for a free terminal state, whose specification is provided in the Appendix, that jointly constitute the so-called canonical system.

Questions of interest include: how does pollution affect the growth rate of this economy in the steady state? And particularly, what is the optimal level of dirtiness? The basic feature of such a steady state implies that:

Remark 1 Along a sustainable balanced growth path (BGP):

1) The marginal rate of substitution between C and E is constant, $MRS_{C,E}=\varepsilon<0$.

2) Both C and E grow at a constant rate, $g=\frac{\rho-(\gamma+\varphi)}{1-2\sigma}$.

3) The degree of dirtiness, z, is constant.

4) The BGP is non-degenerate and the growth rate of the economy is positive.

In particular, from FOC's we can easily derive the following Bernoulli's differential equation for z,

$$\gamma\dot{z}+\phi z=\varepsilon(1+\gamma)z^{1+\gamma}$$

where $\phi=\gamma+\varphi-\theta$. More interestingly, a stable steady state occurs when

Remark 2 The rate of new technological advances is lower than the speed at which nature regenerates ($\phi<0$), and the level of dirty emissions converges to a positive minimum threshold, $\tilde{z}$ (stable equilibrium).

If we concentrate on the stable solution, evolutionary path for $z(t)$ follows consequently.

As depicted in Figure 1, when approaching the steady state the level of dirtiness (z) lowers, but never collapses to zero, $\tilde{z}=\left[\frac{\phi}{\varepsilon(1+\gamma)}\right]^{\frac{1}{\gamma}}$. It can be interpreted as an economy that moves along a long run sustainable path thanks

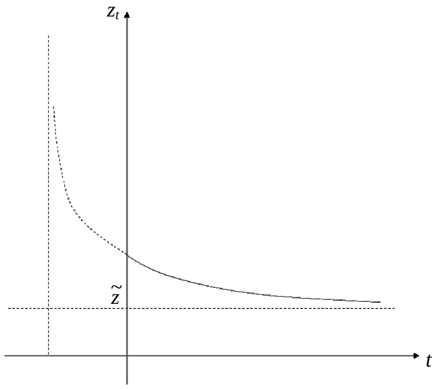


Figure 1. Evolution of dirty emissions

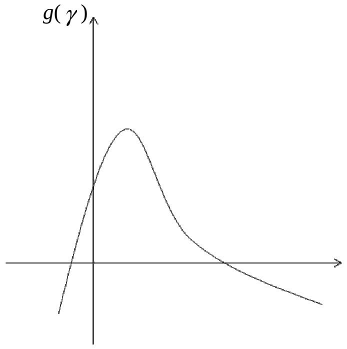


Figure 2. The BGP growth rate

[14] Appendix A derives the optimality conditions, which will be discussed in the rest of this section

[15] Necessary condition for a maximum can be checked by studying the sign of all principal minors of the Hessian matrix for the control variables of the problem, whose determinant is formed by the following signs

to a positive value of dirty emissions, unless we admit a stop in the economic development. This is consistent with the behavior of an advanced economy where, despite the presence of a high demand for environmental protection and a rise in technological innovations, it can be noted nonetheless a substitution amongst pollutants, whose pressure on the ecosystem is far away from disappearing.

This economy seems to mimic one where, to achieve balanced growth, pollution grows at the same level of output. However, we conclude that this economy behaves in a sustainable way only if natural capital grows more than technological sector. To summarize, it can be thought as a simple parable to explain why rich economies, despite their preferences towards clean air, and the presence of a technological sector that permits to substitute inputs in production, still achieve high levels of output, though associated with higher levels of emissions (CO_2, for example) to the atmosphere of the environment they live.

3.1 The Reduced Model

We can reduce the dimension of the canonical system given so far through the following convenient variable substitution

$$\frac{C}{K} = x$$
$$\frac{Y}{K} = m$$
$$\frac{C}{E} z^{\gamma} = q$$

and consequently end up with a new system in three dimensions, x, m, and q:

$$\dot{x} = \xi x - \left(\frac{1-\sigma}{\sigma}\right) mq + (\beta - 1)mx + x^2$$
$$\dot{m} = \eta m - \delta m^2 + \frac{1}{\beta\sigma} mq$$
$$\dot{q} = \xi q - \left(\frac{1-2\sigma}{\sigma}\right)\frac{mq^2}{x} + (1+\gamma)q^2 + \beta(1-\sigma)mq$$

where

$$\xi = \frac{(1-\sigma)\theta - \rho}{\sigma}$$
$$\beta = \left(\frac{\gamma}{1+\gamma}\right)\left(\frac{1-\alpha}{\sigma}\right)$$
$$\eta = \frac{\theta + \alpha\gamma(\varphi + \gamma)}{\gamma(1-\alpha)}$$
$$\delta = \frac{1+\alpha\gamma}{1+\gamma}$$

The associated Jacobian matrix at the steady state (x^*, m^*, q^*) is then

$$J^* = \begin{bmatrix} x^* + \left(\frac{1-\sigma}{\sigma}\right)\frac{m^* q^*}{x^*} & (\beta - 1)x^* - \left(\frac{1-\sigma}{\sigma}\right)q^* & -\left(\frac{1-\sigma}{\sigma}\right)m^* \\ 0 & -\delta m^* & \frac{1}{\beta\sigma} m^* \\ \left(\frac{1-2\sigma}{\sigma}\right)\frac{m^* q^{*2}}{x^{*2}} & \beta(1-\sigma)q^* - \left(\frac{1-2\sigma}{\sigma}\right)\frac{q^{*2}}{x^*} & (1+\gamma)q^* - \left(\frac{1-2\sigma}{\sigma}\right)\frac{m^* q^*}{x^*} \end{bmatrix}$$

Studying the behavior of this economy while converging to the steady state needs particular attention, especially if we want to control for the presence of undesired outcomes due to the rise of indeterminacy problems. To this end, we apply the neat Routh-Hurwitz criterion to the structure of eigenvalues associated with J^*, and easily verify that $trJ^* > 0$, and $DetJ^* < 0$. In this case, the sequence of signs becomes (−, +, ?, −), the only possibility is thus two positive and one negative eigenvalues. The interior steady state is therefore determinate, or saddle path stable.[16]

This is quite a piece of news when dealing with a Romer-type economy, whose uniqueness of the equilibrium trajectory, largely studied in several papers, showed the need for some parameters of the model to belong to a particular defined set. For example, Asada *et al.* [17] study the stability properties of a social planner version of the Romer model and several modifications of it, including the complementarity of different intermediate goods introduced by Benhabib *et al.* [18], and find the emergence of Hopf bifurcation points and stable periodic solutions [19,20].

More recently, Slobodyan [9] reconsiders a slightly simpler version of Benhabib *et al.* [18], and derives the restrictions on the parameter values necessary to obtain an interior steady state solution. He shows that Hopf bifurcation leading from determinate steady state to a completely stable one does not exist, but that indeterminate steady state can become absolutely unstable (explosive) through Hopf bifurcation.

In this light, we ought to make a deep investigation, by relaxing the assumption made upon the research success parameter, γ, and show that in case we allow it to become negative, some indeterminacy problems may finally arise, and thus complicate the possibility to attain a sustainable equilibrium solution either. The next section is devoted to this end.

3.2 A Numerical Analysis

Without any loss of generality, and for the sake of simplicity, in this section we analyze a simpler version of the model set above, where we constrain $\sigma = 1$. It is indeed like moving back to the *AH* model, where the structure of preferences implies a logarithmic utility function.

Firstly, let us consider the case $\gamma > 0$, then the Jacobian matrix easily reduces to

$$J^* = \begin{bmatrix} x^* & -\delta x^* & 0 \\ 0 & -\delta m^* & \frac{1}{\beta} m^* \\ -\frac{m^* q^{*2}}{x^{*2}} & \frac{q^{*2}}{x^*} & \rho \end{bmatrix}$$

[16] Since the eigenvalues of J^* are the solutions of its characteristic equation

with

$$trJ^{*} = 2\rho > 0$$

$$DetJ^{*} = -\frac{\rho m^{*}}{\beta x^{*}}\left[q^{*2} + \beta\delta x^{*2}\right] < 0$$

the system is still characterized by a two-change of sign (−, +, ?, −), which implies the local stability of the steady state solution.

In particular, stability of system (*S*) needs

$$\begin{aligned} & x^{*} = \rho + \delta m^{*} \\ & \eta + \frac{1}{\beta} q^{*} = \delta m^{*} \\ & \left[(1+\gamma)q^{*} - \rho\right]x^{*} + m^{*}q^{*} = 0 \end{aligned}$$

which implies, solving for m^{*}, the following quadratic equation

$$G(m) = am^{*2} - bm^{*} - c = 0$$

where

$$\begin{aligned} a &= \beta\delta\left[1 - (\beta - 1)(1+\gamma)\right] = \frac{\gamma(1-\alpha)(1+\alpha\gamma)(2+\alpha\gamma)}{(1+\gamma)^{2}} \\ b &= \frac{\left(\alpha\gamma^{2} + \theta\right)\left(2+\alpha\gamma\right)}{\left(1+\gamma\right)} + \frac{\rho\left(1+\alpha\gamma\right)}{1+\gamma}\left[1 - \gamma\left(1-\alpha\right)\right] \\ c &= \rho\left(\rho + \theta + \alpha\gamma^{2}\right) \end{aligned}$$

and given $a > 0$, and $c > 0$, this allows us to understand why there is only one possible positive solution for m^{*} in steady state, and the system is therefore locally stable, whatever the sign of b, as shown in Figure 3.

On the contrary, if we allow the research success parameter (i.e. the degree of pollution abatement), to fall below zero, $\gamma < 0$, then some unexpected economic outcomes may arise. In particular, whenever $-\frac{1}{\alpha} < \gamma < -1$, we conclude that either $a, b < 0$ or $c > 0$, whose graphic representation in Figure 4 clearly shows the presence of two positive solutions for $G(m^{*}) = 0$, and thus consequently signal the emergence of a multiplicity of equilibria.

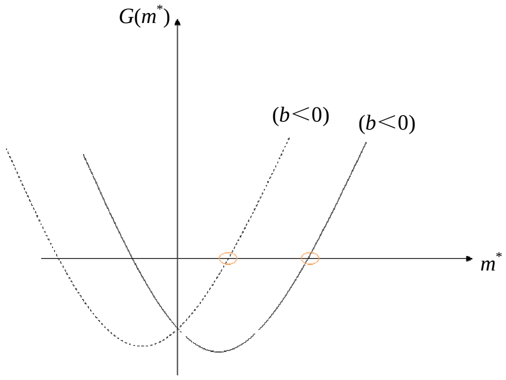


Figure 3. Unique equilibrium

To conclude, the new home-made inventories become a key indicator to achieve a long run sustainable equilibrium. On the one hand, in fact, we have shown so far that as long as an increase in the stock of knowledge is realized, i.e. γ is positive, the economy converges to a saddle path stable steady state. On the other hand, when the home-made research sector experiences a decreasing level of new inventories, which means a negative value for γ, the economy is likely to manifest some indeterminacy problems. In this case, a multiplicity of equilibria is therefore possible to arise, and consequently generate a situation where the economy might be trapped in a lower equilibrium solution. Other non-economic factors are thus possibly acting as a means for equilibria to differ along the transition path towards the steady state.

4. Concluding Remarks

A clear connection between growth and the quality of the environment is complex. Some elements of environmental quality appear to improve with growth; others worsen; still others exhibit deterioration followed by amelioration. Despite this evidence, most studies dealing with the impact of environmental policy on growth ignore the adverse effect of pollution on productivity. The state of the environment may worsen with time if concentrations of pollutants accumulate or if consumer tastes shift towards pollution-intensive goods. The opposite does occur if technological innovations make abatement less costly or if increasing awareness causes an autonomous shift in public demands for environmental safeguards. To this end, only if technological progress has to provide the means to reducing the over-exploitation of natural resources, a sustainable growth can be possible.

To bridge the existing gap we set up a model close to Aghion and Howitt [6] and examine the problem of sustainable growth in presence of dirty (i.e. polluting)

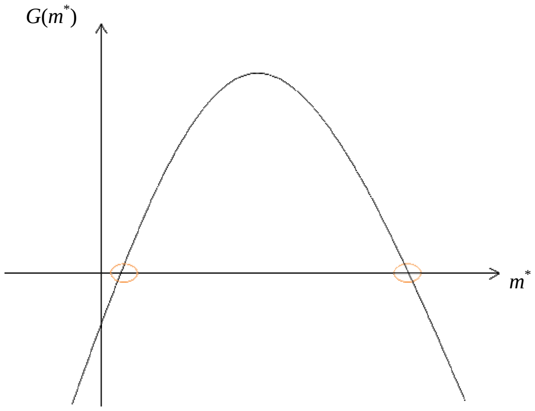


Figure 4. Multiple equilibria

production processes. We show that under certain conditions a sustainable growth is always attainable. The main difference with respect to our analysis regards the definition of a non-separable utility function (where both consumption and the environment are seen as two substitutes), and a particular technological sector where both home-made and outsourcing research activities are considered.

Particular attention has been devoted to the transitional dynamics of the model around the steady state, where the role of home-made research has turned out as a key device for stability and uniqueness of equilibrium solutions. Indeed, if the home-made research parameter is allowed to be negative, some indeterminacy problems arise, and multiple equilibria are likely to emerge. In this latter case, some non-economic factors become crucial in the solution of our decision making problem. This is consistent to how real economies nowadays behave, whenever their different cultural backgrounds impinge on the approach used to tackle the problem of a sustainable allocation of the available natural resources amongst generations. That is also likely to influence the development path of these economies towards the steady state, and eventually trap the systems into an unavoidable low equilibrium level.

REFERENCES

[1] D. I. Stern, “The rise and fall of the environmental kuznets curve,” World Development 32, pp. 1419−1439, 2004.

[2] S. Smulders, “Environmental policy and sustainable economic growth: An endogenous growth perspective,” De Economist 143, pp. 163−195, 1995.

[3] A. L. Bovenberg and S. Smulders, “Environmental quality and pollution-augmenting technological change in a two-sector endogenous growth model,” Journal of Public Economics 57, pp. 369−391, 1995.

[4] A. L. Bovenberg and S. Smulders, “Transitional impacts of environmental policy in an endogenous growth model,” International Economic Review 37, pp. 861−893, 1996.

[5] K. Pittel, “Sustainability and endogenous growth,” Edward Elgar, Cheltenham, 2003.

[6] P. Aghion and P. Howitt, “Endogenous growth theory (2nd edition),” MIT Press, Cambridge, Massachusetts, 1998.

[7] A. Grimaud and F. Ricci, “The growth-environment trade-off: horizontal vs vertical innovations,” Fondazione ENI Enrico Mattei Working Paper n, pp. 34−99, 1999.

[8] P. Schou, “Polluting non renewable resources and growth,” Environmental and Resource Economics 16 (2), pp. 211−227, 2000.

[9] S. Slobodyan, “Indeterminacy and stability in a modified Romer model,” Journal of Macroeconomics 29, pp. 169−177, 2007.

[10] A. Grimaud, “Pollution permits and sustainable growth in a schumpeterian model,” Journal of Environmental Economics and Management 38, pp. 249−266, 1999.

[11] S. I. Restrepo-Ochoa and J. Vazquez, “Cyclical features of the Uzawa--Lucas endogenous growth model,” Economic Modelling 21, pp. 285−322, 2004.

[12] M. A. Gomez, “Transitional dynamics in an endogenous growth model with physical capital,” Human Capital and R&D. Studies in Nonlinear Dynamics & Econometrics 9 (1), article 5, 2005.

[13] S. Smulders, “Economic growth and environmental quality,” In: Folmer, H., Gabel, L. (Eds), Principles of Environmental and Resource Economics, Edward Elgar, Chapter 20, pp. 602−664, 2000.

[14] S. Smulders, “Entropy, environment, and endogenous economic growth,” International Tax and Public Finance 2, pp. 319−340, 1995.

[15] I. Musu, “Transitional dynamics to optimal sustainable growth,” Fondazione ENI Enrico Mattei Working Paper n. 50.95, 1995.

[16] R. J. Barro and X. Salai-Martin, “Economic growth,” McGraw-Hill, New York, 1995.

[17] T. Asada, W. Semmler, and A. J. Novak, “Endogenous growth and the balanced growth equilibrium,” Research in Economics 52, pp. 189−212, 1998.

[18] J. Benhabib, R. Perli, and D. Xie, Monopolistic competition, indeterminacy and growth. Ricerche Economiche 48, pp. 279−298, 1994.

[19] L. G. Arnold, “Endogenous technological change: A note on stability,” Economic Theory 16, pp. 219−226, 2000.

[20] L. G. Arnold, “Stability of the market equilibrium in Romer's model of endogenous technological change: A complete characterization,” Journal of Macroeconomics 22, pp. 69−84, 2000.

(Edited by Vivian and Ann)

Appendix A

The current value Hamiltonian for the maximization problem is given by

$$H_c = \frac{(CE)^{1-\sigma}-1}{1-\sigma} + \lambda\left[A^{\alpha}(1-S_A)^{\alpha}K^{1-\alpha}z - C\right] + \qquad (1)$$
$$+\mu\left[\theta E - A^{\alpha}(1-S_A)^{\alpha}K^{1-\alpha}z^{1+\gamma}\right] + \vartheta\left[(\varphi+\gamma S_A)A\right]$$

where λ, μ and ϑ denote the costate variables associated with the accumulation of physical capital, natural capital and knowledge capital, respectively.

First order conditions can be written as:

1.a $\left[\frac{\partial H_c}{\partial C}=0\right]$:

$$\frac{\partial H_c}{\partial C} = C^{-\sigma}E^{1-\sigma} - \lambda = 0 \;\Rightarrow\; C^{-\sigma}E^{1-\sigma} = \lambda \qquad (2)$$

1.b $\left[\frac{\partial H_c}{\partial z}=0\right]$:

$$\frac{\partial H_c}{\partial z} = \lambda A^{\alpha}(1-S_A)^{\alpha}K^{1-\alpha} - \mu(1+\gamma)A^{\alpha}(1-S_A)^{\alpha}K^{1-\alpha}z^{\gamma} = 0 \qquad (3)$$

that is simply

$$\lambda = \mu(1+\gamma)z^{\gamma} \qquad (4)$$

1.c $\left[\frac{\partial H_c}{\partial S_A}=0\right]$:

$$\frac{\partial H_c}{\partial S_A} = -\lambda\alpha A^{\alpha}(1-S_A)^{\alpha-1}K^{1-\alpha}z + \mu\alpha A^{\alpha}(1-S_A)^{\alpha-1}$$
$$K^{1-\alpha}z^{1+\gamma} + \vartheta\gamma A = 0 \qquad (5)$$

or rather, using (A.3a) it becomes

$$\mu\gamma\alpha A^{\alpha}(1-S_A)^{\alpha-1}K^{1-\alpha}z^{1+\gamma} = \vartheta A \qquad (6)$$

Equation of motion for each costate variable is given by

$$\dot{\lambda} = -\frac{\partial H_c}{\partial K} + \lambda\rho \qquad (7)$$

$$\dot{\mu} = -\frac{\partial H_c}{\partial E} + \mu\rho \qquad (8)$$

$$\dot{\vartheta} = -\frac{\partial H_c}{\partial A} + \vartheta\rho \qquad (9)$$

and we can simply derive, by means of the conditions obtained above:

$$\frac{\dot{\lambda}}{\lambda} = -\left(\frac{\gamma}{1+\gamma}\right)(1-\alpha)A^{\alpha}(1-S_A)^{\alpha}K^{-\alpha}z + \rho$$
$$\frac{\dot{\mu}}{\mu} = -\frac{C^{1-\sigma}E^{-\sigma}}{\mu} - \theta + \rho \qquad (10)$$
$$\frac{\dot{\vartheta}}{\vartheta} = -\frac{\mu}{\vartheta}\gamma\alpha A^{\alpha-1}(1-S_A)^{\alpha}K^{1-\alpha}z^{1+\gamma} - \gamma S_A + \rho$$

by substituting out condition (A.4a) into the law of motion of ϑ, it follows

$$\frac{\dot{\vartheta}}{\vartheta} = \rho - (\varphi+\gamma) \qquad (11)$$

whereas, taking logs in (A.2) and differentiating, we have

$$\frac{\dot{\lambda}}{\lambda} = -\sigma\frac{\dot{C}}{C} + (1-\sigma)\frac{\dot{E}}{E} \qquad (12)$$

From condition (A.6) derives

$$\dot{\mu} = -C^{1-\sigma}E^{-\sigma} - \mu\theta + \mu\rho \qquad (13)$$

or, alternatively,

$$\frac{\dot{\mu}}{\mu} = -\frac{U_E}{\mu} - \theta + \rho \qquad (14)$$

given that $\frac{\partial U(\cdot)}{\partial E} = C^{1-\sigma}E^{-\sigma} = U_E$. But substituting out μ in the RHS, by means of (A.3a), we obtain

$$\frac{\dot{\mu}}{\mu} = -\frac{U_E}{\lambda}(1+\gamma)z^{\gamma} - \theta + \rho \qquad (15)$$

Since $\lambda = U_C$, from FOC, we have

$$\frac{\dot{\mu}}{\mu} = -\frac{U_E}{U_C}(1+\gamma)z^{\gamma} - \theta + \rho \qquad (16)$$

and finally, since equilibrium requires that $\frac{U_E}{U_C} = \frac{C}{E}$, it follows

$$\frac{\dot{\mu}}{\mu} = \rho - \theta - \frac{C}{E}(1+\gamma)z^{\gamma} \qquad (17)$$

- Arrow sufficiency theorem holds since the maximized Hamiltonian, evaluated along the optimal control variables, is concave in all the state variables, as we can simply check through the sing of the minors of the Hessian matrix, whose determinant implies the following sings

$$|H| = \begin{vmatrix} - & 0 & + \\ 0 & - & 0 \\ + & 0 & - \end{vmatrix} \qquad (18)$$

- hence, $|H_1|<0$, $|H_2|>0$, and $|H_3|<0$ *iff* $A>1$, that is the number of designs must necessarily be greater that one.
- Transversality conditions for a free terminal state hold for all shadow prices, and are given by

$$\lim_{t\to\infty}\lambda K e^{-\rho t} = \tilde{\lambda}e^{(1-2\sigma)gt}\tilde{K}e^{gt}e^{-\rho t} = \tilde{\lambda}\tilde{K}e^{-(2\sigma g+\rho)t} = 0$$
$$\lim_{t\to\infty}\mu E e^{-\rho t} = \tilde{\mu}e^{(1-2\sigma)gt}\tilde{E}e^{gt}e^{-\rho t} = \tilde{\mu}\tilde{E}e^{-(2\sigma g+\rho)t} = 0 \qquad (19)$$
$$\lim_{t\to\infty}\vartheta A e^{-\rho t} = \tilde{\vartheta}e^{(\rho-\varphi-\gamma)t}\tilde{A}e^{gt}e^{-\rho t} = \tilde{\vartheta}\tilde{A}e^{(g-\varphi-\gamma)t} = 0$$

- Where $\tilde{\lambda}$, $\tilde{\mu}$, $\tilde{\vartheta}$, and $\tilde{K}$, $\tilde{E}$, $\tilde{A}$, are the shadow prices and the state-values on the balanced growth path;

- Moreover, for free time t, we need to show that $\lim_{t\to\infty} H = 0$, which is always verified due to convergence towards zero of both the discounted utility function, $\lim_{t\to\infty} U(\cdot)e^{-\rho t} = 0$, and all the multipliers, as proved above.

Transitional dynamics of the problem can be studied by applying the Routh-Hurwitz criterion to the autonomous system

$$\dot{x} = \xi x - \left(\frac{1-\sigma}{\sigma}\right)mq + (\beta-1)mx + x^2$$

$$\dot{m} = \eta m - \delta m^2 + \frac{1}{\beta\sigma}mq$$

$$\dot{q} = \xi q - \left(\frac{1-2\sigma}{\sigma}\right)\frac{mq^2}{x} + (1+\gamma)q^2 + \beta(1-\sigma)mq$$

and the associated Jacobian matrix, evaluated along the steady state

$$J^* = \begin{bmatrix} J^*_{11} & J^*_{12} & J^*_{13} \\ J^*_{21} & J^*_{22} & J^*_{23} \\ J^*_{31} & J^*_{32} & J^*_{33} \end{bmatrix}$$

where

$$J^*_{11} = x^* + \left(\tfrac{1-\sigma}{\sigma}\right)\frac{m^*q^*}{x^*}$$

$$J^*_{12} = (\beta-1)x^* - \left(\tfrac{1-\sigma}{\sigma}\right)q^*$$

$$J^*_{13} = -\left(\tfrac{1-\sigma}{\sigma}\right)m^*$$

$$J^*_{21} = 0$$

$$J^*_{22} = -\delta m^*$$

$$J^*_{23} = \tfrac{1}{\beta\sigma}m^*$$

$$J^*_{31} = \left(\tfrac{1-2\sigma}{\sigma}\right)\frac{m^*q^{*2}}{x^{*2}}$$

$$J^*_{32} = \beta(1-\sigma)q^* - \left(\tfrac{1-2\sigma}{\sigma}\right)\frac{q^{*2}}{x^*}$$

$$J^*_{33} = (1+\gamma)q^* - \left(\tfrac{1-2\sigma}{\sigma}\right)\frac{m^*q^*}{x^*}$$

which implies consequently

$$trJ^* = 2\frac{m^*q^*}{x^*} + 2(1+\gamma)q^* > 0$$

and

$$DetJ^* = \left[x^* + \left(\frac{1-\sigma}{\sigma}\right)\frac{m^*q^*}{x^*}\right]\left\{-\delta(1+\gamma)m^*q^* + \delta\left(\frac{1-2\sigma}{\sigma}\right)\frac{m^{*2}q^*}{x^*} - \left(\frac{1-\sigma}{\sigma}\right)m^*q^* + \left(\frac{1-2\sigma}{\beta\sigma^2}\right)\frac{m^*q^{*2}}{x^*}\right\} + \left(\frac{1-2\sigma}{\sigma}\right)\frac{m^*q^{*2}}{x^{*2}}\left[\left(\frac{\beta-1}{\beta\sigma}\right)m^*x^* - \left(\frac{1-\sigma}{\beta\sigma^2}\right)m^*q^* - \delta\left(\frac{1-\sigma}{\sigma}\right)m^{*2}\right] < 0$$

which implies also

$$\left(\frac{1-2\sigma}{\sigma}\right)\frac{m^*q^{*2}}{x^{*2}}\left[\left(\frac{\beta-1}{\beta\sigma}\right)m^*x^* - \left(\frac{1-\sigma}{\beta\sigma^2}\right)m^*q^* - \delta\left(\frac{1-\sigma}{\sigma}\right)m^{*2}\right] < \left[q^* + \left(\frac{1-\sigma}{\sigma}\right)\frac{m^*q^{*2}}{x^{*2}}\right]\left\{\left[\delta(1+\gamma) + \left(\frac{1-\sigma}{\sigma}\right)\right]m^*x^* - \left(\frac{1-2\sigma}{\beta\sigma^2}\right)m^*q^* - \delta\left(\frac{1-2\sigma}{\sigma}\right)m^{*2}\right\}$$

since it is always verifiable that

$$\frac{1-\sigma}{\sigma} > \frac{1-2\sigma}{\sigma}$$

$$\frac{\beta-1}{\beta\sigma} < \delta(1+\gamma) + \left(\frac{1-\sigma}{\sigma}\right)$$

or rather

$$\left[\alpha(1-\alpha)\gamma^2 + (1+\gamma)\right]\sigma > 0$$

J. Serv. Sci. & Management, **2009, 2: 56-60**
Published Online March 2009 in SciRes (www.SciRP.org/journal/jssm)

Scientific
Research
Publishing

A Fuzzy Model for Evaluating Cultivation Quality of Talents of Software Engineering at the Campus Universities

Yongzhong Lu[1], Danping Yan[2], Bo Liu[1]

[1]School of Software Engineering, Huazhong University of Science and Technology, Wuhan 430074, P. R. China; [2]School of Public Administration, Huazhong University of Science and Technology, Wuhan 430074, P. R. China.
Email: hotmailuser@163.com

Received January 6th, 2009; revised February 9th, 2009; accepted March 2nd, 2009.

ABSTRACT

In order to measure the quality of talent cultivation at the school of software engineering, a quality evaluation model based on fuzzy theory is put forward. In the model, a three-layer architecture, which is composed of overall goal layer, second goal layer, and attribute layer, is set up. It places emphasis on the demand of talents with practicability and engineering in the field of software engineering. Then a case is used in the model to illustrate its effectiveness. The experimental results show that the model can comparatively better evaluate the quality of talent cultivation, reach the expected objective, and fulfill the practical demand. According to the model, a quality evaluation software system is developed while a rainfall lifecycle development model and Microsoft Visual C++ Development Studio are utilized.

Keywords: *software engineering, cultivation quality evaluation, fuzzy computing model*

1. Introduction

In order to fulfill the urgent social demands of software talents with high quality, practical experiences and comprehensive engineering skills in China, we have carried out a series of reform and innovation pertaining to the teaching contents and approaches, courses system, and management institution and operational mechanism. Up to now, we have come to deeply recognize that training talents of software engineering is similarly deemed to a item of talent production project. In the course of the teaching reform and innovation, it is significantly vital to lay emphasis on its training quality and effect which are the progress signpost in the forthcoming days. Generally speaking, it is rather difficult to measure the quality and effect of bringing up software talents quantitatively because they are closely related to numerous determinants [1,2]. Therefore, an accurate quality evaluation model about the training project of software talents at the universities is still not set up. Based on the social demands for software talents in China, we first put forward a qualitative model of quality evaluation of talent cultivation at the universities, and then exploit a fuzzy approach to give the quantitative computational results. Subsequently a case is used to testify its effectiveness. At last, a quality evaluation software system is developed while a rainfall lifecycle development model and Microsoft Visual C++ Development Studio are utilized.

2. A Fuzzy Quality Evaluation Model for Software Talents

We have referred to the generic ability evaluation standard of engineering graduates in UK, the USA and other European developed countries [3,4,5,6,7,8]. In addition, we have combined it with present practical situation at the campus schools and amended it properly. As a result, a quality evaluation model of training software talents is presented in Figure 1. In the model there are three layers: the top one is called overall objective layer and expressed by matrix A, the middle layer is called second objective layer and expressed by matrix B, and the lowest layer is called third attribute layer and expressed by matrix C, but it does not mean this layer is no importance. The corresponding statements are shown in Table 1.

3. A Fuzzy Evaluation Approach

It's quite difficult to get the exact values of the attributes in the model above. The fuzzy evaluation approach adapts to solve the problem well. Therefore it is used here to work out the solution to the problem. Its process is described as follows.

1) Establish the evaluation expert group

Different types of software experts are adopted to probe into the quality of training the software talents. They are usually composed of several experts such as field experts, senior managers, and users, and so forth. After the selection of evaluation expert group, a comment set is required to be determined. Supposing that the hierarchical rank of software products is classified into five levels which correspond to a comment set *V*: *V*= ("excellent", "good", "medium", "passed", "bad") =(v_1,v_2,v_3,v_4, v_5).

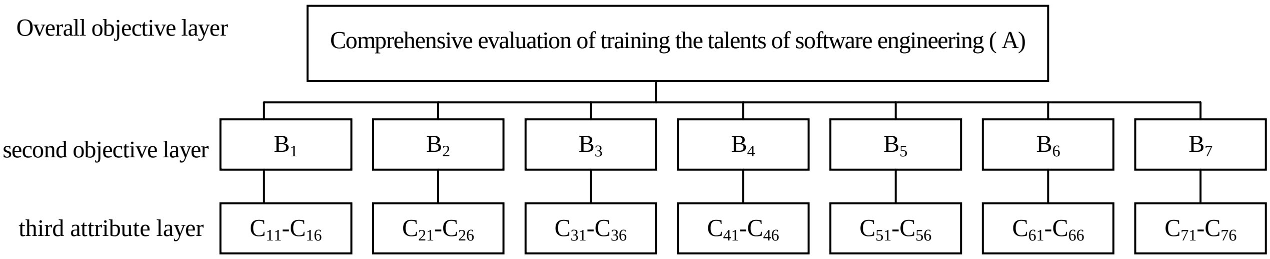


Figure 1. An evaluation model of training the talents of software engineering

Table 1. The generic statements corresponding to the Figure 1

1 Ability to exercise Key Skills in the completion of software engineering-related tasks at a level implied by the benchmarks associated with the following statements (B_1) a) Communication (C_{11}) b) Information Technology (C_{12}) c) Application of Number (C_{13}) d) Working with Others (C_{14}) e) Problem Solving (C_{15}) f) Improving Own Learning and Performance (C_{16})
2 Ability to transform existing software systems into conceptual models (B_2) a) Elicit and clarify client's true needs (C_{21}) b) Identify, classify and describe software engineering systems (C_{22}) c) Define real target software systems in terms of objective functions, performance specifications and other constraints (ie, define the problem) (C_{23}) d) Take account of risk assessment, and social and environmental impacts, in the setting of constraints (including legal, and health and safety issues) (C_{24}) e) Resolve difficulties created by imperfect and incomplete information (C_{25}) f) Derive conceptual models of real target software systems, identifying the key parameters (C_{26})
3 Ability to transform conceptual models into determinable models (B_3) a) Construct determinable models over a range of complexity to suit a range of conceptual models (C_{31}) b) Use mathematics and computing skills to create determinable models by deriving appropriate constitutive equations and specifying appropriate boundary conditions (C_{32}) c) Use industry standard software tools and platforms to set up determinable models (C_{33}) d) Recognise the value of Determinable Models of different complexity and the limitations of their application (C_{34})
4 Ability to use determinable models to obtain system specifications in terms of parametric values (B_4) a) Use mathematics and computing skills to manipulate and solve determinable models; and use data sheets in an appropriate way to supplement solutions (C_{41}) b) Use industry standard software platforms and tools to solve determinable models (C_{42}) c) Carry out a parametric sensitivity analysis (C_{43}) d) Critically assess results and, if inadequate or invalid, improve knowledge database by further reference to existing software systems, and/or improve performance of determinable models (C_{44})
5 Ability to select optimum specifications and create physical models (B_5) a) Use objective functions and constraints to identify optimum specifications (C_{51}) b) Plan physical modelling studies, based on determinable modelling, in order to produce critical information (C_{52}) c) Test and collate results, feeding these back into determinable models (C_{53})
6 Ability to apply the results from physical models to create real target software systems (B_6) a) Write sufficiently detailed specifications of real target software systems, including risk assessments and impact statements (C_{61}) b) Select production methods and write method statements (C_{62}) c) Implement production and deliver products fit for purpose, in a timely and efficient manner (C_{63}) d) Operate within relevant legislative frameworks (C_{64})
7 Ability to critically review real target software systems and personal performance (B_7) a) Test and evaluate real software systems in service against specification and client needs (C_{71}) b) Recognise and make critical judgements about related environmental, social, ethical and professional issues (C_{72}) c) Identify professional, technical and personal development needs and undertake appropriate training and independent research(C_{73})

2) Determine the single weights of the statements

AHP (Analytical Hierarchy Process) is adopted to figure out the weights of the statements. The detailed steps are followed below.

● According to the model above, a proper questionnaire is well-prepared for the experts. They determine the mutual weights among the statements in three layers. The weight matrix between overall objective layer *A* and second objective layer B_i is shown in Table 2. The matrix is usually called determinant matrix. We can obtain other determinant matrixes in the same way. Thereafter they fill out the comments about the attribute layer statements as Table 3.

● Construct the single determinant matrix

The AHP constructs the determinant matrix by terms of relationship among the statement items, and their proportional scales are among 1-9 [9]. Supposing that *A* represents the object set, *U* the evaluation item set, u_i (i=1,2,…,n) the evaluation item, and u_{ij} represents mutual weight between u_i and u_j (j=1,2,…,n), the determinant matrix is expressed below.

$$\begin{array}{ccccccc} U & u_1 & u_2 & \cdots & u_i & \cdots & u_n \\ u_1 & u_{11} & u_{12} & \cdots & u_{1j} & \cdots & u_{1n} \\ u_2 & u_{21} & u_{22} & & u_{2j} & & u_{2n} \\ \vdots & & \cdots & & \cdots & & \\ u_n & u_{n1} & u_{n2} & & u_{nj} & & u_{nn} \end{array} \qquad (1)$$

● Calculate the normalized weights of all evaluation items above

The geometric average method is used to gain the eigenvector corresponding to the most characteristic root $\lambda_{\max}$ of matrix *U* above. And it is normalized and shaped into the weights of all evaluation items. The detailed formula is following

$$W_i = (\prod_{j=1}^{n} u_{ij})^{1/n} \Big/ \sum_{i=1}^{n} (\prod_{j=1}^{n} u_{ij})^{1/n} \qquad (2)$$

where *i, j* = 1, 2 ,…, n. The result $W = (W_1, W_2, \cdots, W_n)^T$ is the above-mentioned eigenvector.

Table 2. Weight matrix of A and B

A	B_1	B_2	B_3	B_4	B_5	B_6	B_7
B_1							
B_2							
B_3							
B_4							
B_5							
B_6							
B_7							

Table 3. Subjection degrees about attribute layer statements

Attributer layer statements	Comment set				
	excellent	good	Medium	passed	bad
C_{11}					
C_{12}					
C_{73}					

● Consistency testing

Supposing that *U* is a matrix with n ranks, u_{ij} ($1 \le i \le n$, $1 \le j \le n$) is an element in *U*, if all elements of *U* have a property of transitivity, that is to say $u_{ij} \times u_{jk} = u_{ik}$, the matrix *U* is called a consistency matrix. A consistency matrix can be verified by the formula (3)

$$CR = CI / RI \qquad (3)$$

where *CR* is called the random consistency ratio of the determinant matrix, RI is called the average random consistency ratio of the determinant matrix, and *CI* is called the general consistency item which can be expressed by the formula (4)

$$CI = (\lambda_{\max} - n)/(n-1) \qquad (4)$$

where *n* is the rank of the determinant matrix. $\lambda_{\max}$ is decided by the following formulae (5) and (6)

$$\lambda_{\max} = \frac{1}{n} \sum_{i=1}^{n} \frac{(PW)_i}{W_i} \qquad (5)$$

$$PW = \begin{bmatrix} (PW)_1 \\ (PW)_2 \\ \cdots \\ (PW)_n \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1j} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2j} & \cdots & u_{2n} \\ & \cdots & & \cdots & & \\ u_{n1} & u_{n2} & \cdots & u_{nj} & \cdots & u_{nn} \end{bmatrix} \begin{bmatrix} W_1 \\ W_2 \\ \cdots \\ W_n \end{bmatrix} \qquad (6)$$

when *CR*<0.10, it can be concluded that the determinant matrix has a satisfactory property of consistency, that is to say that the distributed weights are proper, vice versa.

● Calculate the comprehensive weights

The distributed weights of the second objective layer to the third attribute layer are obtained by the formula (3). The distributed weights of the overall objective layer to the second objective layer is calculated by the formula (7)

$$W = \sum_{j=1}^{n} WB_j WC_{ij} \qquad (7)$$

where WB_j is the important weight of B_j (1<*j*<7) corresponding to A, and WC_{ij} is the important weight of C_{ij} corresponding to B_j. When B_j has no bearing with C_{ij}, WC_{ij} =0.

3) Determine the subjection degrees of the quality evaluation

When carrying out the evaluation of talent cultivation of software engineering, field experts, together with senior manager (policy-makers) and customers, give the decisive subjection degree according to the defined comment set above. It can explicitly be expressed by the subjection degree matrix R below

$$R = (r_{ij})_{m \times k} \qquad (8)$$

where r_{ij} is the percentage of regarding the i-th evaluation statement as the j-th comment class. And it is also expressed by $r_{ij} = d_{ij} / d$ where d_{ij} is the number of the members of drawing the conclusion that the i-th evaluation statement belongs to the j-th comment class, *d* is the total of the members, *m* is the number of the statements, and *k* is the evaluation rank.

4) Calculate the final evaluation result

After attaining the subjection degree matrix R, we calculate the comprehensive evaluation vector *S* of talent cultivation of software engineering. Then we adopt the Weighted Average Model of comprehensive evaluation− M (*,+) in order to consider all relevant factors appropriately and remain their information. The comprehensive evaluation vector *S* and the comprehensive evaluation result *P* are displayed in (9) and (10) respectively

$$S = W_c^a \times R \tag{9}$$

$$P = V \times S^T \tag{10}$$

In the formula (9), W_c^a is the comprehensive weights of third attribute layer C corresponding to overall objective layer *A*. As a result, the quality level of talent cultivation of software engineering at campus universities can easily be performed by the formula (10) and the task of quality evaluation of talent cultivation of software engineering is successfully completed.

4. Illustration

In order to testify the effectiveness of the presented model above, we take a practical case for example. Based on the model, we perform the demonstration in accordance with the following steps.

1) Calculate the single weights of the statements

The AHP is exploited to construct the single determinant matrixe as Table 2 and normalized by Formula (2). Then the consistency testing is done by Formula (3). If the *CR* is less than 0.1, the comprehensive determinant matrix is obtained by Formula (7). The two results are shown in Tables 4 and 5.

2) Calculate the subjection degree matrix

After the mutual weights of three layers are decided, 15 relevant members give their evaluation opinions to the quality of talent cultivation of software engineering with the aid of the comment set above. The subjection degree matrix $R_{30\times5}$ is gotten by Formula (8) and normalized into Formula (11).

$$R = \begin{bmatrix} 0.195 & 0.636 & 0.564 & 0.081 & 0.455 & 0.091 & 0.182 & \cdots \\ 0.455 & 0.564 & 0.345 & 0.273 & 0.273 & 0.345 & 0.455 & \cdots \\ 0.564 & 0 & 0.091 & 0.345 & 0.273 & 0.273 & 0.564 & \cdots \\ 0 & 0 & 0 & 0.091 & 0 & 0.091 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \end{bmatrix} \tag{11}$$

3) Calculate the comprehensive evaluation value

$$S = W_c^a \times R = (0.2175, 0.4635, 0.2123, 0.0415, 0.067) \tag{12}$$

$$P = V \times S^T = (5,4,3,2,1) \times (0.2175, 0.4635, 0.2123, 0.0415, 0.067)^T = 3.728 \tag{13}$$

From Formulae (12) and (13), we find that if the subjection degree is 0.2175, the quality is excellent; if the

Table 4. The single weights of the second objective layer B corresponding to the third attribute layer *C*

B_1(0.476)						B_2(0.266)	…
C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	…	…
0.299	0.141	0.105	0.168	0.127	0.160	…	…

Table 5. The comprehensive weights of the second objective layer *B* and the third attribute layer *C* corresponding to the overall objective layer *A*

B_1(0.476)						B_2(0.266)	…
C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	…	…
0.138	0.089	0.049	0.078	0.059	0.096	…	…

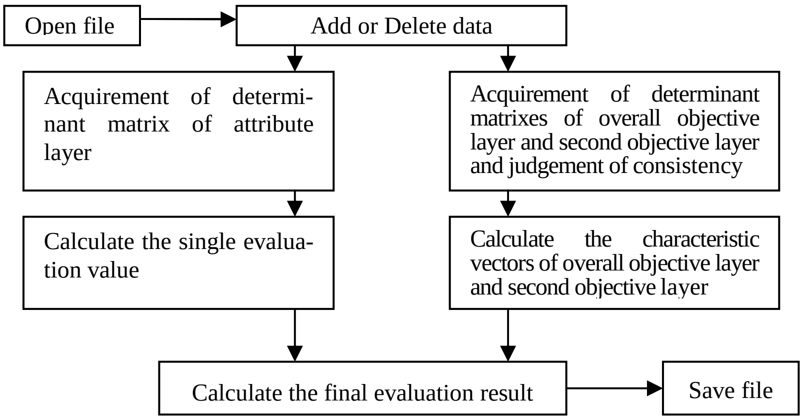
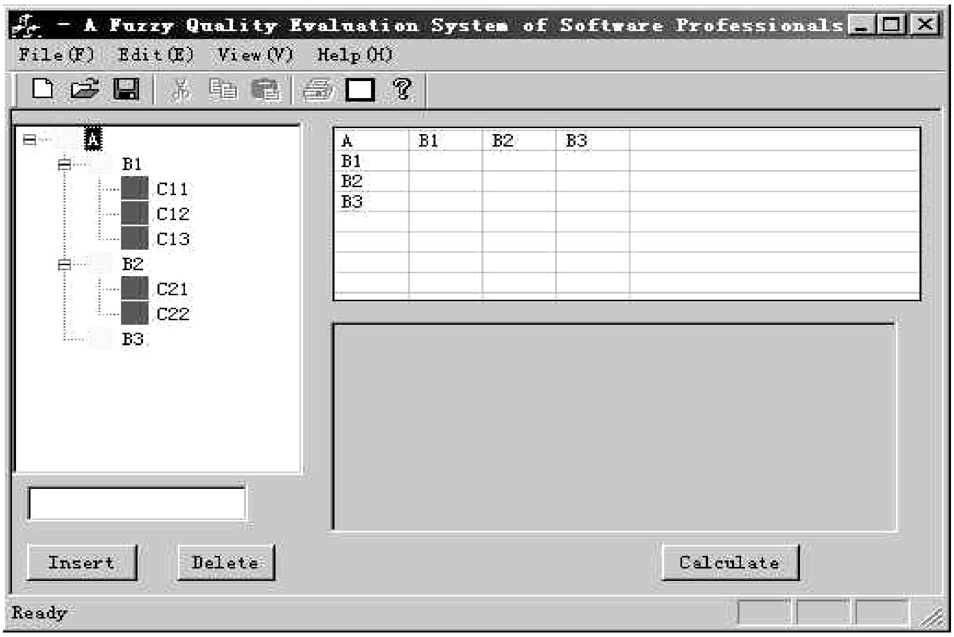

Open file
Add or Delete data
Acquirement of determinant matrix of attribute layer
Acquirement of determinant matrixes of overall objective layer and second objective layer and judgement of consistency
Calculate the single evaluation value
Calculate the characteristic vectors of overall objective layer and second objective layer
Calculate the final evaluation result
Save file

Figure 2. The system workflow



A	B1	B2	B3
B1			
B2			
B3			

Figure 3. The quality evaluation system

subjection degree is 0.4635, the quality is good; if the subjection degree is 0.2123, the quality is medium; if the subjection degree is 0.0415, the quality is passed; if the subjection degree is 0.067, the quality is bad. If $V = \{5,4,3,2,1\}$ is quantified, the comprehensive evaluation value is 3.728 and its final evaluation quality is “medium”.

5. Developing the Quality Evaluation System

The workflow of the quality evaluation system is described as Figure 2. In the figure, we divide the system into five modules which include Add or Delete module, Calculate the single evaluation value of certain attribute module, Consistency testing module, Calculate the characteristic vector module, and Calculate the final evaluation value module.

The quality evaluation software system is developed as Figure 3 while a rainfall lifecycle development model and Microsoft Visual C++ Development Studio are utilized.

6. Conclusions

Based on the quality evaluation model of talent training of software engineering, a fuzzy quality evaluation system of talent training of software engineering is developed. It can easily measure the quality level of talent cultivation of software engineering and provide a good evaluation platform for software talent cultivation. However, some aspects on the consistency testing and determinant matrix construction will be further addressed in the future.

7. Acknowledgement

The support from the Natural Science Foundation at Huazhong University of Science and Technology grant number 2007Q006B is gratefully acknowledged for this work by the authors.

REFERENCES

[1] J. S. J. Lin, “Study of university curriculum development for human resource management and the expectations of business managers,” International Journal of Business and Systems Research, 2(4), pp. 418–430, 2008.

[2] P. Xiao, Y. Cheng, and Y. Yang, “Higher education human resources development criterion inquisition,” Future and Development, 29(2), pp. 9–12, 2008.

[3] N. E. Gibis, “The SEI education program the challenge of teaching future software engineers,” Communications of ACM, 32(5), pp. 594–605, 1989.

[4] O. Hazzan, “The reflective practitioner perspective in software engineering education,” Journal of Systems and Software, 63(3), pp. 161–171, 2002.

[5] A. J. Cowling, “What should graduating software engineers be able to do?” IEEE Proceedings of the 16th Conference on Software Engineering Education and Training, pp. 88–98, 2003.

[6] D. T. Holt and Mitchell Crocker, “Prior negative experiences: their impact on computer training outcomes,” Computers & Education, 35(4), pp. 295–308, 2000.

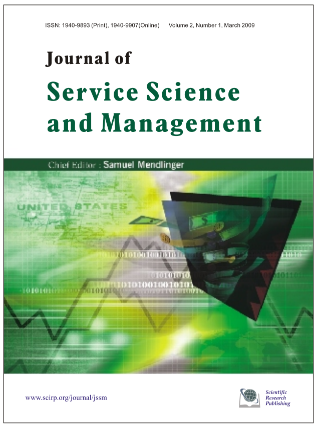
[7] M. M. Boulet, C. Dupuis, and N. Belkhiter, “Selecting continuous training program and activities for computer professionals,” Computers & Education, 36(1), pp. 83–94, 2001.

[8] D. Garlan, “Making formal methods education effective for professional software engineers,” Information and Software Technology, 37(5–6), pp. 261–268, 1995.

[9] Q. Wang, “Practical fuzzy mathematics,” Science and Technology Documentary Press, Peking, 1995.

(Edited by Vivian and Ann)

Call for Papers


ISSN: 1940-9893 (Print), 1940-9907 (Online) Volume 2, Number 1, March 2009
Journal of
Service Science
and Management
Chief Editor : Samuel Mendlinger
www.scirp.org/journal/jssm
Scientific Research Publishing

Journal of Service Science & Management (JSSM)

ISSN 1940-9893 (Print), 1940-9907 (Online)

www.scirp.org/journal/jssm/

JSSM is an international multidisciplinary journal with the emphasis laid on the service innovation in the global economy and entrepreneurship, the latest management technologies. It also explores the contributions of knowledge discovery and information management research and applications. The goal of this journal is to keep a record of the state-of-the-art research and promote the fast moving service science and management technologies.

Editor-in-Chief

Editorial Board (According to Alphabet)

Subject Coverage

All manuscripts must be prepared in English, and are subject to a rigorous and fair peer-review process. Accepted papers will immediately appear online followed by printed in hard copy. The journal publishes the highest quality, original papers included but not limited to the fields:

- **Service Science**
- **Business Intelligence**
- **Operational Research**
- **Computational Economics**
- **Financial Engineering**
- **Decision Support System**
- **Business Process Re-engineering**
- **Data Mining and Knowledge Discovery**
- **Innovation and Entrepreneurship**
- **Risk Management**
- **Quality Management**
- **Project Management**
- **Supply Chain Management**
- **Software Engineering Management**
- **Environment and Energy Management**
- **Knowledge Management and Semantic Web**
- **Information System Management**
- **Customer Capital Management**
- **Human Resources Management**

We are also interested in:
Short reports—Discussion corner of the journal:
2-5 page papers where an author can either present an idea with theoretical background but has not yet completed the research needed for a complete paper or preliminary data.
Case studies—Rather than present a classical paper.
Book reviews—Comments and critiques.

Notes for Intending Authors

Submitted papers should not have been previously published nor be currently under consideration for publication elsewhere. Paper submission will be handled electronically through the website. All papers are refereed through a peer review process. For more details about the submissions, please access the website.

Contact Us:

E-Mail: jssm@scirp.org

J. Serv. Sci. & Management, **2009, 2**
Published Online March 2009 in SciRes (www.scirp.org/journal/jssm)

CONTENTS

Volume 2 Number 1 **March 2009**

9771940989003 04