Quantitative Methods for Hospital Planning and Control

A thesis presented by

Luca Pontecorvi

in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Biomedical Engineering

UNIVERSITÀ CAMPUS BIO-MEDICO DI ROMA
Facoltà dipartimentale di Ingegneria

Coordinator
Prof. Giulio Iannello

Supervisor
Prof. Marco Papi

April 2017

Quantitative Methods for Hospital Planning and Control

A thesis presented by

Luca Pontecorvi

in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Biomedical Engineering

UNIVERSITÀ CAMPUS BIO-MEDICO DI ROMA
Facoltà dipartimentale di Ingegneria

Coordinator
Prof. Giulio Iannello

Supervisor
Prof. Marco Papi

April 2017

A Carlo

Contents

List of Figures

List of Tables

Abstract

The provision of hospital resources, such as beds, operating theatres and nurses, is a matter of considerable public and political concern and has been the subject of widespread debate. The political element of healthcare emphasises the need for objective methods and tools to inform the debate and provide a better foundation for decision-making. An appreciation of the dynamics governing a hospital system, and the flow of patients through it, point towards the need for sophisticated capacity models reflecting the complexity, uncertainty, variability and limited resources.

A common current practice is to plan and manage hospital capacities through a simple deterministic models using average patient flows, average needs, average length-of-stay, average duration of surgical operations etc. Average analysis can be misleading since the underlying distribution is not symmetric. To overcome such limit it is described the probability distribution of some of the most important proxies for measuring the consumption of hospital resources such as discharge rate, admission rate, number of hospitalized patients, and Length of Stay (LoS). While model proposed to describe the LoS is an innovative generalization of models previously applied in this area, the model for the description of the discharge and admission rate is borrowed from the financial mathematics. It is assumed that exist an analogy between the default of a financial institution and discharge of a patient. This approach come up with a simple and closed-form formula for the distribution function of the discharge rate and the admission rate. Moreover some risk metrics, used in financial mathematics, are applied in order to analysis the tail of the distributions.

In order to investigate the activities of the hospital departments, a deterministic analysis of numeric indicators is performed. Among the common measured clinical parameters, a robust metrics, characterizing the constituent entities and the best opportunity tools for the characterization of the results, have been identified. Using this approach is

provided an application in the evolution of the department. Particularly, the attention is focused on the evolving of the medical team.
Deterministic approached turns out to be not suitable for the development of decision support tools, mathematically speaking, a hospital corresponds to a complex stochastic system so that the common deterministic approach for planning and managing the system can be expected to be inadequate.
Hence it is also provided a methodological approach to optimize the hospital resource allocation based on stochastic dynamic programming (SDP). SDP approach is well-positioned to model these types of problems because of the explicitly sequential nature of the decision policies they produce. The aim is to reduce the probability of having a number of patients different from a fixed level over a define interval of time. It is shown that the optimal policy proposed performs better than an empirical policy.
In modern societies, the cost of healthcare are increasing year by year. The requirement is to cut costs without diminishing the quality of care. One solution is to increase efficiency; hospital need to plan their operations to use available resources in an optimal fashion. In order to analysed the relation between LoS, risk score, and costs, Real Option Approach (ROA) is applied. Physician has the right but not the obligation to discharge a patent if some efficiency conditions are not verified. In according financial yield curve models, a cost function is estimated and the results are compared whit value obtain from ROA.
It is also proposed an application of the Lotka-Volterra model and an extension of the Heston model.
The thesis has the following outline: Chapter 1 presents an introduction to healthcare systems and briefly discusses sources of fund and issues affecting healthcare quality and costs. It also highlights the importance of using quantitative models to analyse different healthcare delivery strategies and optimize costs. Chapter 2 focuses on healthcare financing systems in different countries and describes different methods of paying for healthcare providers. The strengths and weaknesses of the discussed methods are also pointed out. Chapter 3 introduces a generic framework for healthcare planning. This framework, encompassing 4 hierarchical levels of control and four managerial areas, is used to identify external and internal environmental characteristics affecting the organization of healthcare systems. Chapter 4 provides an overview of the main

mathematical theories used in subsequent chapters. These include stochastic differential equations, real option analysis, option pricing and Poisson processes. Chapter 5 introduces a statistical model to describe the length of stay of hospital patients. The proposed model overcomes some of the limitations of previous models by using a Phase-Type Gamma distribution which is able to capture the data characteristics in a more accurate way. The model is tested on a case study based on the Campus Bio-Medico hospital database. Chapter 6 introduces some quality indicators as a tool to evaluate health systems performance and quality. A dynamic stochastic optimization model is then proposed to optimize hospital bed occupancy. Three different models to describe patient discharge probabilities are also proposed and then used to evaluate the optimal policies. Chapter 7 introduces three financial-like models to describe the variable costs associated with patient hospitalization: a Nelson-Siegel model; a Black-Scholes model and a Cox-Ingersoll-Ross model.

The second part of the dissertation, Chapters 8 and 9, tackles different problems that the student has investigated during his doctoral studies and which are not related to healthcare system planning. Specifically, Chapter 8 describes a model to optimize the consumption of financial inspection resources for tax evasion by analysing the interaction between prevention/control activities and illegal behaviours. Chapter 9 proposes a new stochastic volatility model for the calibration of option prices.

Glossary

DRG Diagnosis Related Groups is a system to classify hospital cases into groups.

ES EuroSCORE (European System for Cardiac Operative Risk Evaluation) is a method of calculating predicted operative mortality for patients undergoing cardiac surgery.

GDF Guardia di Finanza is an Italian law enforcement agency under the authority of the Minister of Economy and Finance MEF and part of the Italian armed forces.

GDP Gross Domestic Product is a monetary measure of the value of all final goods and services produced in a period (quarterly or yearly).

LoS Length of Stay is a term to describe the duration of a single episode of hospitalization. Inpatient days are calculated by subtracting day of admission from day of discharge.

NPV Net Present Value is the difference between the present value of cash inflows and the present value of cash outflows.

OECD Organisation for Economic Co-operation and Development is an international economic organisation of 34 countries, founded in 1961 to stimulate economic progress and world trade.

PPP Purchasing Power Parity. PPP is used to compare the income levels in different countries. The theory aims to determine the adjustments needed to be made in the exchange rates of two currencies to make them at par with the purchasing power of each other. In other words, the expenditure on a similar commodity must be same in both currencies when accounted for exchange rate. The purchasing power of each currency is determined in the process.

ROA Real Option Analysis is a mathematical approach that calculates the value of options associated with a decision.

STEEPLED is an abbreviation for the following external environment factors: Social factors, Technology, Economic factors, Environmental factors, Political factors, Legislation, Ethical factors, and Demographics.

VaR Value at Risk is a measure of the risk of investments. It estimates how much a set of investments might lose, given normal market conditions, in a set time period such as a day.

VLoS is the smallest time such that the probability of discharge is greater than a fixed level.

WHO World Health Organization is a specialized agency of the United Nations that is concerned with international public health.

1

Introduction

This Chapter presents an introduction to healthcare systems and briefly discusses sources of fund and issues affecting healthcare quality and costs. It also highlights the importance of using quantitative models to analyse different healthcare delivery strategies and optimize costs.

Healthcare systems are a prominent focus for national leaders and policy makers in most countries today. This fact reflects concerns about the availability of necessary health services for the population, as well as about the efficiency and costs of current health systems in delivering those services. The degree of importance of this issue in any given country is directly related to the size of the healthcare system relative to the national economy. Nearly all decisions of policy makers about national health systems must be based on the quantitative aspects of the options available, and the impact of any decisions taken. A quantitative description of the current health system and projection of the impact of changes is also critical. Hence, the ability to quantitatively describe health systems as well as to create a range of *what if* scenarios based on new directions for those systems is increasingly important in all countries.

The WHO defines health as *a state of complete physical, mental and social well-being, and not merely the absence of disease or infirmity* [3]. People's health is influenced by the quality of the air they breathe, the cleanliness of the water they drink, the types of food they consume, their hygiene, their habitat and their environment. All these factors are related to the economic situation of individuals and nations. Health generally deteriorates most where national economies are unable to generate adequate incomes or provide stable social systems, infrastructure and services (including primary healthcare), or where the environment and use of natural resources are poorly managed. The most critical healthcare needs are not only determined directly by the prevalent disease burden, but also indirectly by general conditions that either cause or prevent the introduction and transmission of disease.

Every country has an epidemiological profile that determines its healthcare needs, measured by changes in such factors as life expectancy, the number of inhabitants living in urban and peri-urban settlements, birth rates, mortality rates, survival rates of infants and mothers, the extent of non-communicable diseases, the degree of exposure to diseases, as well as epidemics, disability and mental disorders.

Changes in economic and living conditions have had a profound effect in changing healthcare needs in most countries. The evolution of medical knowledge and biomedical technology, the escalating proportions of elderly people and the different perceptions of health explained the increase of the healthcare costs. Table 1.1 shows that in the last decade there is an increasing of the life expectancy and a reduction of mortality rate. Data also show that this behaviour is common in different countries. One result

	Life expectancy[1]		**Mortality rate[1]**		**Percentage people over 65[4]**	
	2000	**2013**	**2000**	**2013**	**2000**	**2013**
Australia	80	83	77	61	12.35%	14.36%
Brazil	71	75	183	147	5.05%	7.33%
Canada	79	82	81	66	12.55%	15.25%
Germany	78	81	94	71	16.20%	20.95%
India	62	66	239	201	4.41%	5.38%
Israel	79	82	79	56	10.02%	10.83%
Italy	80	83	76	54	18.08%	21.59%
Japan	81	84	73	62	17.18%	25.01%
Norway	79	82	85	61	15.17%	15.69%
Russian Federation	65	69	312	232	12.44%	13.16%
Saudi Arabia	73	76	112	80	2.91%	2.76%
Sierra Leone	39	46	536	433	2.50%	2.67%
Spain	79	83	86	63	16.64%	18.09%
Switzerland	80	83	77	53	15.30%	17.60%
Turkey	70	75	150	109	5.99%	7.27%
United States of America	77	79	114	102	12.32%	14.00%

Table 1.1: Life expectancy, adult mortality rate (probability of dying between 15 and 60 years per 1000 population), and population ages 65 and above (% of total)

of social and economic development has been an increase in the percentage of people over 65 years of age in most countries. Even if the elderly are healthier today than in the past, they are still more susceptible to non-communicable diseases (such as cardiovascular conditions, cancer, diabetes or tobacco-related diseases) and to physical or mental disability. As a result, they require nursing care services to replace traditional but rapidly eroding family-support structures. Urbanization, an ongoing and increasing phenomenon, also changes health needs because it requires capacity increases and structural changes in the medical and public health infrastructure.
The provision of healthcare services is perhaps one of the largest and most complex industries worldwide. As one of the essential necessities to sustain life, it faces the consequences of increasing demand in times of limited financial resources and competing social needs. Providing the appropriate medical care involves decision-making in terms of planning and management of healthcare resources. All countries, rich and poor, struggle to raise the funds required to pay for the health services their populations need or demand (which is sometimes a different matter). No country, no matter how rich, is able to provide its entire population with every technology or intervention that may improve health or prolong life. But while rich countries health systems may face budget limitations, often exacerbated by the dual pressures of ageing populations and shrinking workforces, spending on health remains relatively high. Many richer countries will also need to raise additional funds to meet constantly evolving health demands, driven partly by ageing populations and the new medicines, procedures and technologies being developed to serve them. A key aspect of this complex issue is the diminishing workingage population in some countries. Dwindling contributions from income taxes or wage-based health insurance deductions (payroll taxes) will force policy-makers to consider alternative sources of funding. There are three main ways to raise additional funds or diversify sources of funding:

Make health a higher priority in existing spending, particularly in a governments budget Governments in the Americas, the European and Western Pacific Regions, on average, allocate more to health than the other regions. African countries as a group are increasing their commitment to health as are those in the European and Western Pacific Regions. In South-East Asia, the relative priority given to health fell in 2004 and 2005, but is increasing again, while governments in the WHO Eastern

Mediterranean Region have reduced the share allocated to health since 2003. Some of the variation across regions can be explained by differences in country wealth. Generally, health accounts for a higher proportion of total government spending as countries get richer. Chile is a good example, having increased its share of government spending on health from 11% in 1996 to 16% a decade later during a period of strong economic growth [5]. But a countries relative wealth is not the only factor at play. Substantial variations across countries with similar income levels indicate different levels of government commitment to health.
There are several reasons countries do not prioritize health in their budgets, some fiscal, some political, some perhaps linked to the perception in ministries of finance that ministries of health are not efficient. In addition, the budget priority governments give to health reflects the degree to which those in power care, or are made to care, about the health of their people. Dealing with universal health coverage also means dealing with the poor and the marginalized, people who are often politically disenfranchised and lack representation.
This is why making health a key political issue is so important and why civil society, joined by eminent champions of universal coverage, can help persuade politicians to move health financing for universal coverage to the top of the political agenda [6]. Improving efficiency and accountability may also convince ministries of finance, and increasingly donors, that more funding will be well used.

Find new or diversified sources of domestic funding The international community has taken several important steps since 2000 to raise additional funding to improve health in poor countries. They are outlined briefly here because they offer ideas for countries to raise domestic funds as well. These developments have helped pinpoint new sources of funds and maintained the momentum for increased international solidarity in health financing. However, discussions on innovative financing have so far ignored the needs of countries to find new sources of domestic funds for their own use: low- and middle-income countries that simply need to raise more and high-income countries that need to innovate in the face of changing health needs, demands and work patterns. Not all the options will be applicable in all settings, and the income-generating potential of those that are will also vary by country. Though we do make some suggestions about the likely level of funding that could be raised at the country level [7]. For example so-called solidarity taxes on specific goods and services are another promising option, offering a proven capacity to generate income, relatively low administration costs and sustainability. With political support, they can be implemented quickly. Introducing mechanisms that involve taxes can be politically sensitive and will invariably be resisted by particular interest groups. A tax on foreign exchange transactions, for example, may be perceived as a brake on the banking sector or as a disincentive to exporters/importers. Meanwhile, so-called sin taxes have the advantage of raising funds and improving health at the same time by reducing consumption of harmful products such as tobacco or alcohol. Studies in 80 countries have found that the real price of tobacco, adjusted for purchasing power, fell between 1990 and 2000. Although there have been some increases since 2000, there is great scope for revenue raising in this area, as advocated by the WHO Framework Convention on Tobacco Control [8].

Increase external financial support Prior to the global economic downturn that started late in 2008, development assistance for health from richer to poorer countries was increasing at a robust rate. Countries saw funding from external sources rise on average from 11.1% of their total health expenditures in 2000 to 17.3% in 2013 [1]. According to the databases maintained by the OECDs development assistance committee, government commitments for health reported by bilateral donors jumped from about US$ 4 billion in 1995 to US$ 17 billion in 2007 and US$ 20 billion in 2008. This may represent a significant underestimate given that the committee database does not capture all contributions from non-OECD governments, such as China, India and some Middle-Eastern countries; reports data for only a limited number of multilateral institutions; and does not collate funds provided by key private players in the health domain and nongovernmental organizations. A recent study suggested that the combined contribution from all these sources might have been about US$ 21.8 billion, almost US$ 5 billion greater than reported to the OECD in 2007 [9]. However, in at least four key ways, the outlook for aid-recipient countries is less positive than these numbers might suggest. First, despite the increase in external support, total health expenditures remain pitifully low insufficient to ensure universal access to even a basic set of health services in many countries. Second, even though external funding has increased substantially, about half of the countries reporting their development assistance disbursements to OECD are on track to meet the targets they have committed to internationally (for overall development, including health) [2]. The other countries are failing to meet their pledges, some by a long way. Slow progress towards fulfilling these commitments comes at a huge human cost. Third, the development assistance for health numbers reported above represent commitments; actual disbursements are lower. In addition, some of the funds that donors report as disbursed do not arrive in recipient countries for them to spend. Sometimes considerable proportion of aid is devoted to so-called technical cooperation. Finally, concerns have also been expressed recently that some of the aid arriving in countries is subject to spending constraints. Macroeconomic and monetary targets set for inflation and the level of foreign exchange reserves are based on a concept of prudent macroeconomic management. A more sustainable option is for external partners to reduce the volatility of their aid flows. This would, at a minimum, allow government budget ceilings in health to be relaxed and more aid could be used to improve health.

1.1 Optimize healthcare cost

Healthcare is important for all countries. Receiving quality care when needed to ensure a long and healthy life is a basic tenet of life. Securing the correct solution that balances financing and quality care is now the subject of great debate. Many reforms have sought to ensure that all population have access to affordable quality care and insurance. Providers struggle to deliver quality care at a reasonable price. Insurers struggle to provide a fair and affordable funding solution. Government officials are trying to balance the needs of all participants to ensure a viable, long term, stable solution.
Healthcare issuers are operating in a volatile and fluid environment today. Increased demand, combined with increasing healthcare costs, have forced healthcare issuers to manage administrative expenses. With evolving complexity of product offerings, understanding costs become essential information for managing the business. All of these issues are creating challenges and opportunities for the healthcare issuers. Increasing the efficiency will allow insurers to maintain stability in their operations and rates. It will also allow them to react quickly to changes in regulation and demand. Receiving accurate information and using adequate model will allow to make better decisions regarding their operations [10], [11] and [12].

The key emphasis here is on developing health support systems that are capable of providing, analysing, evaluating and distributing information necessary for health management. Information is indispensable for modelling as well as for progress in policy-making and implementation activities. Developments in science, technology and clinical practice have resulted in the less costly but equally effective ambulatory treatment of a growing number of conditions that were formerly treated on an inpatient basis. Meeting healthcare needs requires the elaboration of strategies to improve health. Since resources are invariably limited, priorities should be established to reflect the needs of the population, the general health situation, and national health priorities as revealed by information analysis.

For health systems to be sustainable, they must have a reliable source of finance and ensure equal access to a basic level of health services. Models can help in analysing various national healthcare delivery strategies, such as patient allocation strategy and bed planning model, their effects on health expenditure, and perhaps even their effects on aggregate indicators of health status. However, no model can select a strategy. Decision makers should use modelling to evaluate more clearly the implications of their policy options and decisions.

2

Economic environment of the health sector

This Chapter focuses on healthcare financing systems in different countries and describes different methods of paying for healthcare providers. The strengths and weaknesses of the discussed methods are also pointed out.

A country's health system has an effect on its economic growth and labour productivity. This is because the health status of the population affects the work force, which is an important factor in determining its productivity. Productivity, in turn, has a strong influence on economic growth. Many factors relating to healthcare services will have a significant impact on morbidity and mortality patterns - these include the type of services provided, the quantity and quality of these services, the method of their distribution, and the extent of their accessibility by the population.

Financing is a critical element determining the quantity, distribution and quality of health services. Financing also has an enormous effect on operational efficiency, and the ability to provide necessary health services according to need rather than ability to pay. Therefore, governments have considerable means to influence the health status of their citizens through their choice of health financing policies.

healthcare goods and services are exchanged on the healthcare market. There is ample reason to believe that the market is distorted by a variety of factors, most importantly by the asymmetry of information between consumers and providers and by the need to insure against potentially substantial healthcare costs. Third-party payment systems, together with inevitably under-informed consumers, lead to a market situation in which providers of care have a dominant influence on the volume and structure of demand. The individual need for health services is highly uncertain, which implies that the demand for services is also uncertain, and possibly difficult to predict. Demand is only an approximate function of need; it also depends on the availability and affordability of services.

The most important aspect for the modeller regarding health expenditure is that it is highly *income elastic*. It is obviously easy to persuade consumers with incomes increasing in real terms to spend more and more on healthcare. This microeconomic relationship obviously aggregates into a macroeconomic relationship. The level of national economic development is also obviously a significant determinant of the level of national health expenditure. Table 2.1 shows the national health expenditure as a percentage of GDP, and per capita GDP in PPP US dollars. This Table shows that national policy, commitment to health, political and historical factors and other influences results in substantial differences, even among countries within the same region or at similar levels of development. Several observations can be made about overall health expenditure at the national level:

1. There are great differences in health expenditure between countries. In Europe total per capita health expenditures vary from less than PPP $2098 per capita in Greece to over PPP $6468 in Switzerland [1].

2. Total spending depends on economic output. Although there are differences, as some countries spend more on health than others, GDP levels are directly related to the level of health expenditures.

3. As national income rises, the proportion of total health expenditure accounted for by the public sector increases.

4. The lower public spending on health in developing countries is the result of a lower share of government spending devoted to health.

5. External funding in the form of foreign aid is a major source of financing for health in Africa and some countries in other regions.

6. The level of insurance coverage varies widely. In Americas, over 60 per cent of the population is covered by health insurance. In Africa, the percentage is in the single digits (see Table 2.2).

The level of expenditure on healthcare in a country depends on many factors, including the need for services (i.e. the existence of medical problems for which there is a useful intervention), the ability to afford care, the efficiency of the financing mechanisms in mobilizing resources, and the efficiency and costs of the health sector. It is important to understand that the level of health spending is a matter of policy and choice, although it is constrained by the wealth of the country and other priorities competing for use of the country's resources. One main factor that determines the choices made on the level of health spending is income. The proportion of a country's income spent on healthcare tends to be around 6.12% in countries with very low yearly GDP per capita (yearly GDP per capita < \$3600), around 6.78% for medium-income countries (yearly GDP per capita > \$3600 and < \$12000), and up to 7.58% in the richest countries (yearly GDP per capita > \$12000).

2.1 National healthcare financing system

All national healthcare systems are pluralistic, which means they consist of a variety of schemes or subsystems. These schemes are distinguished by their pattern of financing and delivery, the scope of their benefits, and their population coverage. In principle, combinations of the following characteristics are possible:

- Population coverage
- Benefit range
- Benefit delivery
- Financing

An important distinction must be made between the financing and delivery of health services. Services may be provided in both the public and private sectors. Thus, it is possible that services would be financed by the public sector, but provided by the private sector. The provision of health services by the public sector may occur at government health facilities or at social health insurance facilities. Private-sector healthcare providers include hospitals, practitioners, and pharmacies operated by non-governmental organizations (NGOs) or the not-for-profit sector, as well as those operated for profit. Health services may also be provided directly by employers. However,

with respect to financing sources and population coverage, only a few types of healthcare schemes are typically dominant components of national healthcare systems.
National Health Service systems are characterized by public tax financing, a mix of public and private delivery, full or almost full scope (range of benefits) and universal coverage of the population. Social insurance schemes combine public contribution financing, public and/or private delivery, usually a full benefit range and less than full population coverage. Private insurance schemes in many countries combine partial population coverage, less than full scope, private financing and private delivery. A private insurance subsystem or scheme, however, may also operate alongside a dominant public sector scheme. Health services may be financed through public or private expenditure. Traditionally, the primary source of financing for the health sector in many countries has been the government, although other sources of financing have more recently increased in importance. Where health services are paid for with taxes or compulsory insurance (through individual and/or employer contributions), they are counted as public expenditure. Private expenditure includes payments by individuals and employers which are generally voluntary, with the rare exception of mandatory healthcare savings schemes. Funding for recurrent operating and long-term development costs for health services may come from any of three primary sources:

1. Public sources of financing
    - Direct government contributions to finance the provision of health services, through national or local budgets
    - Social health insurance, sponsored by the government (may be mandatory)
    - Community financing schemes for health services
2. Private sources of financing
    - Direct payment by patients (fee-for-service or other household expenditure)
    - Private, voluntary health insurance (indirect individual and employer payments)
    - Employer-based health insurance
    - Payments by community and other voluntary local organizations that finance health services

- healthcare savings schemes, in which individuals save a stipulated amount each month to cover healthcare costs in case of need
- Mutuals or cooperative-based insurance schemes

3. External financing

   - Donor monies for health services (institutional aid, foreign aid or development loans)

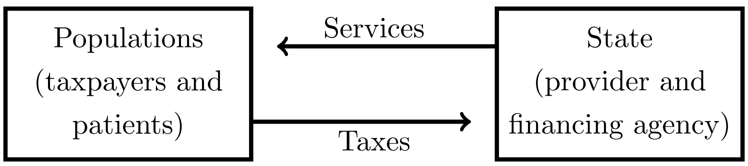
It is possible to categorize the various types of healthcare systems along the main characteristics of their financing dimension. One type of financing will generally dominate a national system. The relative importance of each funding source varies dramatically among countries and within regions. Table 2.1 provides examples of the public/private mix in healthcare financing in countries. The Table reflects the pluralistic structures of virtually all national healthcare systems. In the United States, for example, which is generally regarded as having a privately financed healthcare system, the public share nevertheless exceeds 50 per cents of total expenditures. Meanwhile, the healthcare systems of European Union, which are generally dominated by public financing, still show on average 25 per cent share of private financing in total healthcare expenditure.
In sum, the relative proportion of public spending and private spending in national health expenditure varies widely between regions and countries. The pluralism of national health financing systems is an important factor, there is interaction between the various subsystem in every national healthcare financing system. The reason is simple: many providers operate in both the public and the private segment of the market, for example, physicians in public hospitals might also see private patients.

2.2 Public Healthcare financing system

Publicly financed healthcare systems remain the backbone of healthcare financing in most countries. There are basically two approaches to the public financing of healthcare: the *public health service approach* (including national health service and public service health systems), and the *social health insurance approach*. In the former type of system, the public sector is both the financing agent and provider of health services. In the latter, the government is the financier, but may or may not be the provider. These are the two main paradigms to which the modelling tools developed later in this volume can be applied.

There are two crucial characteristic of public service health system. First on the microeconomic level, *the money does not follow the patient*. Provider units are financed by budgetary allocation, and not paid for specific services rendered to specific patients. Second, on the macroeconomic level, there is no relationship between the amount of taxes paid be the population and the volume and structure of services delivered each year. This is because the financial resources of the health sector come from general revenues, and are determined by an annual budget procedure in which the health sector has to compete with alternative uses for resources. Therefore, increased overall tax payments do not automatically result in increased allocations to the health sector.

The public health service model is a **defined-income scheme**, where the volume and structure of services available to the public is largely defined by whatever income the public health service can obtain from the general budget. Resources are allocated to different categories of care in different regions through the bureaucratic budgeting mechanism. The structure of the flow of money and services is portrayed graphically below.



2.2.1 The outflow of funds: Paying for service delivery

Healthcare financing schemes generally have two main categories of expenditure: administrative costs and medical benefits. In the latter category, medical services may be purchased from providers (i.e. provided indirectly), or delivered in the scheme's own facilities (i.e. provided directly). All healthcare financing systems incur administrative costs, even if they are not directly visible. Particularly in public service healthcare systems, the administrative cost structure of the scheme may lack transparency. These schemes are administered by ministries of health, and costs for the provision of services are incurred in various institutions and facilities, such as the supervising ministry, the health section of the ministry of finance, other ministries, and a whole hierarchy of healthcare facilities, each with its own administration. In social insurance schemes, these administrative costs are (or at least should be) visible in the scheme accounts.

The level of administrative costs varies greatly, and depends on the complexity of the benefit package, the system of reimbursement for services, the means of contribution collection, and the maturity of the respective scheme (young schemes generally require more administrative input than mature schemes). Administrative costs can be a frequent source of waste, and when they exceed 5-10 per cent of total benefit outgo, an in-depth audit of administrative practice is always warranted. The bulk of the funds of a social health insurance scheme should be spent on the purchase or direct delivery of benefits. A critical determinant of benefit expenditure is the way providers of care are paid by the scheme. We explore some alternatives in provider payment systems, which are largely independent of the type of financing system. Virtually all of these systems could be applied within national health service systems. Payment mechanisms determine the amount of financial flows from third party payers or patients (or both) to providers of care, in exchange for healthcare services. Essentially, the mechanisms define two critical items:

- the unit or basket of service for which a provider is paid per individual payment
- the price of the unit or basket.

The units or baskets used as a basis for payment exhibit a wide variety of aggregation. Sometimes they include all services rendered by a provider in a given period of time, such as one month in the case of payment by salary, or three months (or more) in the case of providers receiving budgets, such as hospitals. At the other end of the spectrum, payment may be for a single act performed by a professional, for example for an injection under a fee-for-service system. Prices might be implicitly or explicitly negotiated, they might be related to costs, or they might reflect other objectives such as deterring or encouraging utilization. They may also be linked to the quantity of units or baskets of services provided. Table 2.3 lists the most common *pure* forms of payment mechanisms and their most common applications, in an ascending order of aggregation: Payment methods are usually an element in the *service contract* between third-party payers and providers, but contracts generally regulate more than just payment issues. There may be various combinations of payment systems and relationships between payers and providers. One example of a mixed payment arrangement would be a doctor who receives a salary, but who is an employee either of the payer (e.g. the social insurance scheme) or of an institution which provides services to the payer (e.g.

government healthcare units). If this doctor is employed directly by the payer, the scheme may contain additional rules and incentives, apart from the payment system, to ensure that quality and efficiency standards are met. For example, there may be rules on the minimum number of patients to be treated in a given period of time, a maximum average cost per patient, or rules regarding professional advancement. If the doctor were employed by a provider that supplies services to the payer (such as a hospital), the payer would have to develop other means of influencing the provider, which in turn would have to develop internal procedures to ensure that standards are maintained.
Pure payment methods are rarely found in healthcare systems. In most cases, several payment methods are combined to form a more or less complex payment system. There are two general categories of payment systems:

- In an indirect delivery system, different payment methods can be used at various *levels*. For example, where an insurance scheme pays a public hospital, and the hospital pays individual providers (e.g. doctors), the hospital might be paid by budget or by daily charge, while the doctors are paid by salary or on a fee-for-service basis.
- Combined methods are also found within a single level of contractual relations. For example, a budget may be paid for primary care, but special services not covered by the budget may be paid through additional fees, according to a fee schedule. Fee-for-service payments may also be restricted by a budgetary cap. Per diem charges in hospitals may only cover current variable costs, while investments are reimbursed through lump-sum payments, possibly split between different payers.

2.2.2 Methods of paying healthcare providers

Combined payment systems are used to avoid the apparent weaknesses of pure payment methods. Since an understanding of the mechanisms of provider payment systems is of crucial importance to the modeller, the main strengths and weaknesses of pure provider payment methods are discussed.

Fee-for-service payments have three main characteristics:

1. Definition of a list of services.
   The initial problem in a fee-for-service system Is the description of the units of services for which the providers are individually paid. These are listed In an official list of services. This list can either be defined through negotiation between providers' organizations and healthcare funds, or administered by an independent authority. The list must be updated regularly, in line with developments in medical technology and practice as well as in consideration of the financial resources available. However, frequent or numerous changes may complicate the handling of payment procedures.

2. Determination of the price of a specific service unit.
   The prices for units of care may be the result of negotiations, set by an independent authority or. In extreme cases set directly by the social insurance schemes. The list of services and corresponding fees is called a *fee schedule*.

3. Definition of special rules and restrictions.
   Each fee schedule must be accompanied by instructions for its use, defining specific aspects of delivery, billing and payment. There may be quantity restrictions, e.g. limits on the number of units paid per treatment, per case or for a period of time. There may also be regulations concerning the conditions for payment or non-payment of fees. An annex to the fee schedule might list services which are not paid for if billed simultaneously (such as a consultation fee and fee for a specific treatment), or services that are only paid for certain categories of patients (such as children). Other services might be paid only if billed in connection with a specific diagnosis, or only if approved by the fund's management.

There are at least three central disadvantages of a fee-for-service payment system.

1. Creating and maintaining a fee schedule is complex and requires considerable administrative capacity, as do payment calculations and billing control.

2. Fee-for-service payments encourage a greater quantity of care.

3. Inadequate absolute or relative prices in a fee schedule might create distortions in medical practice, e.g. by moving provider behaviour towards high-profit items on the fee schedule (which is often the case for laboratory and diagnostic services).

One advantage of the fee-for-service method is that it will hardly lead to an under supply of services. However, international experience has shown that fee-for-service systems require very sophisticated calculations if these calculations are not exact, resources are wasted and desired distributive effects are not achieved. Generally, a fee-for-service system requires supplementary mechanisms, such as total expenditure ceilings and quantity restrictions.

Case payment may be made to providers (e.g. to doctors or health centres in primary care) on a fixed basis (or flat-rate-per-case payment system) per patient contact, or according to specific fees for the treatment of specific illnesses, as classified by a list of diagnoses.[1] Under this type of payment system, morbidity risk is born by the payer, while the provider bears the risk of expenditure per case. If the rate per case is uniform, total expenditure (for a category of services, such as treatment by GPs) is the product of the number of cases and the per case rate. If the rate varies by diagnosis, expenditure is also determined in part by the morbidity structure. The risk of varying treatment costs per diagnosis is then borne by the provider.
When looking at provider incentives, the case payment method has crucial advantages over the fee-for-service method. Providers cannot influence their income by supplying "too many", or "incorrect", or "inefficient" service packages. Their net income can be maximized only by minimizing treatment costs with respect to the given morbidity structure.
The case payment method, however, has at least two disadvantages. First, there is no incentive for providers to ensure a minimum standard of quality, since quality will usually increase costs. Thus, strict quality control procedures are necessary. An element of competition between providers might be helpful to ensure quality. If patients can choose between providers, the providers are forced to ensure a certain standard demanded by the patients. Competition between providers as a quality-enhancing mechanism has

[1]The first major payment system based on diagnoses (DRGs or diagnosis-related groups) was developed by researchers at Yale University during the 1970s. The system has been in use for the reimbursement of hospital care for Medicare patients in the United States since October 1983.

clear limitations, however, due to the patients' lack of information and limited ability to pass an educated judgement on the quality of care they are receiving. Thus, additional quality control mechanisms must be installed by the payer. Second, adverse selection of patients can be a problem. If primary care is paid according to a flat-rate structure, providers have an incentive to transfer patients to specialists or hospitals. This risk can be mitigated, for example, by including the average costs of specialist care and (if necessary) fees for specialist care in the per case payment. The unit providing primary care would then pay the specialist. Similarly, hospital and drug expenditures may be paid by the primary care provider.

Payment through fees related to diagnoses are a sophistication of the crude case payment method. Prerequisites for a functioning diagnosis-related payment system are a precisely defined system of diagnosis classification and a limit on the number of items on the fee schedule. The larger the number of items on the schedule, the greater room providers have to influence the diagnosis of a case to change their income. Diagnosis-related fee systems might be successfully applied in hospitals, if the diagnosis is made before the patient is transferred to the hospital. Even in this case, however, alternative or additional diagnoses may emerge during the patient's hospital stay.

Per diem fees (daily charges) are used in hospitals, either to pay for the entire service package delivered per day of stay, or for restricted packages complemented by additional charges for special services. This system has two main problems. First, a hospital can (and, according to experience, tends to) influence the length of a patient's stay. Generally, the marginal cost at the end of a hospital stay - the cost produced by an additional day - is less than the average daily cost. If charges are the main source of funds, the daily charge must cover at least the average daily costs. The hospital therefore makes a (marginal) profit by extending the average length of stay at the end of a treatment. Additional contractual regulations are necessary to eliminate this incentive, e.g. by instituting a restriction on the number of billable days, in total or per case.

A second problem is that the daily rate must be negotiated. If the rate is calculated by the hospital - even by a non-profit hospital - the hospital might finance inefficient equipment or treatment procedures. The healthcare scheme needs a basis for its own calculations, with at least the following information:

1. A classification of contracted hospitals by level of care (ranging from general secondary care facilities to tertiary care facilities such as university hospitals).

2. A minimum infrastructure standard for each class of hospital (e.g. equipment, patient / staff ratios, or minimum number of departments).

3. A calculation scheme providing guidance on which per diem fees are considered reasonable for each class of hospital, including variables such as infrastructure, wages, and price levels in the area.

The cost structure of an individual hospital is difficult to calculate or determine from the "outside", especially for insurance schemes which might not even have access to tax data or other government data sources.

Capitation payment method is generally used in primary care, especially for private providers (general practitioners). Under this system, a flat-rate fee (perhaps differentiated by age and sex of the covered persons) is paid for every covered person enrolled with an individual provider. This fee covers all care that covered persons require during a defined period of time. Generally, covered persons can choose from among more than one provider. (If they were allocated to providers by an administrative procedure, there would be no difference between capitation and budget payments.) In a capitation system, morbidity risk is borne by the providers, at least in the short run. The system can calculate a fixed amount of expenditure per insured person and per period. Economic disadvantages of this method are in general the same as under a flat-rate-per-case payment system. Given a defined list of patients, the net income of providers is maximized by minimizing their production costs. Thus, quality control (e.g. on the equipment used in provider facilities) and competition among providers are necessary to ensure a minimum quality standard. The capitation method is often combined with bonus and fee-for-service elements, in order to steer provider behaviour to less costly health-promoting activities, such as preventive care.

Bonus payments are global, flat-rate payments made to providers for executing specific duties, or as reimbursement for the purchase and operation of a particular piece of equipment. They are often used to complement capitation methods.

Budgeting under this method, a generally prospective budget is fixed for a provider unit (usually a hospital). The budget is meant to cover all expenses over a defined period (usually one year). Under certain limiting conditions (e.g. with respect to unexpected changes in the caseload), the provider can maintain possible profits, but must also shoulder potential losses. Budgets are usually renegotiated annually.
Morbidity risk as well as variance in treatment costs are borne to a large extent by the provider, and the cost of the scheme is fairly predictable in the long run. The potential for manipulating the mechanism is limited to the renegotiation process. Quality control mechanisms are necessary under budgeting systems to ensure minimum standards of quality.

Salary is usually paid for a certain period of time and a certain number of working hours. The income of a salaried provider, therefore, does not depend on the volume or structure of services provided, or the number of patients treated in a given period of time. To a large extent, it does not even directly depend on the quality of treatment.
The provider is paid for supplying labour. The employer is responsible for demanding services, i.e. for ensuring that the provider is employed during working hours. The employer might also decide on treatment patterns and intensity, as well as on the technology employed.
In fact, employers normally do not have the means to control the provider's behaviour completely. Providers, like all professionals, have room for discretion with respect to the services they supply, following their own preferences and judgement. Salaried providers normally do not have financial incentives to treat more than a minimum number of patients, or to achieve more than a minimum standard of quality. However, the employer might influence the provider's behaviour through a system of evaluation and promotion, or by offering bonus payments. Ethics and professional reputation might also be variables that figure prominently in the provider's behaviour. As a result, the influence of the salary payment system on quality is ambiguous. Table[1] 2.4 shows the incentive effects generally attributed to the various systems discussed above, with respect to several variables of crucial importance for the modeller. Providers are

[1]Refers to first-level direct effects. ++ is a strong incentive to increase activity, + is an incentive to increase activity, 0 is neutral incentive or not applicable, − is an incentive to decrease activity, −− is a strong incentive to decrease activity, and ? is inconclusive incentive effect.

assumed to attempt to achieve maximum (or target) income. No hypothesis is made with respect to non-financial behavioural incentives.
Under a fee-for-service mechanism, there is an incentive for providers to maximize total services and to prefer services with a high profit margin. Under the case payment and daily charge methods, providers will normally attempt to maximize the number of paid units (cases or days of stay), and to avoid costly treatment. The effect of these methods on the quality of care is ambiguous, as efficient early treatment or secondary preventive measures might be substituted for long-term treatment in the case of daily charge payments.
Bonus payments are generally used to spur certain activities. The size of the effect of bonus payments is difficult to identify as they vary based on the purpose for which the bonus is paid.
Under the budget and salary payment methods, there is no first-level direct incentive for providers to increase quantities or the intensity of care. Rather, the opposite is true. Long-term strategic considerations, however, might result in certain incentives to increase the number of patients for competitive reasons. The number of cases or patients is often used to demonstrate the demand for the services of the facility or of an individual provider.

According to the healthcare characteristics describe above, is clear that the use of decision support systems represent a critical aspect in healthcare financing system. In order to face with complex dynamic system that change constantly, optimization models managing the complexity of the problem are required. Health system has several crucial characteristics that make fundamental the optimization of available resources. Considering that provider units are financed by budget allocation and that there is a week relationship between the amount of resources allocated and the volume and the quality of service delivered, is reasonable to assume that optimizing the consumption process is a proper strategy to provide a better service.

	Total expenditure on health as % of GDP		**General government expenditure on health as % of total expenditure**		**Private government expenditure on health as % of total expenditure**		**Per capita total expenditure on health (PPP int.$)**	
	2005	**2014**	**2005**	**2014**	**2005**	**2014**	**2005**	**2014**
Australia	8.45%	9.42%	66.97%	67.04%	33.03%	32.96%	3031	4357
Brazil	8.27%	8.32%	41.51%	46.04%	58.49%	53.96%	899	1318
Canada	9.57%	10.45%	70.24%	70.93%	29.76%	29.07%	3469	4641
Germany	10.52%	11.30%	76.13%	76.99%	23.87%	23.01%	3384	5182
India	4.28%	4.69%	26.49%	30.04%	73.51%	69.96%	123	267
Israel	7.44%	7.81%	59.29%	60.85%	38.73%	39.15%	1829	2599
Italy	8.71%	9.25%	76.31%	75.61%	23.69%	24.39%	2587	3239
Japan	8.18%	10.23%	81.37%	83.59%	18.63%	16.41%	2491	3727
Norway	8.89%	9.72%	83.54%	85.49%	16.46%	14.51%	4317	6347
Russian Federation	5.21%	7.07%	61.98%	52.20%	38.02%	47.80%	616	1836
Saudi Arabia	3.42%	4.68%	72.49%	74.52%	27.51%	25.48%	1181	2466
Sierra Leone	12.25%	11.09%	21.85%	16.99%	78.15%	83.01%	128	224
Spain	8.12%	9.03%	72.37%	70.88%	27.63%	29.12%	2229	2966
Switzerland	10.86%	11.66%	59.46%	66.00%	40.54%	34.00%	4027	6468
Turkey	5.45%	5.41%	67.84%	77.45%	32.16%	22.55%	625	1036
United States of America	15.15%	17.14%	44.36%	48.30%	55.64%	51.70%	6741	9403

Table 2.1: Health expenditure as a share of GDP and per capita GDP in 16 countries [1].

	2005	2014
Australia	23.60%	25.36%
Brazil	35.53%	49.70%
Canada	42.26%	43.40%
Germany	38.12%	38.80%
India	1.19%	2.54%
Israel	16.36%	26.42%
Italy	3.67%	3.71%
Japan	13.16%	14.79%
Norway	8.22%	8.12%
Russian Federation	8.24%	3.50%
Saudi Arabia	11.84%	22.32%
Sierra Leone	0.59%	0.24%
Spain	18.93%	15.07%
Switzerland	22.16%	25.75%
Turkey	*nan*	*nan*
United States of America	63.56%	64.20%

Table 2.2: Private Insurance as % of Private Health Expenditure [1]

Payment mechanism	**Definition of unit**	**Paying the**		
		Hospital	**Health center**	**Individual provider unit**
Fee for service schedule	Single act	✓	✓	✓
Case payment	Patient cases, according to a fee schedule	✓	✓	✓
Daily charge	Patient day	✓		
Flat rate allowance	For certain investments	✓	✓	✓
Capitation	All potential service for one person during a defined period	✓	✓	✓
Salary	A period of work (month)			✓
Budget	All service provided in a given period		✓	✓

Table 2.3: Pure payment methods and their application.

Payment mechanism	**Quality services**		**Structurel composition of service provided**	
	Number of patients	Number of billable units of care	Number of acts · per units · per patients ·per period	Cost-boosting substitutions of high-profit for low profit services
Fee-for-service	+	++	++	++
Case payment	0	++	−	+
Per diem fee	+	++	−	−
Flat rate	?	0	?	0
Capitation	++	0	−	−−
Salary	−	−	−−	+
Budget	−	0	−−	−−

Table 2.4: Incentive structure of payment mechanisms. The incentive effects generally attributed to the various system with respect to several variable of crucial importance.

3

A generic framework for healthcare planning and control

This Chapter introduces a generic framework for healthcare planning. This framework, encompassing 4 hierarchical levels of control and four managerial areas, is used to identify external and internal environmental characteristics affecting the organization of healthcare systems.

The provision of healthcare services is perhaps one of the largest and most complex industries worldwide. As one of the essential necessities to sustain life, it faces the consequences of increasing demand in times of limited financial resources and competing social needs. Providing the appropriate medical care involves decision making in terms of planning and management of healthcare resources.

We describe a framework for healthcare planning and control that integrates all managerial areas involved in healthcare delivery operations and all hierarchical levels of control. It is applicable broadly, to an individual department , an entire healthcare organization, and to a complete supply chain of cure and care providers. The framework can be used to identify and position various types of managerial problems, to demarcate the scope of organization interventions, and to facilitate a dialogue between clinical staff and managers.

Healthcare planning and control lags far behind with respect to other fields in which planning and control procedure are applied [13]. Common reasons stated in the literature include:

- healthcare organizations are professional organizations which often lack cooperation between, or commitment from, involved part. These groups have their own, sometimes conflicting and objectives [14], [15].

- Due to the state of information systems in healthcare, crucial information required for planning and control is often not available [16] . Although DRG and electronic health record systems have spurred the need for financial and clinical information management systems, these systems tend to be poorly integrated with operational information systems. This lack of integration is impeding the advance of integrated planning and control in healthcare, both organization-wide and between organizations [17], [18].

- Since large healthcare providers such as hospitals generally consist of autonomously managed departments, managers tend not to look beyond the border of their department , and planning and control is fragmented [19], [18].

- The Hippocratic Oath taken by doctors forces them to focus on the patient at hand, whereas planning and control addresses the entire patient population, both within and beyond the scope of an individual doctor [20], [21].

- While healthcare managers are generally dedicated to provide the best possible service, they lack the knowledge and training to make the best use of the available resources [16].

- As healthcare managers often feel that investing in better administration diverts funds from direct patient care [16], managerial functions are often ill defined, over looked, poorly addressed, or functionally dispersed.

The proposed framework serves as a tool to structure and break down all functions of healthcare planing and control. In addition, it can be used to identify **planning and control problems** an to demarcate the scope of organization interventions. It is applicable broadly, from an individual hospital department to an entire hospital, or to complete supply chain of care providers. The framework facilitates a dialogue between clinical staff and managers to design the planning and control mechanisms. These mechanisms are necessary to translate the organizations objectives into effective and efficient healthcare delivery processes. It covers all managerial areas involved in

healthcare delivery operations and all levels of control, to ensure completeness and coherence of responsibilities for every managerial area.
According to [13] we propose a four-by-four generic framework for healthcare planing and control which spans four hierarchical levels of control, and four managerial areas.

3.1 Managerial Areas

Most existing framework in the literature focus on the managerial area [19], [18], [22]. In the proposed framework is included the following managerial areas for the healthcare planning and control:

Medical planning The role of engineers/ process planners in manufacturing is performed by clinicians in healthcare. We refer to healthcare version of *technological planning* as medical planning. Medical planning comprises decision making by clinicians regarding for example medical protocols, treatments, diagnoses, and triage. It also comprises development of new medical treatments by clinicians. The more complex and unpredictable the healthcare processes, the more autonomy is required for clinicians. For example, activities in acute care are necessarily planned by clinicians, whereas in elective care (e.g. ambulatory surgery), standardized and predictable activities can be planned cent rally by management.

Resource capacity planning Resource capacity planning addresses the dimensioning, planning, scheduling, monitoring, and control of renewable resources. These include equipment and facilities (e.g. MRI s, physical therapy equipment , bed linen, sterile instruments, operating theatres, rehabilitation rooms) , as well as staff.

Materials planning Materials planning addresses the acquisition, storage, distribution and retrieval of all consumable resources/materials, such as suture materials, prostheses, blood, bandages, food, etc. Materials planning typically encompasses functions like warehouse design, inventory management and purchasing.

Financial planning Financial planning addresses how an organization should manage its costs and revenues to achieve its objectives under current and future organizational and economic circumstances. Since healthcare spending has been increasing steadily [2], market mechanisms are being introduced in many countries as an incentive to encourage cost efficient healthcare delivery [23]. An example is the introduction of DRG, which enables the comparison of care products and their prices. As healthcare systems differ per country, so does financial planning in healthcare organizations. As financial planning heavily influences the way the processes are organized and managed, we include this managerial area in our framework. For example, Wachtel and Dexter [24] argue that in the US, the tactical allocation of temporary expansions in operating theatre capacity should be based on the contribution margin of the involved surgical specialities. This criterion is not likely to be used in countries with a non-competitive healthcare system, such as the UK or the Netherlands. Financial planning in healthcare concerns functions such as investment planning, contracting (with e.g. healthcare insurers) , budget and cost allocation, accounting, cost price calculation, and billing.

3.2 Hierarchical decomposition

Decision making disaggregates as time progresses and information gradually becomes available. We build upon the *classical* hierarchical decomposition often used in manufacturing planning and control, which discerns strategic, tactical, and operational levels of control [25]. We extend this decomposition by discerning between **offline** and **online** on the operational level. This distinction reflects the difference between **in advance** decision making and **reactive** decision making. We explain the resulting four hierarchical levels below, where the tactical level is explained last. The tactical level is often considered less tangible than the strategic and operational levels.
In this first stage we do not explicitly give the decision horizon length for any of the hierarchical planning levels, since these depend on the specific characteristics of the application. An emergency department for example inherently has shorter planning horizons than a long-stay ward in a nursing home.

Strategical level Strategic planning addresses structural decision making. These decisions are the bricks and mortar of an organization [26]. It involves defining the organization mission, and the decision making to translate this mission into the design, dimensioning, and development of the healthcare delivery process. Inherently, strategic planning has a long planning horizon and is based on highly aggregated information and forecasts. Examples of strategic planning are resource capacity expansions, developing and/ or implementing new medical protocols, forming a purchasing consortium, a merger of nursing homes, and contracting with health insurers.

Offline operational level Operational planning (both *offline* and *online*) involves the short-term decision making related to the execution of the healthcare delivery process. There is low flexibility on this planning level, since many decisions on higher levels have demarcated the scope for the operational level decision making. The adjective *offline* reflects that this planning level concerns the in advance planning of operations. It comprises the detailed coordination of the activities regarding current (elective) demand. Examples of offline operational planning are: treatment selection, appointment scheduling, nurse rostering, inventory replenishment ordering, and billing.

Online operational level The stochastic nature of healthcare processes demands for reactive decision making. *Online* operational planning involves control mechanisms that deal with monitoring the process and reacting to unforeseen or unanticipated events. Examples of online planning functions are: triaging, add- on scheduling of emergencies, replenishing depleted inventories, rush ordering surgery instrument sterilization, handling billing complications.

Tactical level In between the strategic level, which sets the stage (regarding e.g. location and size), and the operational level, which addresses the execution of the processes, lies the tactical planning level. We explain tactical planning in relation to strategic and operational planning.
While strategic planning addresses structural decision making, tactical planning addresses the organization of the operations/execution of the healthcare delivery process. In this way, it is similar to operational planning, however, decisions are made on a longer planning horizon. The length of this intermediate planning horizon lies somewhere between the strategic planning horizon and operational planning horizon. Following the concept of hierarchical planning, intermediate, tactical planning has more flexibility than operational planning, is less detailed, and has less demand certainty. Conversely, the opposite is true when compared to strategic planning. For example, while capacity

is fixed in operational planning, temporary capacity expansions like overtime or hiring staff are possible in tactical planning. Also, while demand is largely known in operational planning, it has to be (partly) forecasted for tactical planning, based on (seasonal) demand, waiting list information, and the *downstream* demand in care pathways of patients currently under treatment. Due to this demand uncertainty, tactical planning is less detailed than operational planning. Examples of tactical functions are admission planning, block planning, treatment select ion, supplier select ion and budget allocation.

3.3 Framework for health planning and control

Integrating the four managerial areas and the four hierarchical levels of control shapes a four-by- four positioning framework for healthcare planning and control. While the dimensions of the framework are generic, the content depends on the application at hand. The framework can be applied anywhere from the department level (for example to an operating theatre department) to organization-wide, or to a complete supply chain of care providers. Depending on the context, the content of the framework may be very different. Above scheme shows the content of a general framework that can be applied to a general hospital as a whole.

	Medical planning	Resource capacity planning	Materials planning	Financial planning
Strategic	Research, development of medical protocols	Case mix planning, capacity dimensioning, workforce planning	Supply chain and warehouse design	Investment plans, contracting with insurance companies
Tactical	Treatment selection, protocol selection	Block planning, staffing, admission planning	Supplier selection, tendering	Budget and cost allocation
Offline operational	Diagnosis and planning of an individual treatment	Appointment scheduling, workforce scheduling	Materials purchasing, determing order sizes	DRG billing, cash flow analysis
Online operational	Triage, diagnosing emergencies and complications	Monitoring, emergency coordination	Rush ordering, inventory replenishing	Billing complications and changes

The content of the framework should be accommodated to the context of the applica-

tion. We discern the **external** and the **internal** environment characteristic.
The way healthcare organizations are organized is perhaps influenced by its **external environment**. For example a *STEEPLED* analysis (an extension of *PESTEL* [27]) can be done to identify external factors that influence healthcare planning and control, now or in the future. STEEPLED is an abbreviation for the following external environment factors:

- Social factors e.g. education, social mobility, and religious attitudes.
- Technology e.g. medical innovation and transport infrastructure.
- Economic factors e.g. change in health finance system.
- Environmental factors e.g.ecological and recycling.
- Political factors e.g. change of government policy and privatization.
- Legislation e.g. business regulations and quality regulations.
- Ethical factors e.g. business ethics, confidentiality, and safety
- Demographics e.g. graying population, life expectancy, and obesity.

These factors largely explain the differences amongst countries in the management approach of healthcare organization.
The **internal environment** characteristics are scoped by the boundaries of the organization. This involves all characteristics that affect planning and control regarding for example patient demand (e.g. variability complexity arrival intensity, medical urgency, recurrence), organizational culture and structure.
Hospital length of stay (**LoS**) is considered to be a reliable and valid proxy for measuring the consumption of hospital resources [28]. The average length of stay in hospitals (ALoS) is often used as an indicator of efficiency [2]. All other things being equal, a shorter stay will reduce the cost per discharge and shift care from inpatient to less expensive post-acute settings. The ALoS refers to the average number of days that patients spend in hospital. It is generally measured by dividing the total number of days stayed by all inpatients during a year by the number of admissions or discharges. Day cases are excluded. The indicator is presented both for all acute care cases and for childbirth without complications. Another characteristics that effect the healthcare

	2013	
	Average length of stay	**Discharge number**
Australia	6.5	17240
Brazil	nan	nan
Canada	7.5	8438
Germany	7.6	25224
India	nan	nan
Israel	5.3	16118
Italy	6.8	12377
Japan	17.2	nan
Norway	5.5	16924
Russian Federation	nan	nan
Saudi Arabia	nan	nan
Sierra Leone	nan	nan
Spain	6.0	9947
Switzerland	5.9	nan
Turkey	3.8	16074
United States of America	5.4	nan

Table 3.1: Yearly average length of stay for acute care patients and discharge number measured per 100000 inhabitants [2].

planning and control are the hospital discharge and the hospital admission. Hospital discharge is defined as the release of a patient who has stayed at least one night in hospital. It includes deaths in hospital following inpatient care. Same-day discharges are usually excluded [2]. Similarly hospital admission is defined as the number of cases of a specified disease or condition admitted to hospitals, related to the population of a given geographical area. In order to provide information on the efficiency of hospital departments and give operational guidelines for medical staff hospital **discharge rate** (**DR**) and hospital **admission rate** (**AR**) are also introduced [29]. Discharge rate is defined as the number of discharge divided by the maximum number of discharge recorded in the department. Similarly admission rate is defined. The distribution of discharge rate supplies the probability that a given number of patients is discharged in one day; likewise the admission rate provides the probability that a given number of patients is admitted in one day.

3.4 Operational modelling of hospital resources

Healthcare costs continue to grow faster than the economy, and the health share of the Gross Domestic Product has maintained its upward trend. Policymakers are among those who are increasingly concerned with the growing burden of medical care expenses to governments, consumers, and insurers. Hospital costs are often the focus of this concern, because they constitute the largest single component of healthcare spending [30]. Therefore once introduced a generic framework for the healthcare planning and control, hospital resource consumption are analysed.
The internal dynamic of a hospital represent a complex non-linear structure. To plan and manage the day-today running of a hospital requires a thorough understanding of the system together with detailed information for decision-making. An appreciation of the dynamics governing a hospital system, and the flow of patients through it, point towards the need for sophisticated capacity models reflecting the complexity, uncertainty, variability and limited resources. Within the proposed framework, necessarily detailed hospital capacity models incorporate time-dependent demand profiles and meaningful statistical distributions that capture the inherent variability in a number of patient factors, such as lengths of stay and operation times.

3.4.1 Hospital dynamics and modelling approaches

A common current practice is to plan and manage hospital capacities through a simple deterministic approach using average patient flows, average needs, average length-of-stay, average duration of surgical operations etc. Patient flows, patient needs, and utilisation of hospital capacities involve complexity, uncertainty, variability, constraints, and scarce resources. Mathematically speaking, a hospital corresponds to a complex stochastic system so that the common deterministic approach for planning and managing the system can be expected to be inadequate [31].
The dynamics governing a hospital, and the flow of patients through it, means that the necessary models should reflect the complexity, uncertainty, variability and limited resources. Examples of these conditions, themselves visibly evident within the participating hospitals, are listed below [32]:

Complexity Rules governing patient admissions into hospital, e.g., keep some beds free for emergency patients only; elective patients may only be deferred so many times before increasing their priority.
Patient-flows through the hospital, e.g., when there is no available bed, we try to admit the patient into another suitable, and available, hospital bed albeit on a different ward; intensive care patients may be discharged early to high dependency care if, and only if, various complex criteria are satisfied.
Constraints imposed by other hospital services, e.g., patients cannot go to theatre if there is no available inpatient bed for them in the first place; operations themselves are subject to theatre space and surgeons hours.

Uncertainty Demand is likely to be a function of time, e.g., elective (planned) patient arrivals can be controlled and are often therefore highly correlated with the month of the year, day of the week, and hour of the day; planned scheduled admissions to hospital though must also account for emergency patients who arrive at random, often in quick succession, and who must be admitted with the minimum of delay.

Variability Patient LoS varies enormously between and within different hospital specialties. For example, Paediatric care length of stay is frequently biased towards shorter LoS, but occasionally a child might stay a very long time, which can cause a disproportionate *blocking* effect. LoS for Geriatric care, however, can be expected to show very different characteristics from that of Paediatric care. Here LoS is much longer and the blocking effect can be extreme when elderly patients stay for months rather than days and become so called *bed-blockers*.

Limited resources Hospitals must treat increasing number of patients through diminishing bed numbers. There is a need to efficiently and effectively plan and manage all hospital resources with particular emphasis on inpatient beds, operating theatres, hospital workforce, and expensive critical care resources. Furthermore, the participating hospitals requested models to aid with both the planning and management of resources. Appropriate detailed models that can evaluate a variety of scenarios could be powerful tools for good planning and management decisions.

Planning tools Capacity planning in hospitals is largely a strategic decision. For example the total number of beds in a new hospital and the number of beds in various specialities are very major concerns; here the planning horizon could be about 10 years. Planning tools should allow hospital staff to examine in detail the likely capacities required over time, for example the annual hospital business planning cycle each financial year. They should enable the user to identify the likely consequences of changes in numbers, and distribution across the hospital, of beds, theatres and workforce.
Planning for services across a region, e.g., the number and location of outpatient clinics to serve the populations needs; the number and distribution of critical care beds in a geographical region.

Management tools Management of available capacities could be from day to day or over longer periods such as winter months and summer months. An example would be a planned transfer of surgical beds to elderly medical patients in winter.
Management tools should allow the user to examine resources in detail over smaller time intervals in order to maximise their utilisation and provide a more efficient use healthcare resources. For example, the consequences of changing daily arrival patterns for elective patients and re-scheduling of nurses.

All of these features point towards a need for sophisticated hospital capacity models. There is considerable scope for Operational Research models to be widely used for this purpose. Indeed, since the early 80s was created and utilized different approaches for a wide range of scenarios in hospital resource modelling. Within bed modelling alone, a number of operational approaches have been utilised, including queueing models [33] and [34], integer programming [35], forecasting [36] and simulation [37], [38], and [39]. A similar array of operational methods have been used in operating theatre modelling [40] and [41] and workforce planning [42], [43], [44], [45], and [46].

3.5 Identification of managerial deficiencies

Once the content of the framework has been established for a given application, further analysis of this content may identify managerial problems. In the remainder of this section, we discuss examples of four kinds of typical problems.

Deficient or lacking planning functions Overlooked or poorly addressed managerial functions can be encountered on all levels of control [16], but are often found on the tactical level of control [18]. In fact, to many, tactical planning is less tangible than operational planning and even strategic planning. Inundated with operational problems, managers are inclined to solve problems at hand (i.e., on the operational level) . We refer to this phenomenon as the *real-time hype* of managers. A claim for *more capacity* is the universal panacea for many healthcare managers. It is, however, often overlooked that instead of such drastic strategic measures, tactically allocating and organizing the available resources may be more effective and cheaper. Consider for example a *master schedule* or *block plan*, which is the tactical allocation of blocks of resource time (e.g. operating theatres, or CTscanners) to specialties and patient categories during a week. Such a block plan should be periodically revised to react on variat ions in supply and demand. However, in practice, it is more often a result of historical development than of analytical considerations [47].

Inappropriate planning approaches There are many logistical paradigms, such as Just-In-Time (JIT), Kanban, Lean, Total Quality Management (TQM), and Six Sigma, all of which have reported success stories. As these paradigms are most ly developed for indust ry, they generally cannot be simply copied to healthcare without impunity. The tendency to uncritically embrace a solution concept , developed for a rather specific manufacturing environment , as the panacea for a variety of other problems in totally different environments has led to many disappointments [48]. The structure provided by the framework helps to identify whether a planning approach is suitable for a planning function in a particular organizational environment. Planning approaches are only suitable if they fit the internal and external characteristics of the involved application. They have to be adapted to the characteristics that are unique for healthcare delivery [49], such as:

1. patient participation in the service process
2. simultaneity of product ion and consumption
3. perishable capacity
4. intangibility of healthcare outputs

5. heterogeneity

Lack of coherence between planning functions The effectiveness and efficiency of healthcare delivery is not only determined by how the various planning functions are addressed; this is also determined by how they interact. As healthcare providers such as hospitals are typically formed as a cluster of autonomous departments, planning is also often functionally dispersed. The framework structures planning functions, and provides insight in their horizontal (cross-management) and vertical (hierarchical) interactions. Horizontal interact ion between managerial areas in the framework provides that required medical information and protocols, and all involved resources and materials, are brought together to enable both effective and efficient healthcare delivery. Downward vertical interact ion concerns concretizing higher level objectives and decisions on a shorter planning horizon.

Planning functions that have conflicting objectives As argued, the framework structures planning functions and their horizontal and vertical interact ions. The framework can thus identify conflicting objectives between planning functions. For example, minimal-invasive surgery generally results in significant reduced length of stay in wards and improved quality of care, but results in higher costs and increased capacity consumption for the operating theatre department. These departments are often managed autonomously and independently, which leads to sub-optimal decision making from both the patient's and the hospital's point of view.

4

Mathematical framework

This Chapter provides an overview of the main mathematical theories used in subsequent chapters. These include stochastic differential equations, real option analysis, option pricing and Poisson processes.

A model is a simplified representation of a complex system designed to focus in on a specific question. In general, modelling techniques used in health are adaptations from other fields such as telecommunications and traffic engineering. In health service planning, modelling techniques derived from queuing theory can be used to forecast the effects of changes on access to services and to calculate the required capacity of services given assumptions about patterns of demand and levels of utilisation; techniques derived from the physics of gravitation may be used to estimate catchment areas of new facilities; models utilising network analysis may be used to study patients travel requirements to services, models based on Markov chains can be used to assess patients progress though treatment and also in economic assessments. Other modelling techniques are used for optimize the resource consumption. They can also help to identify where there may be problems or inefficiency, identify priorities and focus efforts.

In this thesis is proposed the use of different mathematical model in order to optimize and describe the healthcare resource consumption. In particular the usage of financial model is proposed. In this chapter we introduce the mathematical concepts required to describe the operational models proposed.

4.1 Stochastic Differential Equations

Let $b(x,t) = (b_i(x,t))$ and $\sigma(x,t) = (\sigma_{ij}(x,t))$ where $1 \leq i \leq m$ and $1 \leq j \leq d$ be measurable functions defined on $\mathbb{R}^m \times [0,T]$ and $\mathbb{R}^m$ and $M(m,d)$ value receptively.

Definition 4.1.1. *The process* $(\Omega, \mathcal{F}, \{\mathcal{F}\}_{t\in[0,T]}, \{X\}_{t\in[0,T]}, (B_t)_t, \mathbb{P})$ *is solution of the stochastic differential equation*

$$\begin{cases} dX_t = b(t,X_t) + \sigma(t,X_t)dB_t \\ X_u = X_0 \in \mathbb{R}^m \end{cases} \tag{1.1}$$

if

(i) $(\Omega, \mathcal{F}, \{\mathcal{F}\}_t, (B_t)_t, \mathbb{P})$ *is a standard d-dimensional Brownian process*

(ii) *for all* $t \in [u,T]$

$$X_t = X_0 + \int_u^t b(s,X_s)dt + \int_u^t \sigma(s,X_s)dB_s. \tag{1.2}$$

We call σ *diffusion and* b *drift.*

Definition 4.1.2. *The stochastic differential Equation (1.1) has strong solution if for all standard Brownian motions* $(\Omega, \mathcal{F}, \{\mathcal{F}\}_t, (B_t)_t, \mathbb{P})$ *there exists a process* X *such that* $(\Omega, \mathcal{F}, \{\mathcal{F}\}_t, \{X\}_t, (B_t)_t, \mathbb{P})$ *is a solution of (1.1).*

Definition 4.1.3. *There is uniqueness in law for the solution of (1.1) if, given two solutions* $X^i = (\dot{\Omega}, \mathcal{F}^i, \{\mathcal{F}^i\}_t, \{X^i\}_t, (B^i_t)_t, \mathbb{P}^i)$, $i = 1, 2$, X^1 *and* X^2 *have the same law.*

Definition 4.1.4. *There is uniqueness of the trajectories for the (1.1) if, given two solution* $(\Omega^i, \mathcal{F}^i, \{\mathcal{F}^i\}_t, \{X^i\}_t, (B^i_t)_t, \mathbb{P}^i)$, $i = 1, 2$, *(that are defined on the same probability space and on the same Brownian motion), then* $\mathbb{P}(X^1_t = X^2_t \text{ for all } t \in [u,T]) = 1$.

Example 4.1.5. *Consider a* 1*-dimensional equation*

$$\begin{cases} dX_t = b(t,X_t) + \sigma(t,X_t)dB_t \\ X_0 = x > 0 \end{cases} \tag{1.3}$$

where the drift b *and the diffusion* σ *are linear. Dividing by* X_t *and using the Ito' formula to calculate the stochastic differential of* $\log(X_t)$ *we obtain*

$$d(\log(X_t)) = \frac{dX_t}{X_t} - \frac{1}{2X_t^2}d\langle X\rangle_t \tag{1.4}$$

remembering that $d\langle X\rangle_t = \sigma^2 X_t^2 dt$*, it holds*

$$d(\log(X_t)) = (b - \frac{\sigma^2}{2})dt + \sigma dB_t. \tag{1.5}$$

So

$$X_t = xe^{(b-\frac{\sigma^2}{2})t+\sigma B_t}. \tag{1.6}$$

This process is called a ***geometric Brownian motion*** *(**GBM**) and it is used to describe the evolution of paths that are always positive.*

We say that b and σ satisfy **Hypothesis (A)** if they are (t,x)-measurable and if exist $L > 0$ and $M > 0$ such that for all $x, y \in \mathbb{R}^m$, $t \in [0,T]$

$$|b(t,x)| \leq M(1+|x|) \qquad \sigma(t,x)| \leq M(1+|x|) \tag{1.7}$$

$$|b(t,x) - b(t,y)| \leq L|x-y| \qquad |\sigma(t,x) - \sigma(t,y)| \leq L|x-y| \tag{1.8}$$

Theorem 4.1.6. *([50]) If* b *and* σ *are* (t,x)*-measurable and satisfy Hypothesis (A) and* $X \in \Lambda^2([0,T])$ *is solution, for* $t \leq T$*, of* $X_t = SX_t$

$$X_t = (SX)_t = \eta + \int_u^t b(s, X_s)dt + \int_u^t \sigma(s, X_s)dB_s. \tag{1.9}$$

then

$$\mathbb{E}\Big[\sup_{u\leq t\leq T} |X_t|^2\Big] \leq C(M,T)(1 + \mathbb{E}|\eta|^2). \tag{1.10}$$

In particular if $\eta \in L^2$ *then* $X \in M^2([0,T])$*.*

Theorem 4.1.7 (Existence and uniqueness theorem)**.** *Let* $u \leq 0$ *and* η *be random variables* $\mathcal{F}_u$*-measurable,* $\mathbb{R}^m$*-valued into* L^2*. If* b *and* σ *satisfies Hypothesis (A) then exist* $X_t \in M([u,T])$ *such that*

$$X_t = \eta + \int_u^t b(s, X_s)dt + \int_u^t \sigma(s, X_s)dB_s. \tag{1.11}$$

Moreover if exist another solution $\tilde{X}_t$ *for (1.11) then*

$$\mathbb{P}(X_t = \tilde{X}_t \text{ for all } t \in [u,T]) = 1. \tag{1.12}$$

Furthermore it can be proven that exits a process that satisfies (1.11) if b and σ are not *uniformly Lipschitz continuous* in the variable x (Condition 1.8).

Theorem 4.1.8 (Uniqueness in law)**.** *Let* $B_i = (\,\dot{\Omega}, \mathcal{F}^i, \{\mathcal{F}^i_t\}_t, \{B^i_t\}_t, (B^i_t)_t, \mathbb{P}^i)$*,* $i = 1, 2$*, standard Brownian motion d-dimensional,* η_i*,* $i = 1, 2$ *random variables m-dimensional in* $L^2(\Omega_i, \mathcal{F}^i_u), \mathbb{P}^i$ *having the same law. Suppose that* b *and* σ *verify Hypothesis (A) and* X^i *are the solutions of*

$$X^i(t) = \eta_i + \int_u^t b(s, X^i(s))dt + \int_u^t \sigma(s, X^i(s))dB^i_s. \tag{1.13}$$

Then the processes (X^i, B^i)*,* $i = 1, 2$ *have the same law.*

Now we analyse how the solution depends on the initial data η. From now we consider that η is $\mathcal{F}_u$ -measurable and square integrable.

Proposition 4.1.9. *If* X *is solution of (1.1) and apply the Hypothesis (A), for all* $t \in [0, T]$*,*

$$\mathbb{E}\Big[\sup_{u \le s \le t} |X_s - \eta|^2\Big] \le c(T, M)(t - u)(1 + \mathbb{E}|\eta|^2) \tag{1.14}$$

Proposition 4.1.10. *In the Hypothesis (A) if* X_i*,* $i = 1, 2$ *is solution of*

$$\begin{cases} dX_i(t) = b(t, X_i(t)) + \sigma(t, X_i(t))dB_t \\ X_i(u) = \eta_i \end{cases} \tag{1.15}$$

then for $t \in [0, T]$

$$\mathbb{E}\Big[\sup_{u \le s \le t} |X_1(s) - X_2(s)|^2\Big] \le 3\mathbb{E}(|\eta_1 - \eta_2|^2)e^{c(L,T)(t-u)} \tag{1.16}$$

4.1.1 Markovs's Properties

In this section we call $X^{x,s}_t$ the solution of

$$\begin{cases} dX_t = b(t, X_t) + \sigma(t, X_t)dB_t \\ \\ X_s = x \end{cases} \tag{1.17}$$

we suppose that $X^{x,s}_t$ is continuous for $x, s, t,\ t \le s$. If Γ is the borelian set of $\mathbb{R}^m$, $x \in \mathbb{R}^m$ $t \le s$ we say that

$$p(s, t, x, \Gamma) = \mathbb{P}(X^{x,s}_t \in \Gamma). \tag{1.18}$$

Proposition 4.1.11. *In Hypothesis (A),* p *is a transition function and* $(\Omega^i, \mathcal{F}, \{\mathcal{F}_t\}_t,$ $(X^{x,s}_t)_t, \mathbb{P})$ *is a Markov's process whit initial time* s *and initial law* δ_t *associate to transition function* p*.*

4.1.1.1 Feynman-Kac Formula

Feynman-Kac formula [51], [52], named after Richard Feynman and Mark Kac, establishes a link between parabolic partial differential equations (PDEs) and stochastic processes. It offers a method of solving certain PDEs by simulating random paths of a stochastic process. Conversely, an important class of expectations of random processes can be computed by deterministic methods.
Consider two real functions ϕ and f, continuous and define respectively on $\mathbb{R}^m$ and $\mathbb{R}^m \times [0,T]$. We also consider two hypothesis

$$|\phi(x)| \leq M(1+|x|), \quad |f(x,t)| \leq M(1+|x|) \quad (x,t) \in \mathbb{R}^m \times [0,T]. \tag{1.19}$$

Also define the differential operator L_t on $\mathbb{R}^m \times [0,T]$ as

$$L_t = \frac{1}{2}\sum_{i,j=1}^{m} a_{ij}(x,t)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{m} b_i(x,t)\frac{\partial}{\partial x_i} \tag{1.20}$$

when $a = \sigma\sigma^T$.

Theorem 4.1.12. *Suppose that*

i) b and σ satisfy Hypothesis (A), exists $\lambda > 0$ such that $\langle a(x,t)z, z\rangle \geq \lambda |z|^2$ for all $(x,t), \in \mathbb{R}^m \times [0,T]$, $z \in \mathbb{R}^m$.

ii) $\phi : \mathbb{R}^m \to \mathbb{R}$ is continuous and satisfies (1.19).

iii) function f and c, define on $\mathbb{R}^m \times [0,T]$, are locally Lipschitz continuous; c has a lower bound, f satisfies (1.19).

Then there exists a function u, continuous on $\mathbb{R}^m \times [0,T]$ and $C^{2,1}(\mathbb{R}^m \times [0,T[)$ that is solution of

$$\begin{cases} L_t u + \frac{\partial u}{\partial t} - cu = f \quad on \ \mathbb{R}^m \times [0,T[\\ u(x,T) = \phi(x). \end{cases} \tag{1.21}$$

Moreover it holds

$$u(x,t) = \mathbb{E}^{x,t}\Big[\phi(X_T)e^{-\int_t^T c(s,X_s)ds}\Big] - \mathbb{E}^{x,t}\Big[\int_t^T f(X_s,s)e^{-\int_t^s c(v,X_v)ds}ds\Big]. \tag{1.22}$$

u is the only solution that has polynomial increase.

4.1.2 Affine Models

Affine term structure models have gained significant attention in literature, mainly due to their analytical tractability and statistical flexibility. They find large application in quantitative finance and in many other contexts for their versatility also because they have explicit formulas easy to calculate.

The class of Affine Models was introduced by Duffie and Kan the core is the framework of Duffie and Kan [53]. Based on mathematical structure developed in finance, where they have found growing interest due to their computational tractability, we use these models in a completely new content. Let $D \subset \mathbb{R}^d$ be the closure of a (possibly unbounded) starshaped open set and we consider functions $\mu : D \to \mathbb{R}^d$, $\sigma : D \to \mathbb{R}^{d\times d}$, respectively continuous and measurable, such that $x \in D \mapsto \sigma(x)\sigma^\top(x)$ is continuous. Let the uncertainty be generated by a d-dimensional Brownian process B, defined on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$.

Then, for every $x \in D$, consider the stochastic differential equation:

$$dX_t = \mu(X_t)dt + \sigma(X_t)dB_t, \quad X_0 = x. \tag{1.23}$$

A solution X of (1.23) is affine if the $\mathcal{F}_t$-conditional characteristic function of X_T is exponential affine in X_t, for all $0 \leq t \leq T$. That is, there exist $\mathbb{C}$- and $\mathbb{C}^d$-valued functions $\phi(t, i\xi)$ and $\psi(t, i\xi)$, respectively, with jointly continuous t-derivatives such that

$$\mathbb{E}\left[e^{i\langle\xi, X_T\rangle}\middle|\mathcal{F}_t\right] = e^{\phi(T-t,i\xi)+\langle\psi(T-t,i\xi)Y_t\rangle}, \tag{1.24}$$

for all $\xi \in \mathbb{R}^d$, $t \in [0, T]$. We observe that the real part of $\phi(T-t, i\xi) + \langle\psi(T-t, i\xi)X_t\rangle$ has to be negative, as the conditional characteristic function is bounded. The set D is called the state space of X. When X is affine, the matrix $\sigma(x)\sigma^\top(x)$ and $\mu(x)$ are affine in x:

$$\sigma(x)\sigma^\top(x) = A + \sum_{i=1}^{d} x_i B_i, \qquad \mu(x) = \alpha + \sum_{i=1}^{d} x_i \beta_i, \tag{1.25}$$

where A, B_i are $d \times d$ real matrices and α,$\beta\ _i \in \mathbb{R}^d$, for $i = 1, \ldots, d$.

Model parameters cannot be chosen arbitrarily; as discussed in [54]; there are admissibility restrictions required for the existence of the process X_t. The authors prove the existence, for each value of d, of $d+1$ disjoint admissible regions of the parameter space. In each of these families different restrictions are imposed on the parameters. Moreover, affine models do not have a unique representation, that is, there exist different choices of the model parameters that generate identical behaviour of the process. The functions ϕ and $\psi = (\psi_1, \dots, \psi_d)$ solve the system of Riccati equations

$$\begin{cases} \partial_t \phi(t,\eta) = \frac{1}{2}\langle A\psi(t,\eta), \psi(t,\eta)\rangle + \langle \alpha, \psi(t,\eta)\rangle \\ \partial_t \psi_i(t,\eta) = \frac{1}{2}\langle B_i\psi(t,\eta), \psi(t,\eta)\rangle + \langle \beta_i, \psi(t,\eta)\rangle, \quad i = 1, \dots, d, \end{cases} \tag{1.26}$$

with the initial conditions $\phi(0,\eta) = 0$ and $\psi(0,\eta) = \eta$, for all $\eta \in i\mathbb{R}^d$. The existence and uniqueness for the system (1.26) can also be established.

Theorem 4.1.13. (in [55])

i) For every $\eta \in \mathbb{C}^d$, there exists a unique solution $(\phi(\cdot,\eta), \psi(\cdot,\eta))$ of the Riccati Equations (1.26), defined in $t \in [0, T(\eta))$, for some $T(\eta) \in (0, +\infty]$. Moreover $T(0) = +\infty$.

ii) The domain $\mathcal{D} = \{(t,\eta) \in \mathbb{R}_+ \times \mathbb{C}^d : \ t < T(\eta)\}$ is open in $\mathbb{R}_+ \times \mathbb{C}^d$ and, for all $\eta \in \mathbb{C}^d$ either $T(\eta) = +\infty$ or $\lim_{t \uparrow T(\eta)} |\psi(t,\eta)| = +\infty$, respectively.

iii) For all $t \in \mathbb{R}_+$, the t-section $\mathcal{D}(t) = \{\eta \in \mathbb{C}^d \mid (t,\eta) \in \mathcal{D}\}$, is an open neighbourhood of 0 in $\mathbb{C}^d$. Moreover $\mathcal{D}(t_1) \supseteq \mathcal{D}(t_2)$ for $0 \le t_1 \le t_2$.

iv) ϕ and ψ are analytic functions on $\mathcal{D}$.

As a function of t, the characteristic function of an $\mathbb{R}^d$-valued affine process Y solves a generalized Riccati equation as it is shown in [56]. The following result provides an applicable condition to establish the existence of exponential moments of Y.

Theorem 4.1.14. (in [55]) *Let $\eta \in \mathbb{R}^d$ and $s > 0$. Suppose that $\phi(\cdot,\eta) \in C^1([0,s];\mathbb{R})$, $\psi(\cdot,\eta) = (\psi_1(\cdot,\eta), \dots, \psi_d(\cdot,\eta)) \in C^1([0,s];\mathbb{R}^d)$ satisfy the System (1.26) with the initial datum $(0,\eta)$. Then the solution process X of (1.23) satisfies*

$$\mathbb{E}\left[e^{\langle \eta, X_s\rangle} \,\middle|\, \mathcal{F}_t\right] = e^{\phi(s-t,\eta) + \langle \psi(s-t,\eta), X_t\rangle}, \qquad \forall\, 0 \le t \le s. \tag{1.27}$$

Although there are several different ways to define affine models, we shall use the following definition. The dynamics of the state vector X is affine with

$$\sigma(x) = \sqrt{I_d + \mathrm{diag}(\langle x, \gamma^1\rangle, \dots, \langle x, \gamma^d\rangle)}. \tag{1.28}$$

for vectors $\gamma^i \in \mathbb{R}^d$, $i = 1, \ldots, d$. Note that the requirement that σ is diagonal does not result in a loss of generality. An application of an invertible linear transformation to the state process allows to reduce the diffusion coefficient to the form (1.28). In fact, this changes the form of the stochastic differential Equation (1.23). The relation with the matrices A and B_i, $i = 1, \ldots, d$ in (1.25) is the following:

$$A = I_d \text{ (identity matrix)} \qquad B_i = \text{diag}(\gamma_i^1, \ldots, \gamma_i^d), \quad i = 1, \ldots, d. \tag{1.29}$$

4.2 Real option analysis

A key responsibility of hospital executives is to make important decisions concerning budgeting, investment and resource allocation decisions. They have to decide which medical technologies to invest in and how to balance these investments with other projects, such as building construction or information technology. These decisions must be made in the context of ever changing healthcare policies, insurance reimbursements, patient demand, and competition with other hospitals.
The uncertainties at the hospital level and market place should be explicitly addressed and accounted for in executive decision-making processes. Furthermore, past decisions affect the range of decisions that can be taken in the future. For example, renting compared to buying medical equipment can result in different sets of possible future actions. In the same way discharge policies strongly influence the admission number.
The possibility to take an action in the future is called an option. Options give executives flexibility to respond to future events and manage risk. The flexibility of adjusting ones plan of action has value that should be accounted for in the analysis. **Real Options Analysis** (**ROA**) is a mathematical approach that calculates the value of options associated with a decision. ROA determines optimal investment scope and timing taking future decisions and flexibility into account. ROA originated from options theory, which determines the value of financial options that give option holders the right to buy or sell stocks at a previously set price.

4.2.1 Literature review and real option basics

The beginning of options theory starts with the works of [57] and [58] on the pricing of financial options and the development of closed-form solutions for the value of call and

put options. Scholes and Merton received the 1997 Nobel Prize in Economics for their contributions. Alternatives to closed form solutions are partial differential equation models, simulations, or portfolio optimization techniques [59].
The holder of a **call option** has the right, but not the obligation, to buy the underlying asset of the option, often a stock, within a specified period at a given price, called the strike price. Similarly, a put option allows the selling of the underlying asset at the strike price [60]. The parallels between financial options and real options were first discussed by [61]. He called them growth options, which later became known as real options. Early literature on real options focused on determining the value of one specific type of option at a time, such as the option to delay or modify the operating scale of a project ([62], [63], [64]). Later research focused on combining these options to model more complex options such as duopoly settings [65] or compound options [66]. Financial and real options share common terminology and concepts but also have differences [67]. In financial options, the value of the underlying asset (e.g., a company stock) is its market price, while in real options it is the investments value, which is typically regarded as the **Net Present Value** (**NPV**) of future cash flows. The strike price in financial options is the price for which the underlying asset can be bought or sold when the option is exercised. In real options, the strike price is the additional investment required to exercise the option, such as expanding or abandoning an investment or project. The expiration time of a financial option is the point in time when the right to exercise the option ends, which is similar to real options, where the right or possibility to take actions ceases to exist. For both financial and real options, risk-free rates are used for discounting. Dividends paid by a financial option correspond in real options to cash inflows and outflows of the project. Uncertainties regarding the future value of the underlying asset (financial option) or project (real option) are measured through volatility in both cases. A major difference between real and financial options is that management decisions can affect the real options value, while a financial options value cannot be influenced in that way.
Real options can be regarded as an extension and improvement over NPV analysis. The NPV is the discounted sum of future cash flows. ROA is a more comprehensive approach as ROA takes into account that future decisions are not static and predetermined, but are made in response to future events that are only known as uncertain variables in the present [68]. NPV does not account for managerial flexibility in responding to

new information. As uncertainties evolve over time, the optimal action depends on the actual values of the uncertain variables. Consequently, the option of delaying, expanding or abandoning a project at some point after the initial decision has a value that NPV does not capture because it is based on averages and static decisions.
ROA has been applied to a variety of applications in numerous industries [69]. However, only a few papers discuss the application of ROA to healthcare problems. Furthermore, all of these journal papers fall into one of two categories:

1. Mathematical papers that contribute to theory and merely use healthcare as motivation or background

2. Conceptual papers that discuss applications of ROA to healthcare, but do not provide mathematical analyses

Papers in both of these categories fail to address the needs of practitioners and academics who seek a mathematically accessible and real-world focused guide for applying ROA.
An example of a theoretical contribution is the work by [70], in which the authors analysed the effect of various contract variations on the investment decisions of a profit-maximizing hospital. The healthcare system was modeled as a production process, where the inputs of medical care and technology led to patient recovery, subject to uncertainties based on personal characteristics. [71] adapted the classical real options model to include characteristics of the healthcare sector and study the effect of decreasing drug prices. A ROA model to compare health plans by health maintenance organizations (HMOs) was developed and tested by [72]. These authors showed how the option to receive coverage for out-of-network care increased in value with the severity of health problems. [73] and [74] explored the combination of real options with the analytical hierarchy process (AHP) to incorporate intangible values and game theory to capture competition.
Conceptual papers include the work by [60], who discussed the application of real options to support an investment decision for a hospitals new imaging department. Based on input from stakeholders and analysis of internal and external influence factors, a portfolio of options (e.g., expand, delay, abandon, or contract) with varying market and technological uncertainty was determined for the different investment options. [75] used the real options framework to illustrate that the value of investments by hospitals

can be increased if options are included in contracts that allow for more decision flexibility in the future.

The real options framework has also been applied to medical decision-making. [76] illustrated how *watchful waiting* benefits patients since delaying a treatment, such as a surgery, allows for better clinical decisions once more information becomes available and uncertainties are reduced. Watchful waiting was also discussed in a managed care setting by [77].

We identified only one paper that bridged the gap between theory and practice and addressed the need of hospital executives and analysts who want to apply ROA. [78] showed how ROA can effectively support the decision to invest in either a stationary or mobile lithotripter. While the NPV analysis favored a stationary device, ROA led to a different result. A mobile lithotripter includes the option to rent out the device to other medical institutions, so the hospital can respond more flexibly to changing patient volume. The value of this option makes the mobile device the favored decision alternative.

The value of real options stems from the fact that when investing in risky assets, we can learn from observing what happens in the real world and adapting our behaviour to increase our potential upside from the investment and to decrease the possible downside. In the real options framework, we use updated knowledge or information to expand opportunities while reducing danger. In the context of a risky investment (patient hospitalization can be interpreter as a risky investment), there are three potential actions that can be taken based upon this updated knowledge. The first is that you build on good fortune to increase your possible profits; this is the option to expand. For instance, a market test that suggests that consumers are far more receptive to a new product than you expected them to be could be used as a basis for expanding the scale of the project and speeding its delivery to the market. The second is to scale down or even abandon an investment when the information you receive contains bad news; this is the option to abandon and can allow you to cut your losses. The third is to hold off on making further investments, if the information you receive suggests ambivalence about future prospects; this is the option to delay or wait. You are, in a sense, buying time for the investment, hoping that product and market developments will make it attractive in the future.

The value of learning is greatest, when you and only you have access to that learning and can act on it. After all, the expected value of knowledge that is public, where anyone can act on that knowledge, will be close to zero. We will term this third condition *exclusivity* and use it to scrutinize when real options have the most value.

4.2.2 The Option to Delay an Investment

Investments are typically analysed based upon their expected cash flows and discount rates at the time of the analysis; the net present value computed on that basis is a measure of its value and acceptability at that time. The rule that emerges is a simple one: negative net present value investments destroy value and should not be accepted. Expected cash flows and discount rates change over time, however, and so does the net present value. Thus, a project that has a negative net present value now may have a positive net present value in the future. In a competitive environment, in which individual firms have no special advantages over their competitors in taking projects, this may not seem significant. In an environment in which a project can be taken by only one firm (because of legal restrictions or other barriers to entry to competitors), however, the changes in the projects value over time give it the characteristics of a call option.

We assume that a project requires an initial up-front investment of X, and that the present value of expected cash inflows computed right now is V. The net present value of this project is the difference between the two:

$$NPV = V - X \tag{2.30}$$

Now assume that the firm has exclusive rights to this project for the next n years, and that the present value of the cash inflows may change over that time, because of changes in either the cash flows or the discount rate. Thus, the project may have a negative net present value right now, but it may still be a good project if the firm waits. Defining V again as the present value of the cash flows, the firms decision rule on this project can be summarized as follows:

- $V > X$ Take the project: Project has a positive present value
- $V < X$ Do not take the project: Project has negative net present value

If the firm does not invest in the project, it incurs no additional cash flows, though it will lose what it originally invested in the project. Note that this payoff diagram is that of a call option, the underlying asset is the investment, the strike price of the option is the initial outlay needed to initiate the investment; and the life of the option is the period for which the firm has rights to the investment. The present value of the cash flows on this project and the expected variance in this present value represent the value and variance of the underlying asset.

4.2.3 The Option to Expand

In some cases, a firm will take an investment because doing so allows it either to make other investments or to enter other markets in the future. In such cases, it can be argued that the initial investment provides the firm with an option to expand, and the firm should therefore be willing to pay a price for such an option. Consequently, a firm may be willing to lose money on the first investment because it perceives the option to expand as having a large enough value to compensate for the initial loss. To examine this option, assume that the present value of the expected cash flows from entering the new market or taking the new project is V, and the total investment needed to enter this market or take this project is X. Further, assume that the firm has a fixed time horizon, at the end of which it has to make the final decision on whether or not to take advantage of this opportunity. Finally, assume that the firm cannot move forward on this opportunity if it does not take the initial investment. At the expiration of the fixed time horizon, the firm will enter the new market or take the new investment if the present value of the expected cash flows at that point in time exceeds the cost of entering the market.

4.2.4 The Option to Abandon an Investment

The final option to consider here is the option to abandon a project when its cash flows do not measure up to expectations. The option pricing approach provides a general way of estimating and building in the value of abandonment into investment analysis. To illustrate, assume that V is the remaining value on a project if it continues to the end of its life, and L is the liquidation or abandonment value for the same project at the same point in time. If the project has a life of n years, the value of continuing the project can be compared to the liquidation (abandonment) value. If the value from continuing is higher, the project should be continued; if the value of abandonment is higher, the holder of the abandonment option could consider abandoning the project. Payoff from owning an abandonment option is

- 0 if $V > L$
- $L - V$ if $V \leq L$

Unlike the prior two cases, the option to abandon takes on the characteristics of a put option.

4.2.5 Valuing an Real Option

We need the value of the underlying asset, the variance in that value, the time to expiration on the option, the strike price, the riskless rate and the equivalent of the dividend yield (cost of delay). Actually estimating these inputs for a real option to delay can be difficult, however.

Value Of The Underlying Asset In this case, the underlying asset is the investment itself. The current value of this asset is the present value of expected cash flows from initiating the project now, not including the up-front investment, which can be obtained by doing a standard capital budgeting analysis. There is likely to be a substantial amount of error in the cash flow estimates and the present value, however. Rather than being viewed as a problem, this uncertainty should be viewed as the reason why the project delay option has value. If the expected cash flows on the project were known with certainty and were not expected to change, there would be no need to adopt an option pricing framework, since there would be no value to the option.

Variance in the value of the assets The present value of the expected cashflows that measures the value of the asset will change over time, partly because the potential market size for the product may be unknown, and partly because technological shifts can change the cost structure and profitability of the product. The variance in the present value of cash flows from the project can be estimated in one of three ways.

1. If similar projects have been introduced in the past, the variance in the cash flows from those projects can be used as an estimate. This may be the way that

a consumer product company like Gillette might estimate the variance associated with introducing a new blade for its razors.

2. Probabilities can be assigned to various market scenarios, cash flows estimated under each scenario and the variance estimated across present values. Alternatively, the probability distributions can be estimated for each of the inputs into the project analysis - the size of the market, the market share and the profit margin, for instance - and simulations used to estimate the variance in the present values that emerge.

3. The variance in the market value of publicly traded firms involved in the same business (as the project being considered) can be used as an estimate of the variance. Thus, the average variance in firm value of firms involved in the software business can be used as the variance in present value of a software project.

The value of the option is largely derived from the variance in cash flows - the higher the variance, the higher the value of the project delay option. Thus, the value of an option to delay a project in a stable business will be less than the value of a similar option in an environment where technology, competition and markets are all changing rapidly.

Exercise Price On Option A project delay option is exercised when the firm owning the rights to the project decides to invest in it. The cost of making this investment is the exercise price of the option. The underlying assumption is that this cost remains constant (in present value dollars) and that any uncertainty associated with the product is reflected in the present value of cash flows on the product.

Expiration Of The Option And The Riskless Rate The project delay option expires when the rights to the project lapse; investments made after the project rights expire are assumed to deliver a net present value of zero as competition drives returns down to the required rate. The riskless rate to use in pricing the option should be the rate that corresponds to the expiration of the option. While this input can be estimated easily when firms have the explicit right to a project (through a license or a patent, for instance), it becomes far more difficult to obtain when firms only have a competitive advantage to take a project.

Cost of Delay (Dividend Yield) There is a cost to delaying taking a project, once the net present value turns positive. Since the project rights expire after a fixed period, and excess profits (which are the source of positive present value) are assumed to disappear after that time as new competitors emerge, each year of delay translates into one less year of value-creating cash flows. If the cash flows are evenly distributed over time, and the exclusive rights last n years, the cost of delay can be written as:

$$\text{Annual cost of delay} = \frac{1}{n} \tag{2.31}$$

Thus, if the project rights are for 20 years, the annual cost of delay works out to 5% a year. Note, though, that this cost of delay rises each year , to $\frac{1}{19}$ in year 2, $\frac{1}{18}$ in year 3 and so on, making the cost of delaying exercise larger over time.

4.3 Option pricing

Diffusion processes are used (also) as a model of random phenomena. Among these, a particularly telling example in daily life is given by financial market. In this and the following sections we will see the models to evaluate *derivatives*. We denote S_t the value at time t of a financial instrument and we try to model the evolution of this amount with a diffusion process. In order to choice the *drift* and the *diffusion coefficient* there are two considerations to make. First, it is the amount that must always be positive; therefore $S_t > 0$ for all $t > 0$. Moreover the increases are always considered in a multiplicative; an increase of $p\%$ between time s and t means that $\frac{S_t}{S_s} = 1 + \frac{p}{100}$. Therefore we model the logarithm of the price

$$\frac{dS_t}{S_t} = b(S_t,t)dt + \sigma(S_t,t)dB_t \tag{3.32}$$

where b are vector fields and σ are matrix fields limited and locally Lipschitz. The process $\zeta_t = \log(S_t)$ is solution of

$$\begin{aligned} d\zeta_t &= \big(b(S_t,t) - \tfrac{1}{2}\sigma(S_t,t)\big)dt + \sigma(S_t,t)dB_t = \\ &= \big(b(e^{\zeta_t},t) - \tfrac{1}{2}\sigma(e^{\zeta_t},t)^2\big)dt + \sigma(e^{\zeta_t},t)dB_t \end{aligned} \tag{3.33}$$

The assumptions on of b and σ ensure that the process ζ is defined for all $t > 0$, therefore if $S_0 > 0$ then $S_t > 0$ for all $t \geq 0$. If b and σ are constants and $S_s = x$, the solution of 3.33 is a geometric Brownian motion

$$S_t^{x,s} = xe^{\left(b-\frac{\sigma^2}{2}\right)(t-s)+\sigma(B_t-B_s)}. \tag{3.34}$$

One *option* is what is called derivative: an option to purchase (**call**) is an agreement that gives an investor the right, but not the obligation, to buy a given instrument (*underlying*) at a specified price K (*strike price*) within a specific time T (*maturity*). Therefore, if at time T the underlying price is higher than the strike price K, the option holder will exercise its right and buy the instrument at price K. Otherwise he can buy the instrument on the market at a lower price. This type of contract guarantees the buyer to be able to buy the instrument at price not greater than K. There are options to sell (**put**) that guarantee the right to sell at time T at a price not less than K. Above are described examples of so-called **European options**. There are other types of derivatives, with other rules, such as **American options**. An American option is an option that can be exercised anytime during its life. American options allow option holders to exercise the option at any time prior to and including its maturity date.
Option seller is exposed to a risk: if at time T the underlying price is greater than the exercise price K, the seller would be forced to buy on the market the instrument and sell it at price K. Therefore if S_T is the price at time T, the option seller has a loss of $\big(S_T - K\big)^+$ (where $^+$ indicates the positive part). The loss is $S_T - K$ if $S_T > K$ and is equal to 0 if $S_T < K$.

4.3.1 Hedging strategy

A **portfolio** is a set of assets. Consider a market composed by securities $S_0, S_1, \ldots, S_m$; suppose that S_0 follows the equation

$$dS_0(t) = r_t S_0(t)dt \tag{3.35}$$

where $(r_t)_t$ is a random process. If $S_0(s) = x$, then

$$S_0(t) = xe^{\int_s^t r_u du}. \tag{3.36}$$

S_0 typically corresponds to an investment of the kind of treasury bills or a deposit in a bank account. S_0 is the *risk free* title, even if its evolution is still uncertain. The $(r_t)_t$ process represents the instantaneous interest rate. Other securities follow the equation

$$\frac{dS_i(t)}{S_i(t)} = b_i(S(t), t)dt + \sum_{j=1}^{m} \sigma_{ij}(S(t), t)dB_j(t) \tag{3.37}$$

where $S(t) = \big(S_1(t), \ldots, S_m(t)\big)$.
We say that **Hypothesis (B)** are satisfied if

i) $B = \big(\Omega, \mathcal{F}, (\mathcal{F}_t)_t, (B_t)_t, \mathbb{P}\big)$ is a m-dimensional Brownian motion with natural and complete filtration $(\mathcal{S}_t)_t$.

ii) b and σ are limited and locally Lipschitz. Moreover if $a = \sigma\sigma^*$, the matrix field a is uniformly elliptic; therefore exists $\delta > 0$ such that $\langle a(x,t)z, z\rangle > \delta|z|^2$ for all $(x,t) \in \mathbb{R}^m \times \mathbb{R}^+$, $z \in \mathbb{R}^m$.

iii) The process $(r_t)_t)$ is limited and progressively measurable with respect to the natural and complete filtration $(\mathcal{S}_t)_t$ of B.

The *portfolio process* is

$$V_t = \sum_{i=0}^{m} H_i(t)S_i(t). \tag{3.38}$$

V_t is the value at time t of the amount invested in assets $S_0, \dots, S_m$. $H_0(t), \dots, H_m(t)$ are the amounts invested in each portfolio assets. Assuming that $H_0(t), \dots, H_m(t)$ are real, it is possible to buy non integer quantities of assets. Moreover these amounts change over time and is possible to reallocate resources on portfolio assets.

Definition 4.3.1. *A portfolio is called **self-financing** if* $H_0 \in \Lambda^1_B([0, +\infty[)$, $H_iS_i \in \Lambda^2_B([0, +\infty[)$ *and*

$$dV_t = \sum_{i=0}^{m} H_i(t)dS_i(t). \tag{3.39}$$

Equivalently calling $U_i(t) = H_i(t)S_i(t)$ equation 3.39 becomes

$$dV_T = r_iU_0(t)dt + \sum_{i=1}^{m} U_i(t)b_i(S(t),t)dt + \sum_{j=1}^{m} U_i(t)\sigma_{ij}(S(t),t)dB_j(t) \tag{3.40}$$

and replacing $U_0(t) = V_t - \sum_{i=1}^{m} U_i(t)$,

$$dV_t = r_tV_tdt + \sum_{i=1}^{m} U_i(t)\big(b_i(S(t),t)) - r_t\big)dt + \sum_{j=1}^{m} U_i(t)\sigma_{ij}(S(t),t)dB_j(t). \tag{3.41}$$

In order to evaluate the price of an european option we consider a market composed by S_0 and S_1. We also assume that the dynamics of S_0 and S_1 are described by equations 3.35 and 3.37 where $m = 1$. Let T be the maturity time and $s \leq T$ the initial time. An european option with time to maturity T is a random variable$\Psi \geq 0$ $\mathcal{S}_T$-measurable, where $\Psi = (S_T - K)^+$ if we consider a call option with exercise price K. The **hedging portfolio** is a feasible self-financing portfolio such that $V_T = \Psi$. Portfolio V_T is also called **replicating portfolio**. If V is an hedging portfolio and $V_s = v$, then is possible to invest an amount v and receiving at time T the amount Ψ. Therefore the price of the option Ψ issued at time s is the smallest value v such that exist a portfolio V that replicates Ψ and $V_s = v$. The annualized portfolio $(\tilde{V}_t)_t = R_{s,t}(V_t)_t$ is a supermatigale with respect to the risk neutral measure $\mathbb{Q}$ and

$$\mathbb{E}^{\mathbb{Q}}\Big[R_{s,T}\Psi|\mathcal{S}_s\Big] = \mathbb{E}^{\mathbb{Q}}\Big[\tilde{V}_T|\mathcal{S}_s\Big] \le \tilde{V}_s = V_s \tag{3.42}$$

therefore

$$v_{min} \doteq \mathbb{E}^{\mathbb{Q}}\Big[\tilde{\Psi}|\mathcal{S}_s\Big] \tag{3.43}$$

where $\tilde{\Psi} = R_{s,T}\Psi$.

Theorem 4.3.2. *If Hypothesis (B) are satisfied, let Ψ be an european option such that $\mathbb{E}[\Psi^q] < +\infty$ for some $q > 1$. Then for each hedging portfolio V of Ψ we have $V_s \ge v_{min}$ where v_{min} is defined in equation 3.43. Moreover exist an hedging portfolio V such that $V_s = v_{min}$.*

Therefore if $S_1(s) = x$, the call option price at time s

$$\mathbb{E}^{\mathbb{Q}}\Big[R_{s,T}\big(S_1^{x,s}(T) - K\big)^+|\mathcal{S}_s\Big]. \tag{3.44}$$

Moreover if we assume that $(r_t)_t$ is a deterministic function and given that the random variable $S_1^{x,s}(T)$ and $\mathcal{S}_s$ are independent the price is

$$R_{s,T}\mathbb{E}^{\mathbb{Q}}\Big[\big(S_1^{x,s}(T) - K\big)^+\Big]. \tag{3.45}$$

4.3.2 Arbitrage approach

In this section is proposed a pricing approach based on the arbitrage definition. An **arbitrage** is a financial operation that:

- do not require any capital investment
- the profit probability is > 0 and the loss probability is equal to 0.

In other words, an arbitrage opportunity is a zero-cost strategy which has nonnegative pay-off in all states of nature and strictly positive pay-off in at least one state.

Definition 4.3.3. *A financial market composed by assets $S_0, S_1, \dots$ admits no arbitrage opportunity if all admittable and self-financing portfolios V such that $V_s = 0$, $V_T \geq 0$, $s \leq T$ is respected $V_T = 0$.*

Proposition 4.3.4. *A market composed by assets $S_0, S_1, \dots$ that satisfy equations 3.35 and 3.37 admits no arbitrage.*

We consider a market composed by three assets S_0, S_1, and $C(S_1(t), t)$ and we assume that C is regular enough. Consider the portfolio

$$V_t = H_0(t)S_0(t) + H_1(t)S_1(t) + C(S_1(t), t), \tag{3.46}$$

composed by an option and by variable amount of risk-free assets $S_0(t)$ and assets $S_1(t)$. According to Ito's formula

$$dC(S_1(t), t) = \Big(\bar{L} + \frac{\partial}{\partial t}\Big)C(S_1(t), t)dt + \frac{\partial C}{\partial x}(S_1(t), t)\sigma(S_1(t), t)S_1(t)dB_t \tag{3.47}$$

where

$$\bar{L}_t = \frac{1}{2}\sigma^2(x, t)x^2\frac{\partial^2}{\partial x^2} + b_1(x, t)x\frac{\partial}{\partial x}. \tag{3.48}$$

Assuming that portfolio 3.46 is self-financing so

$$dV_t = H_0(t)dS_0(t) + H_1(t)dS_1(t) + dC(S_1(t), t) \tag{3.49}$$

and considering 3.41

$$\begin{aligned} dV_t = rV_tdt \quad &+ \Big[\Big(\bar{L}_t + \tfrac{\partial}{\partial t} - r\Big)C(S_1(t), t) + H_1(t)S_1(t)\big(b_1(S_1(t), t) - r\big)\Big]dt + \\ &+ \Big[\tfrac{\partial C}{\partial x}(S_1(t), t) + H_1(t)\Big]\sigma(S_1(t), t)S_1(t)dB_t. \end{aligned} \tag{3.50}$$

Choosing

$$H_1(t) = \frac{\partial C}{\partial x}(S_1(t), t) \tag{3.51}$$

the coefficient of dB_t is equal to 0 and portfolio V has the differential

$$\begin{aligned} dV_t \quad &= \Big[rV_t\Big(\bar{L}_t + \tfrac{\partial}{\partial t} - r\Big)C(S_1(t), t) - \tfrac{\partial C}{\partial x}(S_1(t), t)S_1(t)(b_1(S_1(t), t) - r)\Big]dt = \\ &= rV_tdt + \Big(L_t + \tfrac{\partial}{\partial t} - r\Big)C(S_1(t), t)dt \end{aligned} \tag{3.52}$$

where

$$L_t = \frac{1}{2}\sigma^2(x, t)x^2\frac{\partial^2}{\partial x^2} + rx\frac{\partial}{\partial x}. \tag{3.53}$$

In order to avoid arbitrage opportunity for all t

$$\Big(\bar{L}_t + \frac{\partial}{\partial t} - r\Big) C(S_1(t), t) = 0, \tag{3.54}$$

in order that $S_1(t)$ assumes any value > 0 with positive probability, it needs to be

$$\Big(\bar{L}_t + \frac{\partial}{\partial t} - r\Big) C(x, t) = 0, \tag{3.55}$$

for all $x \in \mathbb{R}$, $t \leq T$. Moreover must hold the condition $C(x,T) = \phi(x)$ where $\phi(x) = (x-K)^+$. If ϕ has a polynomial growth, according to Feynman-Kac formula

$$C(x,t) = \mathbb{E}^{\mathbb{Q}}\Big[e^{-r(T-s)}\phi(S_1^{x,t}(T))\Big] \tag{3.56}$$

is solution of the problem 3.55 with final condition $C(x,T) = \phi(x)$. Moreover 3.56 is the only solution with a polynomial growth. With respect to $\mathbb{Q}$, S_1 is solution of 3.57.

Black-Scholes formula In order to evaluate the option price we need to evaluate the expected value $\mathbb{E}^{\mathbb{Q}}\big[\tilde{\Psi}\big]$ where $\tilde{\Psi} = \tilde{V}_T$. Generally the law of the random variable $\tilde{\Psi}$ with respect $\mathbb{Q}$ is not known, therefore simulation method and partial derivatives method are needed. Explicit formula for call and put options are obtained in the **Black-Scholes** model if the following hypothesis are satisfied

i) Volatility σ is constant, $\sigma(x,t) \equiv \sigma$

ii) Risk free is constant, $r_t \equiv r$

iii) Drift b satisfies Hypothesis (B).

If we consider an european buying option, $\Psi = \big(S_T - K\big)^+$. In Black-Scholes model, with respect to $\mathbb{Q}$, the asset $\big(S_1(t)\big)_t$ follows equation

$$\frac{dS_1(t)}{S_1(t)} = r dt \sigma d\tilde{B}_t \tag{3.57}$$

hence with the initial condition $S_1(s) = x$

$$S_1^{x,s}(t) = x e^{\left(r - \frac{\sigma^2}{2}\right)(t-s) + \sigma\left(\tilde{B}_t - \tilde{B}_s\right)}. \tag{3.58}$$

Let $C(x,s)$ be the option price sells at time s with the initial condition $S_1(s) = x$. Then, calling Z the random variable $\mathcal{N}(0,1)$,

$$\begin{aligned} C(x,s) &= e^{r(T-s)}\mathbb{E}^{\mathbb{Q}}\Big[\big(S_1^{x,s}(T) - K\big)^+\Big] = \\ &= e^{r(T-s)}\mathbb{E}\Big[\big(x e^{(r-\frac{\sigma^2}{2})(T-s)+\sigma\sqrt{T-s}Z} - K\big)^+\Big]. \end{aligned} \tag{3.59}$$

The expected value is equal to

$$e^{r(T-s)}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\left(xe^{(r-\frac{\sigma^2}{2})(T-s)+\sigma\sqrt{T-s}z}-K\right)^{+}e^{-\frac{z^2}{2}}dz. \tag{3.60}$$

It is easy to find that the integral is equal to 0 when $z \le \zeta_s$ where

$$\zeta_s = \frac{\left(\log(\frac{K}{x})-(r-\frac{\sigma^2}{2})(T-s)\right)}{\sigma\sqrt{T-s}}. \tag{3.61}$$

Calling with Φ the cumulative distribution of the random variable $\mathcal{N}(0,1)$,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{z^2}{2}}dz = \frac{1}{\sqrt{2\pi}}\int_{-x}^{+\infty}e^{-\frac{z^2}{2}}dz. \tag{3.62}$$

Then

$$\begin{aligned} C(x,s) &= \tfrac{e^{-r(T-s)}}{\sqrt{2\pi}}\int_{\zeta_s}^{+\infty}\big(xe^{(r-\frac{\sigma^2}{2})(T-s)+\sigma\sqrt{T-s}z}-K\big)e^{-\frac{z^2}{2}}dz = \\ &= \tfrac{x}{\sqrt{2\pi}}\int_{\zeta_s}^{+\infty}e^{\frac{1}{2}(z-\sigma\sqrt{T-s})^2}dz - Ke^{-r(T-s)}\Phi(-\zeta_s) \\ &= \tfrac{x}{\sqrt{2\pi}}\int_{\zeta_s-\sigma\sqrt{T-s}}^{+\infty}e^{-\frac{z^2}{2}}dz - Ke^{-r(T-s)}\Phi(-\zeta_s) \\ &= x\Phi(-\zeta_s+\sigma\sqrt{T-s}) - Ke^{-r(T-s)}\Psi(-\zeta_s). \end{aligned} \tag{3.63}$$

Therefore the Black-Scholes formula is

$$C(x,s) = x\Phi\big(-\zeta_s+\sigma\sqrt{T-s}\big) - Ke^{-r(T-s)}\Phi(-\zeta_s). \tag{3.64}$$

4.4 Poisson process

A Poisson process is a simple and widely used stochastic process for modeling the times at which arrivals enter a system. It is in many ways the continuous-time version of the Bernoulli process. For the Bernoulli process, the arrivals can occur only at positive integer multiples of some given increment size (often taken to be 1). For the Poisson process, arrivals may occur at arbitrary positive times, and the probability of an arrival at any particular instant is 0. This means that there is no very clean way of describing a Poisson process in terms of the probability of an arrival at any given instant. It is more convenient to define a Poisson process in terms of the sequence of interarrival times, $X_1, X_2, \ldots$ which are defined to be independent and identically distributed.

Arrival process An arrival process is a sequence of increasing random variable, $0 < S_1 < S_2 < \ldots$, where $S_i < S_{i+1}$ means that S_{i+1} is a positive random variable, i.e., a random variable X such that $F_X(0) = 0$. The random variables $S_1, S2, \ldots$ are called arrival epochs and represent the times at which some repeating phenomenon occurs. Note that the process starts at time 0 and that multiple arrivals can not occur simultaneously (the phenomenon of bulk arrivals can be handled by the simple extension of associating a positive integer random variable to each arrival). We will sometimes permit simultaneous arrivals or arrivals at time 0 as events of zero probability, but these can be ignored. In order to fully specify the process by the sequence $S_1, S_2, \ldots$ of random variables, it is necessary to specify the joint distribution of the subsequences $S_1, \ldots, S_n$ for all $n > 1$.
Although we refer to these processes as arrival processes, they could equally well model departures from a system, or any other sequence of incidents. Although it is quite common, especially in the simulation field, to refer to incidents or arrivals as events, we shall avoid that here. The $n - th$ arrival epoch S_n is a random variable and $\{S_n \leq t\}$, for example, is an event. This would make it confusing to refer to the nth arrival itself as an event.
Any arrival process can also be specified by two alternative stochastic processes. The first alternative is the sequence of interarrival times, $X_1, X_2, \ldots$ These are positive random variables defined in terms of the arrival epochs by $X_1 = S_1$ and $X_i = S_i - S_{i-1}$ for $i > 1$. Similarly, given the X_i, the arrival epochs S_i are specified as

$$S_n = \sum_{i=1}^{n} X_i. \tag{4.65}$$

Thus the joint distribution of $X_1, \ldots, X_n$ for all $n > 1$is sufficient (in principle) to specify the arrival process. Since the interarrival times are independent and identically distributed in most cases of interest, it is usually much easier to specify the joint distribution of the X_i than of the S_i.
The second alternative for specifying an arrival process is the counting process $N(t)$, where for each $t > 0$, the random variable $N(t)$ is the number of arrivals up to and including time t. The counting process $\{N(t); t > 0\}$, is an uncountably infinite family of random variable $\{N(t); t > 0\}$ where $N(t)$, for each $t > 0$, is the number of arrivals in the interval $(0, t]$. Whether the end points are included in these intervals is sometimes

important, and we use parentheses to represent intervals without end points and square brackets to represent inclusion of the end point. Thus (a, b) denotes the interval $\{t : a < t < b\}$, and $(a, b]$ denotes $\{t : a < t \leq b\}$. The counting random variables $N(t)$ for each $t > 0$ are then defined as the number of arrivals in the interval $(0, t]$. $N(0)$ is defined to be 0 with probability 1, which means, as before, that we are considering only arrivals at strictly positive times.

The counting process $\{N(t); t > 0\}$ for any arrival process has the properties that $N(\tau) \geq N(t)$ for all $\tau \geq t > 0$ (i.e., $N(\tau) - N(t)$ is a nonnegative random variable). For any given integer $n \geq 1$ and time $t > 0$, the $n-$th arrival epoch, S_n, and the counting random variable, $N(t)$, are related by

$$\{S_n \leq t\} = \{N(t) \geq n\} \tag{4.66}$$

To see this, note that $\{S_n \leq t\}$ is the event that the n-th arrival occurs by time t. This event implies that $N(t)$, the number of arrivals by time t, must be at least n; i.e., it implies the event $\{N(t) \geq n\}$. Similarly, $\{N(t) \geq n\}$ implies $\{S_n \leq t\}$, yielding the equality in 4.66. An alternate form, which is occasionally more transparent, comes from taking the complement of both sides of 4.66, getting

$$\{S_n > t\} = \{N(t) < n\} \tag{4.67}$$

For example, the event $\{S_1 > t\}$ means that the first arrival occurs after t, which means $\{N(t) < 1\}$ (i.e., $\{N(t) = 0\}$). These relations will be used constantly in going back and forth between arrival epochs and counting random variables. In principle, 4.66 or 4.67 can be used to specify joint distribution functions of arrival epochs in terms of joint distribution functions of counting variables and vice versa, so either characterization can be used to specify an arrival process. In summary, then, an arrival process can be specified by the joint distributions of the arrival epochs, the interarrival intervals, or the counting random variables. In principle, specifying any one of these specifies the others also.

Definition 4.4.1. *A* ***renewal process*** *is an arrival process for which the sequence of inter-arrival times is a sequence of independent identically distributed random variables.*

Definition 4.4.2. *A* ***Poisson process*** *is a renewal process in which the interarrival intervals have an exponential distribution function; i.e., for some real $\lambda > 0$, each X_i has the density*

$$f_X = \lambda e^{-\lambda x}, \quad x \geq 0 \tag{4.68}$$

The parameter λ is called the rate of the process. For any interval of size t, λt is the expected number of arrivals in that interval. Thus λ is called the *arrival rate* of the process.

4.4.1 Doubly stochastic Poisson process

A doubly stochastic Poisson process can be viewed as a two step randomization procedure. A process λ is used to generate another process N by acting as its intensity. This means that N is a Poisson process conditional on λ. The Cox process setup provides us with a very useful framework for modelling hospital bed occupancy time. Many alternative definitions of a doubly stochastic Poisson process can be given. We will offer the one adopted by Bremaud [79].

Definition 4.4.3. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with information structure $\mathcal{F}$. The information structure $\mathcal{F}$ is the filtration, i.e. $\mathcal{F} = \{\mathcal{F}_t, t \in [0, T]\}$. Let λ be a non-negative process, such that λ_t is $\mathcal{F}_t$-measurable, for $t \geq 0$ and assume that*

$$\int_0^t \lambda_s ds < \infty, \quad \mathbb{P} \text{ almost surely (no explosions)} \tag{4.69}$$

A process $N = \{N_t\}_{t \geq 0}$ is called a $\mathcal{F}$-doubly stochastic Poisson process with intensity λ if N is $\mathcal{F}$-adapted and for all $0 \leq t_1 \leq t_2$, $u \in \mathbb{R}$, it holds

$$\mathbb{E}\left\{ e^{iu(N_{t_2} - N_{t_1})} \middle| \mathcal{F}_{t_2} \right\} = \exp\left\{ (e^{iu} - 1) \int_{t_1}^{t_2} \lambda_s ds \right\}. \tag{4.70}$$

$$\mathbb{P}[N_{t_2} - N_{t_1} = k | \lambda_s;\, t_1 \leq s \leq t_2] = \frac{e^{-\int_{t_1}^{t_2} \lambda_s ds} \left(\int_{t_1}^{t_2} \lambda_s \, ds \right)^k}{k!}, \tag{4.71}$$

Hence, the law of iterated expectations gives us the following relation:

$$\begin{aligned} \mathbb{P}[N_{t_2} - N_{t_1} = k] &= \mathbb{E}[\mathbf{1}_{\{N_{t_2} - N_{t_1} = k\}}] \\ &= \mathbb{E}\Big[\mathbb{E}[\mathbf{1}_{\{N_{t_2} - N_{t_1} = k\}} | \mathcal{F}_{t_2}] \Big] \\ &= \mathbb{E}\left[\frac{e^{-\int_{t_1}^{t_2} \lambda_s ds} \left(\int_{t_1}^{t_2} \lambda_s ds \right)^k}{k!} \right]. \end{aligned} \tag{4.72}$$

Now consider the *aggregated process* $X_t = \int_0^t \lambda_s ds$, then by (4.72), choosing $t_2 = T$ and $t_1 = 0$, we deduce the following representation for the density function of N_T, for $T \geq 0$:

$$\mathbb{P}[N_T = k] = \mathbb{E}[f_k(X_T)], \tag{4.73}$$

with

$$f_k(x) = \frac{e^{-x}x^k}{k!}, \qquad \forall\, x \in \mathbb{R}, \;\; k \geq 0. \tag{4.74}$$

Thus, it is easy to note that the problem of finding the distribution of N_T, the point process, is equivalent to the problem of finding the distribution of the aggregated process. In particular we can state the following results based on the Feynman-Kac formula (1.21)

Theorem 4.4.4. *Let $T > 0$ and assume that the stochastic intensity is $\lambda_t = c(Y_t, t)$, $t \geq 0$, where $c : [0,T] \times \mathbb{R}^m \to [0,\infty)$. Let the process Y be $\mathcal{F}$-adapted and $\mathbb{R}^m$-valued which satisfies the stochastic differential equation with initial condition $y \in \mathbb{R}^m$*

$$dY_s^{y,t} = b(s, Y_s^{y,t})dt + \sigma(s, Y_s^{y,t})dB_s, \quad s \geq t \qquad Y_t^{y,t} = y, \tag{4.75}$$

B being an m-dimensional Brownian motion on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If the coefficient functions b, σ and c satisfy the hypotheses of Theorem 4.1.12, then for every $\alpha > 0$, the function

$$u_\alpha(y,t) = \mathbb{E}\left[e^{-\alpha \int_t^T c(Y_s^{y,t})\, ds}\right], \tag{4.76}$$

is the unique $C^{2,1}(\mathbb{R}^m \times [0,T))$ function, continuous on $\mathbb{R}^m \times [0,T]$, which satisfies the final value problem

$$\begin{cases} L_t u_\alpha + \frac{\partial u_\alpha}{\partial t} - c(y,t)u = 0, & \text{on } \mathbb{R}^m \times [0,T), \\ \\ u_\alpha(y,T) = 1. \end{cases} \tag{4.77}$$

L_t being the differential operator defined in (1.20). Moreover, for all $k \geq 0$ and $y \in \mathbb{R}^m$, it holds

$$\mathbb{P}[N_T = k] = \frac{(-1)^k}{k!} \frac{d^k}{d\alpha^k}_{|\alpha=1} u_\alpha(y,0). \tag{4.78}$$

On the basis of equation (4.78), we are able to compute the density function of the point process N. Unfortunately, the partial differential equation in (4.77) does not always admit a closed-form solution. This is available only in few special situations.

This, together with their flexibility, is the reason for the success of using affine models. In an affine model, if the function c is an affine function with respect to Y, then the function u_α can be found by solving a system of ordinary differential equations. The following result describes this property.

Theorem 4.4.5. *Let $T > 0$ and assume that $\lambda_t = c(Y_t)$, $t \geq 0$, where $c : \mathbb{R}^m \to \mathbb{R}$ is an affine function. Suppose that the process Y is also affine, then, for every $\alpha > 0$, there exist functions $z_\alpha \in C^1([0,T];\mathbb{R})$, $w_\alpha = (w_{1,\alpha}, w_{2,\alpha}, \dots, w_{m,\alpha}) \in C^1([0,T]\ \mathbb{R}^m)$ such that*

$$u_\alpha(y,t) = \exp\left(z_\alpha(t) + \langle w_\alpha(t), y\rangle\right) \tag{4.79}$$

$\langle\cdot,\cdot\rangle$ being the standard scalar product on $\mathbb{R}^m$. The functions z_α, w_α satisfies a system of ordinary differential equations of Riccati type.

We present an example of affine structure for the intensity process λ which allows a closed-form expression for the probability in (4.73)

Example 4.4.6. *Let λ be the process satisfying the following stochastic differential equation:*

$$d\lambda_t = k(\theta - \lambda_t)dt + \sigma\sqrt{\lambda_t}dB_t, \qquad \lambda_0 > 0. \tag{4.80}$$

where B is a 1-dimensional Brownian motion and suppose that $c(y) = y$, for all $y \in \mathbb{R}$. Hence $\lambda = Y$. The parameters k,θ and σ are all positive and satisfy Feller's [80] condition $2k\theta \geq \sigma^2$. This condition ensures the positivity of the process λ, whose state space becomes the interval $[0,+\infty)$. This model exhibits mean reversion of the intensity, causing the intensity to be pulled downward when it is above the long run average intensity and be pulled upward when it is below the long run average intensity. The coefficient k is the speed of this mean reversion, θ is the long run average intensity and σ is called volatility.
Model (4.80) was developed in [81], [82] and has been extensively applied in financial mathematics to describe the term structure of interest rates.
The random variable λ_t, conditioning with respect to λ_s, is distributed as a chi-square. By Theorem 4.4.5, a closed form solution for the function u_α can be derived:

$$u_\alpha(\lambda,t) = A_\alpha(T-t;k,\theta,\sigma)e^{-B_\alpha(T-t;k,\theta,\sigma)\lambda}, \tag{4.81}$$

for any $t \in [0,T]$, $\lambda > 0$. Here the function A_α and B_α have the following expressions. If $\alpha = 1$, it holds

$$A_1(T-t;k,\theta,\sigma) = \exp(z_1(t)) = \left[\frac{2\gamma e^{(k+\gamma)(T-t)/2}}{(\gamma+k)(e^{\gamma(T-t)}-1)+2\gamma}\right]^{2k\theta/\sigma^2}, \tag{4.82}$$

$$B_1(T-t;k,\theta,\sigma) = -w_1(t) = \frac{2(e^{\gamma(T-t)}-1)}{(\gamma+k)(e^{\gamma(T-t)}-1)+2\gamma}, \tag{4.83}$$

with $\gamma = \sqrt{k^2+2\sigma^2}$. For the general case, it suffices to replace the coefficients k, θ,σ respectively with $k_\alpha = k$, $\theta_\alpha = \theta\alpha$ and $\sigma_\alpha = \sqrt{\alpha}\sigma$, that is

$$A_\alpha(T-t;k,\theta,\sigma) = A_1(T-t;k,\alpha\theta,\sqrt{\alpha}\sigma), \tag{4.84}$$

and

$$B_\alpha(T-t;k,\theta,\sigma) = B_1(T-t;\alpha\theta,\sqrt{\alpha}\sigma). \tag{4.85}$$

Thus, the density function of N_T can be easily computed using Equation (4.78) in Theorem 4.4.4.

5

Statistical data analysis

This Chapter introduces a statistical model to describe the length of stay of hospital patients. The proposed model overcomes some of the limitations of previous models by using a Phase-Type Gamma distribution which is able to capture the data characteristics in a more accurate way. The model is tested on a case study based on the Campus Bio-Medico hospital database. The purpose is to demonstrate how statistical techniques can be applied to describe healthcare proxies. According to the framework presented in Chapter 3, we propose some tools that support the decision support system for **resource capacity planning**. The aim is to provide guidance for the **tactical** and **operational** level. The tactical planning addresses the organization of the operation and the execution of the healthcare delivery process. The operation planning involves the short term decision making related to the execution of the healthcare delivery process. Length of stay, discharge rate, and admission rate are reliable and valid proxies for the resource consumption. Therefore a statistical analysis of these proxies provides strong indication for the decision process.

It is well recognized that statistical analysis of healthcare resource use and cost data poses a number of difficulties. These non-negative data often exhibit substantial positive skewness, can have heavy tails and are often multimodal. The traditional approach for handling such non-normal data in medical statistics has been to use non-parametric methods, such as rank order statistics.

5.1 Phase-type distributions

Phase-type (PH) distributions constitute a very versatile class of distributions. They have been used in a wide range of stochastic modelling applications in areas as diverse as telecommunications, finance, biostatistics, queueing theory, drug kinetics, and survival analysis. Their use in modelling systems in the healthcare industry, however, has so far been limited. There have been a number of papers written on the application of PH distributions in the healthcare literature, but as we shall see, the number of areas where they have been used is rather limited. Before describing PH distribution we have to briefly introduce the **function of a matrix** definition, the **exponential distribution**, and the **continuous-time Markov chains** (by way of an example).

5.1.1 Function of a matrix

Given a matrix $n \times n$ matrix $\boldsymbol{T}$ with real entries and a scalar function $g : \mathbb{R} \to \mathbb{R}$, we are interested in definition of $g(\boldsymbol{T})$ which specifies $g(\boldsymbol{T})$ to be a matrix of the same dimensions of $\boldsymbol{T}$.

It is a standard result that the matrix $\boldsymbol{T}$ can be expressed in the Jordan canonical form:

$$\boldsymbol{Z}^{-1}\boldsymbol{T}\boldsymbol{Z} = \mathrm{diag}(J_1, J_2, \ldots, J_p), \tag{1.1}$$

$$\boldsymbol{J}_k = \boldsymbol{J}_k(\lambda_k) = \begin{bmatrix} \lambda_k & 1 & & \\ & \lambda_k & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_k \end{bmatrix} \in \mathbb{C}^{m_k \times m_k} \tag{1.2}$$

where $\boldsymbol{Z}$ is nonsingular and $m_1 + m_2 + \cdots + m_p = n$. The Jordan matrix $\boldsymbol{J}$ is unique up to the ordering of the blocks $\boldsymbol{J}_i$, but the transforming matrix $\boldsymbol{Z}$ is not unique. Denote by $\lambda_1, \ldots, \lambda_s$ the distinct eigenvalues of $\boldsymbol{T}$ and let n_i be the order of the largest Jordan block in which λ_i appears, which is called the *index* of λ_i. We need the following terminology.

Definition 5.1.1. *s The function g is said to be defined on the spectrum of $\boldsymbol{A}$ if the values*

$$D^j g(\lambda_i), \qquad j = 0, \ldots, n_i - 1, \quad i = 1, \ldots s \tag{1.3}$$

exist. These are called the values of the function g on the spectrum of A. Here $D^j g$ denotes the jth *derivative of g.*

In most cases of practical interest g is given by a formula. However, the following definition of $g(\boldsymbol{A})$ requires only the values of g on the spectrum of $\boldsymbol{A}$; it does not require any other information about g. It is only when we need to make statements about global properties (such as the continuity) that we will need to assume more about g.

Definition 5.1.2. *(***Jordan canonical form***). Let g be defined on the spectrum of $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ and let $\boldsymbol{A}$ have the Jordan decomposition form (1.1). Then*

$$g(\boldsymbol{A}) = \boldsymbol{Z} g(\boldsymbol{J}) \boldsymbol{Z}^{-1} = \boldsymbol{Z} \mathrm{diag}(g(J_k)) \boldsymbol{Z}^{-1}, \tag{1.4}$$

where

$$g(J_k) = \begin{bmatrix} g(\lambda_k) & Dg(\lambda_k) & \cdots & \frac{D^{m_k-1} g(\lambda_k)}{(m_k-1)!} \\ & g(\lambda_k) & \ddots & \vdots \\ & & \ddots & Dg(\lambda_k) \\ & & & g(\lambda_k) \end{bmatrix} \tag{1.5}$$

we observe that the definition yields a matrix $g(\boldsymbol{A})$ that can be shown to be independent of the particular Jordan canonical form that is used. Second, note that if $\boldsymbol{A}$ is diagonalizable, then the Jordan canonical form reduces to an eigenvalue decomposition $\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^{-1}$, with $\boldsymbol{D} = \mathrm{diag}(\lambda_i)$ and the columns of $\boldsymbol{Z}$ are the eigenvectors of $\boldsymbol{A}$. Hence Definition 2 yields

$$g(\boldsymbol{A}) = \boldsymbol{Z} g(\boldsymbol{D}) \boldsymbol{Z}^{-1} = \boldsymbol{Z} \mathrm{diag}(g(\lambda_i)) \boldsymbol{Z}^{-1}. \tag{1.6}$$

Therefore for diagonalizable matrices $g(\boldsymbol{A})$ has the same eigenvectors as $\boldsymbol{A}$ and its eigenvalues are obtained by applying g to those of $\boldsymbol{A}$.
Finally, we remark that the equation (1.5) can be proved by applying Taylor series considerations.

5.1.2 Exponential distribution

The exponential distribution is ubiquitous in stochastic modelling, mainly because of its simplicity and ability to model random lengths of time reasonably well. For example, it has been used to model the length of stay in a hospital bed, or the time between presentations to an emergency department.

A continuous non-negative random variable T is distributed according to an exponential distribution with parameter $\lambda > 0$, if its *distribution* (or *cumulative distribution*) *function*, defined for $t \geq 0$, is given by

$$F(t) = \mathbb{P}(T \leq t) = 1 - e^{-\lambda t}. \tag{1.7}$$

The *density* (or *probability density*) *function* of T defined for $t \geq 0$, is given by

$$f(t) = \lambda e^{-\lambda t}. \tag{1.8}$$

The *expected value* of T, or its mean, is $\mathbb{E}(T) = \frac{1}{\lambda}$, and its *variance* is $\mathbb{V}(T) = \frac{1}{\lambda^2}$.

The simplicity in using the exponential distribution in stochastic modelling is not only due to its formulation in terms of a single parameter λ , but also because of the so called *memoryless property*. That is, for, $t, s \geq 0$, $\mathbb{P}(T > s + t | T > t) = \mathbb{P}(T > s)$.

The memoryless property enables simple expressions for many performance measures of stochastic models that use the exponential distribution to be given. We also remark here that the exponential distribution is the only continuous distribution that exhibits the memoryless property.

5.1.3 Markov chains

Figure 5.1 shows the *state transition diagram* for a *finite-state continuous time Markov chain*. The Markov chain consists of four states labelled $0, 1, 2$, and 3. States $1, 2$, and 3 are called *transient states*, and state 0 an *absorbing state*. A state is *transient* if once it has been reached, the probability of returning to it is less than one, and a state is *absorbing* if once it has been reached the process stops. We choose any of states, $0, 1, 2$ and 3 according to the probabilities $\frac{1}{10}, \frac{1}{2}, \frac{3}{10}$ and $\frac{1}{5}$ respectively. The probability of being instantaneously absorbed, that is $\frac{1}{10}$, is known as the *point mass at zero*.

Suppose that state 1 has been chosen. We spend an exponentially distributed length of time with parameter $\lambda = 12$ there. This parameter can be interpreted as the (average) rate of movement out of state 1. Once we have completed this time we move to either state 0 or state 2 with (average) rates 8 and 4, respectively. Alternatively, we move from state 1 to state 0 with probability $\frac{8}{12} = \frac{2}{3}$, or to state 2 with probability $\frac{4}{12} = \frac{1}{3}$. If we chose state 0 we stop, but if we chose state 2 we spend an exponentially distributed length of time with $\lambda = 10$ there, and so on until absorption. The various

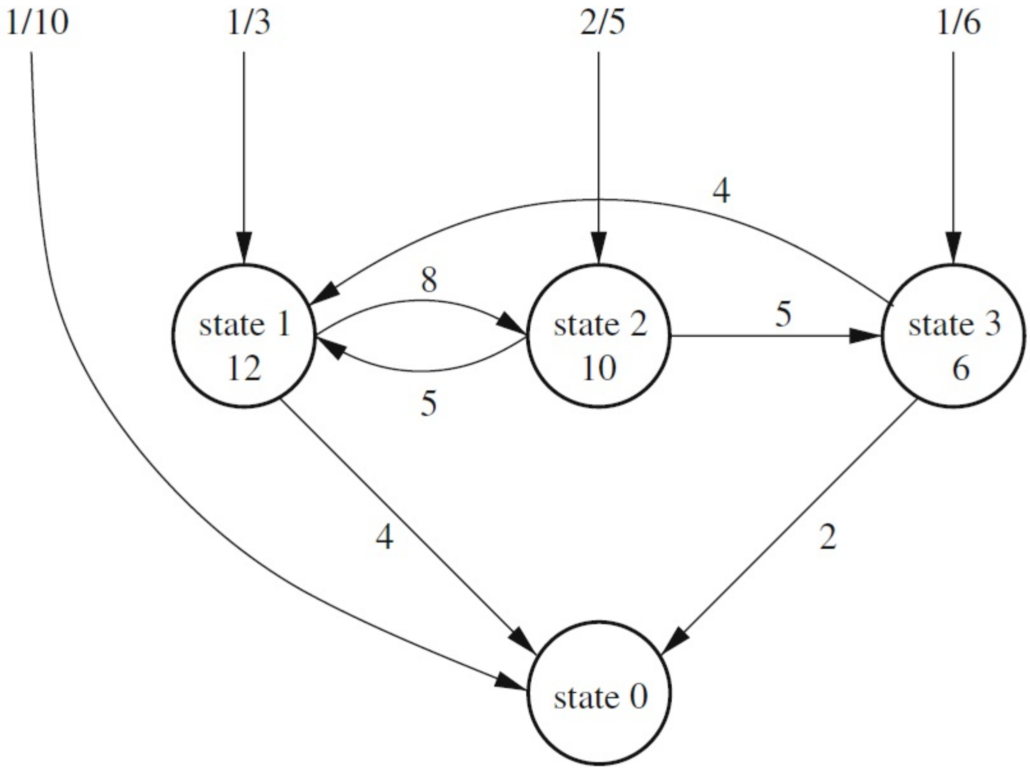


Figure 5.1: Example of Markov chain - State transition diagram of a 4-state continuous-time Markov chain with one absorbing state

rates have been chosen so that absorption occurs with *probability one*.
In order to describe the Markov chain we need three descriptors:

1. A *state space*

$$S = \{0, 1, 2, 3\}. \tag{1.9}$$

2. An *initial state probability distribution*

$$(\alpha_0, \boldsymbol{\alpha}) = \begin{pmatrix} \frac{1}{10} & \frac{1}{3} & \frac{2}{5} & \frac{1}{6} \end{pmatrix} \tag{1.10}$$

which governs the selection of the initial state, α_0 being the point mass at zero being the point mass at zero.

3. An *infinitesimal generator*

$$\mathbf{Q} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 4 & -12 & 8 & 0 \\ 0 & 5 & -10 & 5 \\ 2 & 4 & 0 & -6 \end{bmatrix}$$

which governs the transitions between states.

The rows (labelled $0, 1, 2$ and 3) of $\mathbf{Q}$ correspond to the state we move from, and the columns (labelled $0, 1, 2$ and 3) correspond to the state we move to. The zero-*th* row consists of all zeros because once we have reached state 0 (absorption) we stay there. The remaining diagonal entries are negative and the off diagonal entries non-negative, with all row sums equal to zero. The absorption rates from states $1, 2$ and 3 are $4, 0$ and 2, respectively. The distribution of time from start to finish (absorption), in the Markov chain, is said to have a *PH* distribution. Consider a continuous-time Markov chain on a finite state space $S = \{0, 1, 2, ..., p\}$ where state 0 is absorbing. Let the initial state probability distribution be

$$(\alpha_0, \boldsymbol{\alpha}) = (\alpha_0, \alpha_1, ..., \alpha_p), \quad \text{with} \sum_{i=0}^{p} \alpha_i = 1, \tag{1.11}$$

and the infinitesimal generator be $\mathbf{Q}$. The random variable that is defined as the time to absorption, is said to have a (*continuous*) PH distribution.
The infinitesimal generator for the Markov chain can be written in block-matrix form as

$$\mathbf{Q} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{t} & \mathbf{T} \end{bmatrix}.$$

Here, $\mathbf{0}$ is a $1 \times p$ vector of zeros. The vector $t = (t_{10,t_{20},...,t_{p0}})'$ (the prime denoting transpose) where, for $i = 1, 2, ..., p$, $t_{i0} \geq 0$ with at least one of the t_{i0} positive, is the absorption rate from state i. The $p \times p$ matrix $\mathbf{T} = [t_{ij}]$ is such that, for $i, j = 1, 2, ..., p$, with $i \neq j$

$$t_{ij} \geq 0 \tag{1.12}$$

and

$$t_{ii} = -\sum_{j=0, j\neq i}^{p} t_{ij} \tag{1.13}$$

that is, $\mathbf{t} = \mathbf{Te}$ where $\mathbf{e}$ is a $p \times 1$ vector of ones. The PH distribution is said to have a representation $(\boldsymbol{\alpha}, \mathbf{T})$ of order p. The matrix $\mathbf{T}$ is referred to as a *PH generator*. The point mass at zero α_0 is completely determined by $\boldsymbol{\alpha}$ and therefore does not need to appear in the expression for the representation. Typically representations are not unique and at least one representation of minimal order must exist. Such a representation is known as a *minimal representation*, and the *order of the PH distribution* itself is defined to be the order of any of its minimal representations. To ensure absorption in a finite time with probability one, we require that every non-absorbing state is transient. This statement is equivalent to $\mathbf{T}$ being invertible.

An additional requirement on the PH representation $(\boldsymbol{\alpha}, \mathbf{T})$ is that there are no superfluous phases. That is, each phase in the Markov chain defined by $\boldsymbol{\alpha}$ and $\mathbf{T}$ has a positive probability of being visited before absorption. If this is the case, then we say that the PH representation is *irreducible*. If the representation is reducible, we can form an irreducible representation by simply deleting those states that are superfluous.

A PH distribution with representation $(\boldsymbol{\alpha}, \mathbf{T})$ has distribution function, defined for $t \geq 0$, given by

$$F(t) = \begin{cases} \alpha_0 & t = 0 \\ 1 - \langle \boldsymbol{\alpha}, \exp{(\mathbf{T}t)}\boldsymbol{e} \rangle & t \geq 0 \end{cases}$$

with respect to t gives the corresponding density function, defined for $t > 0$,

$$f(t) = -\langle \boldsymbol{\alpha}, \exp{(\mathbf{T}t)}\boldsymbol{Te} \rangle. \tag{1.14}$$

Now we will give some examples *PH* distribution.

Exponential distribution The minimal representation of an exponential distribution is

$$\boldsymbol{\alpha} = (1) \tag{1.15}$$

$$\mathbf{T} = (-\lambda). \tag{1.16}$$

Generalized Erlang distribution The order p generalized Erlang distribution can be described using a state transition diagram that has p states in series, see the state diagram below

$$\begin{array}{ccccccc} \downarrow & & & & & & \\ \mathbf{1} & \xrightarrow{\lambda_1} & \mathbf{2} & \xrightarrow{\lambda_2} & \cdots & \xrightarrow{\lambda_{p-1}} & \mathbf{p} \\ & & & & & & \lambda_p \downarrow \end{array}$$

It is easy to see, without loss of generality, that the states can be ordered so that the rates $0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_p$. The representation for the generalized Erlang distribution corresponding to the state transition diagram is

$$\boldsymbol{\alpha} = (1\ 0\ ...\ 0)$$

$$\mathbf{T} = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & \cdots & 0 \\ 0 & -\lambda_2 & \lambda_2 & \cdots & 0 \\ 0 & 0 & -\lambda_3 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \cdots - & \lambda_p \end{pmatrix}$$

The density function for an Erlang distribution of order p, defined for $t > 0$ is given by

$$f(t) = \frac{\lambda^p t^{p-1} e^{-\lambda t}}{p!} \tag{1.17}$$

Hyper-exponential distribution The order p hyper-exponential distribution can be described using a state transition diagram with p states in parallel, see below

$$\begin{array}{cccc} \alpha_1 \downarrow & \alpha_2 \downarrow & & \alpha_p \downarrow \\ \mathbf{1} & \mathbf{2} & \cdots & \mathbf{p} \\ \lambda_1 \downarrow & \lambda_2 \downarrow & & \lambda_p \downarrow \end{array}$$

Clearly, without loss of generality, the states can be ordered so that the rates $0 < \lambda_1 < \lambda_2 < ... < \lambda_p$. The corresponding representation is

$$\boldsymbol{\alpha} = (\alpha_1\ \alpha_2\ ...\alpha_{\ p})$$

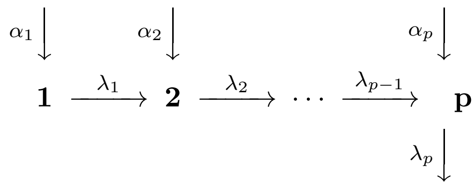
$$\mathbf{T} = \begin{pmatrix} -\lambda_1 & 0 & \cdots & 0 \\ 0 & -\lambda_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots - & \lambda_p \end{pmatrix}$$

with density function, defined for $t > 0$, given by

$$f(t) = \sum_{i=1}^{p} \alpha_i \lambda_i e^{\lambda_i t} \tag{1.18}$$

where, for $i = 1,\ 2, ..., p$, $\alpha_i > 0$ and $\sum_{i=1}^{p} \alpha_i = 1$

Coxian distribution The order p Coxian distribution has the following state transition diagram



These distributions have representations of the form

$$\boldsymbol{\alpha} = (\alpha_1\ \alpha_2\ ...\alpha_{\ p})$$

$$\mathbf{T} = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & \cdots & 0 \\ 0 & -\lambda_2 & \lambda_2 & \cdots & 0 \\ 0 & 0 & -\lambda_3 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \cdots - & \lambda_p \end{pmatrix}$$

Although it is not obvious, in this case, without loss of generality, the states can be ordered so that the rates $0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_p$ [83].

Unicyclic distribution The order p unicyclic distribution has representations of the form

$$\boldsymbol{\alpha} = (\alpha_1\ \alpha_2\ ...\alpha_{\ p})$$

$$\mathbf{T} = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & \cdots & 0 \\ 0 & -\lambda_2 & \lambda_2 & \cdots & 0 \\ 0 & 0 & -\lambda_3 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & -\lambda_{p-1} & \lambda_{p-1} \\ \mu_1 & \mu_2 & \cdots & \mu_{p-1} & -\lambda_p \end{pmatrix}$$

where for $i = 1, 2..., p-1$, $\mu_i \geq 0$, $0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_p$ and $\lambda_p > \sum_{i=1}^{p-1} \mu_i$. It was conjectured in [84] that every PH distribution of order p has a unicyclic representation of the same order, however, [85] showed that this is not, in general, the case.

5.1.4 Tail analysis

In order to evaluate the tail of the probability distribution and to provide an operational guideline in bed planning, we propose two indices, inspired by the theory of risk measures [86], [87], and [88], the **value-at-risk** (**VaR**) and the **expected shortfall** (**ESh**) related to the LoS probability distribution.
In its most general form, the VaR measures the potential loss in value of a risky asset or portfolio over a defined period for a given confidence interval. While VaR can be used by any entity to measure its risk exposure, it is used most often by commercial and investment banks to capture the potential loss in value of their traded portfolios from adverse market movements over a specified period; this can then be compared to their available capital and cash reserves to ensure that the losses can be covered without putting the firms at risk. Taking a closer look at VaR:

- To estimate the probability of the loss, with a confidence interval, we need to define the probability distributions of individual risks, the correlation across these risks and the effect of such risks on value. In fact, simulations are widely used to measure the value-at-risk for asset portfolio.

- The focus in value-at-risk is clearly on downside risk and potential losses.

- There are three key elements of VaR, a specified level of loss in value, a fixed time period over which risk is assessed and a confidence interval. The VaR can be specified for an individual asset, a portfolio of assets or for an entire firm.

Given a confidence level $\alpha \in (0,1)$, the VaR of a portfolio at α over the period t is given by the smallest number $k \in \mathcal{R}$ such that the probability of a loss over a time interval t greater than k is α. In the same way

$$VaR_t(\alpha) = F_{t+1}^{-1}(\alpha) = \inf\{k; F_{t+1}(k) \geq \alpha\} \tag{1.19}$$

where F_t is the cumulative distribution function of k. In our framework the risky asset is represented by the hospitalization cost of a patients. Thus, if the VaR associated to the LoS distribution is \$ 2 days, 95% confidence level, the probability of having a discharge in 2 days is grater than 95%. Understanding that VaR and quantiles are fundamentally related provides a key insight into computing VaR. The $VaR(\alpha)$ is q_α where q_α is the α-quantile of the portfolio. In most cases α is chosen to be some small quantile 5% or 10%.
The **expected shortfall** combines aspects of the VaR methodology with more information about the distribution of returns in the tail. Expected shortfall is defined as the expected value of the portfolio loss given a VaR exceedance has occurred

$$\mathrm{ESh}_\alpha(\mathrm{t}) = \mathbb{E}\big[\mathrm{k}|\mathrm{k} \geq \mathrm{F}^{-1}(\alpha)\big]. \tag{1.20}$$

5.2 A new model for the length of stay of hospital patients

In this section is introduced an innovative approach to draw the hospital length of stay (LoS) distribution.
Hospital LoS is considered to be a reliable and valid proxy for measuring the consumption of hospital resources [28]. However average LoS can be misleading since the underlying distribution is not symmetric. Therefore model based on the average LoS can not describe the underlying distribution of patients [89], [12].
We also show how well know models [90] should be improved in order to capture some statistical features observed in real situations. More specifically the application of a generalization of the phase-type distributions is demonstrated [91]. Our case study is based on a data set originated from the University Polyclinic Campus Bio-Medico and it is shows that standard models are not adequate to describe the flow patients. To overcome such limit we propose a generalization which is based on a few number of parameters and supplies a phenomenological interpretation of data. Furthermore to

afford an operational guideline for planning and management of hospital beds we propose a method to evaluate the forecasting ability. Precisely we reword some measures used in financial risk management like the Value at Risk and Expected Shortfall [86], [87], [88] in order to show the goodness of the model.
In next section a short description of Campus Bio-Medico hospital and some related statistics are summarized. The proceeding section follows with a description of phase-type distributions and the description of our generalization.

5.2.1 Data

The data set analysed originates from the Campus Bio-Medico hospital database and concerns patients of different departments. This hospital is a medium size structure within the context of Italian health care system. Moreover this hospital offers a high quality service compared to the national health system.
This work focuses on the flow of patients within four departments: General Surgery, Geriatrics, Clinical Medicine and Endoscopy. Those departments have highest number of hospitalizations within Campus Bio-Medico hospital. All of the Length of Stay (LoS) and occupancy statistics refer to patients that were hospitalized from January 2007 to December 2011. The basic descriptive statistics of the patients LoS for each department are reported in Table 5.1. The values of skewness and kurtosis show a

	no. patients (N)	**Mean**	**Variance**	**Skewness**	**Kurtosis**
Clinical Medicine	3890	6.45	27.34	1.87	6.72
Endoscopy	3620	3.80	25.32	1.49	5.68
General Surgery	8787	6.04	19.14	2.08	8.77
Geriatrics	3934	8.86	8.69	3.38	17.71

Table 5.1: Descriptive statistics of LoS. The mean value reppresents the average number number of days spent in each department.

strong difference between the distributions of each department. The density function of the departments are asymmetric, they can only take positive values, and leptokurtic, execute acute peak around the mean and have fatter tails. The long gradual tail to the right of the distribution is given by a very small number of patients that stay in hospital for a considerable amount of time. Furthermore Endoscopy department exhibit strongly differences form other departments; kurtosis is bigger and the average time spent is smaller. In Section 5.2.4 is shown that the proposed model achieves good result for all departments.

We remark the importance of longer stay patients, in fact although the majority of patients is discharged after a short period, beds are mostly occupied by patients that stay for a longer period. Moreover Campus Bio-Medico hospital does not have the emergency department.

5.2.2 Model

In order to describe the empirical distribution it is used a generalization of *phase-type distribution* (*PH distribution*). *PH* distribution is a very versatile class of distributions. Introduced by Neuts in [92], phase-type distributions are defined via a continuous time homogeneous Markov chain $X(t)$ on a finite state space $\{0, 1, ...n\}$, $n \geq 0$, see [93]. The state 0 is absorbing and Markov chain is assumed to be irreducible. Then the random variable representing the first hitting time of 0 is

$$\tau = \inf\left\{t \geq 0 : X(t) = 0\right\}, \tag{2.21}$$

which is finite almost everywhere, and its distribution is called a phase-type distribution. Let the initial state probability distribution be $\boldsymbol{\alpha} \in \mathbb{R}^n$ such that

$$\sum_{i=1}^{n} \alpha_i = 1, \quad \alpha_i \geq 0 \quad \forall i = 1, 2, \ldots n. \tag{2.22}$$

Let $\boldsymbol{T} = [\lambda_{ij}]$ for $i, j = 1, 2, \ldots, n$ be a $n \times n$ real matrix such that for $i \neq j$, $\lambda_{ij} \geq 0$, satisfying

$$\lambda_{ii} \leq -\sum_{j=0,\ j\neq i}^{n} \lambda ij. \tag{2.23}$$

The *PH* distribution is said to have a representation $(\boldsymbol{\alpha}, \boldsymbol{T})$ of order n if the probability density function of τ is as follows [83], [94]:

$$ph(t) = -\boldsymbol{\alpha} e^{\boldsymbol{T}t}\boldsymbol{T}\boldsymbol{e}, \quad t > 0. \tag{2.24}$$

where $\boldsymbol{e}$ is a $n \times 1$ vector of ones. When $\boldsymbol{T}$ is diagonal, the density function (2.24) is called an *hyperexponential distribution*.

In this work we shall consider an extension of (2.24). Given the density function $\gamma(t; k,\theta)$ of a Gamma distribution with parameters k, θ and a diagonal matrix $\boldsymbol{T} = -\text{diag}(\lambda_1, \lambda_2, \dots, \lambda_\text{n})$, with $\lambda_i \geq 0$ for $i = 1, 2, \dots n$, we define

$$f(t) = \sum_{i=1}^{n} \alpha_i \lambda_i \gamma(t\lambda_i; k,\theta), \quad \text{for } t > 0, \tag{2.25}$$

that is

$$f(t) = \frac{\theta^k}{\Gamma(k)} \sum_{i=1}^{n} \alpha_i \lambda_i^k t^{k-1} e^{-\theta\lambda_i t}. \tag{2.26}$$

In the rest of the chapter we refer to 2.25 as the density function of *phase-type Gamma* distribution (*PHGamma*).

Remark We observe that equation (2.25) may be generalized by using the following form:

$$h(t) = \sum_{i=1}^{n} \alpha_i g(t\lambda_i)\lambda_i \tag{2.27}$$

where g is a given density function. In the case where g is analytic on $\mathbb{R}$, then for any matrix $\boldsymbol{T}$, $g(\boldsymbol{T}t)$ is well define and h takes the following form:

$$h(t) = -\boldsymbol{\alpha} g(\boldsymbol{T}t)\boldsymbol{T}\boldsymbol{e}. \tag{2.28}$$

which is a generalized version of (2.24). In general, a function $g : \mathbb{R} \to \mathbb{R}$ is a probability density function of a continuous random variable if

- g is a non-negative Lebesgue-integrable function;

- $\int_{-\infty}^{+\infty} g(x)\, dx = 1$.

Then we have the following result

Proposition 5.2.1. *Let g be a probability density function. For every $\alpha_1, \alpha_2, \dots, \alpha_n \geq 0$ with $\alpha_1 + \alpha_2 + \cdots + \alpha_n = 1$, and $\lambda_1, \lambda_2, \dots, \lambda_n > 0$, the function*

$$h(\tau) = \sum_{i=1}^{n} \alpha_i g(-\lambda_i\tau)\lambda_i \tag{2.29}$$

is a probability density function.

Proof. It suffices to prove that $\int_{-\infty}^{+\infty} h(\tau)\, d\tau = 1$. In fact, we have

$$\int_{-\infty}^{+\infty} h(\tau) d\tau = \sum_{i=1}^{n} \alpha_i \int_{-\infty}^{+\infty} g(-\tau\lambda_i)\lambda_i d\tau. \tag{2.30}$$

For any integral, in the sum, we apply the change of variable $\tau = -s/\lambda_i$. Thus

$$\sum_{i=1}^{n} \alpha_i \int_{-\infty}^{+\infty} g(-\tau\lambda_i)\lambda_i d\tau = \sum_{i=1}^{n} \alpha_i \int_{-\infty}^{+\infty} g(s) ds = \sum_{i=1}^{n} \alpha_i = 1. \tag{2.31}$$

□

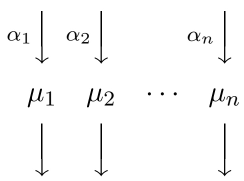
In the same way it is possible to evaluate the mean and the variance of a random variable distributed according to a PHGamma density function:

$$\mathbb{E}[X] = \frac{k}{\theta} \sum_{i=1}^{n} \frac{\alpha_i}{\lambda_i}, \tag{2.32}$$

$$\mathbb{E}[(X - \mathbb{E}[X])^2] = \frac{(k^2 + k)}{\theta^2} \sum_{i=1}^{n} \frac{\alpha_i}{\lambda_i^2} - \left(\frac{k}{\theta} \sum_{i=1}^{n} \frac{\alpha_i}{\lambda_i} \right)^2, \tag{2.33}$$

$$\mathbb{E}[e^{Xt}] = \sum_{i=1}^{n} \alpha_i \left(1 - \frac{1}{\theta} \frac{t}{\lambda_i} \right)^{-k}. \tag{2.34}$$

Here the moment generating function is well defined on the interval $t < \theta \min(\lambda_1, \lambda_2, \ldots, \lambda_n)$. The density function 2.25 provides a useful tool to describe LoS data; the standard PH-distribution can not describe leptokurtic densities; furthermore the use of the PHGamma distribution allows to reach the peak and fit the fat tail of the data. Moreover, like for the Hyperexponential distribution, the PHGamma can be described using the following state transition diagram with n states in parallel:



where the average time spent μ_i in each state is evaluated by

$$\mu_i = \frac{1}{\lambda_i} \frac{k}{\theta}. \tag{2.35}$$

5.2.3 Optimization method

According to [95], [96], and [97] maximum log-likelihood estimation is used in order to fit the parameters of LoS distributions. The number $N^{(j)}$ of patients for each department are reported in Table 5.1, where j represents j-th department. The log-likelihood objective function is given by

$$\arg\max_{\Theta \in \mathcal{D}} \sum_{i=1}^{N^{(j)}} \log(f(t_i; \boldsymbol{\Theta})), \tag{2.36}$$

where $f(\cdot; \boldsymbol{\Theta})$ is the density function of the considered model depending on the set of parameters described by a vector $\boldsymbol{\Theta}$.
The optimization routine is implemented in the MATLAB framework using existing algorithm based on the interior-point method. Parameters calibration starts from a Halton quasi-random sequence of $10 \times n$ initial points [98], [99]. A Halton sequence is a deterministic sequence of numbers that provides well-spaced combinations from an interval and provides negative correlation between simulated probability for individuals. This negative correlation reduces the variance in the log-likelihood function, hence it is a suitable method for the selection of the parameters estimation process starting point. A sequential procedure is adopted whereby increasing numbers of phases n are tried starting with $n = 1$ until little improvements in the fit of the data can be obtained by adding a new phase. The number of the phases that allows the best compromise between model complexity and goodness of the fit is chosen. In Table 5.2 are reported the values of the objective function related to the PHGamma model. The analysis of the data shows that the use of models with more than 3 states provide negligible improvements in the performance. Hence we fix the number of the phases to 3.

5.2.4 Estimation Results

The PHGamma distribution provides a very versatile tool to describe a wide class of density functions. Using a few parameters is possible to describe distributions that exhibit different characteristics. The performance of the model are compared with different phase-type distributions. It is chosen the *Coxian* distribution that is already used in modelling hospital LoS [90], [96]. To underline the difference of using a Gamma density function instead of using an Exponential density function, the *Hyperexponential* distribution is used, see [100]. Moreover in order to complete the analysis, the *Unicyclic*

	$n = 1$	$n = 2$	$n = 3$
Clinical Medicine	-2.6299	-2.5866	-2.5785
Endoscopy	-2.0857	-1.9074	-1.8708
General Surgery	-2.7100	-2.5973	-2.5659
Geriatrics	-2.8736	-2.8603	-2.8484
	$n = 4$	$n = 5$	$n = 6$
Clinical Medicine	-2.5731	-2.5700	-2.5576
Endoscopy	-1.8420	-1.7731	-1.6907
General Surgery	-2.5655	-2.5484	-2.5113
Geriatrics	-2.8464	-2.8422	-2.8353

Table 5.2: Objective log-likelihood function values for each department of Campus Bio-Medico LoS distribution estimation.

distribution is also used, see [101], [85]. In Figures 5.2 and 5.3, a three-states model is estimated for each distribution. In order to obtain the LoS data it is considered the difference between the discharge and the admission date. Data analysis of Table

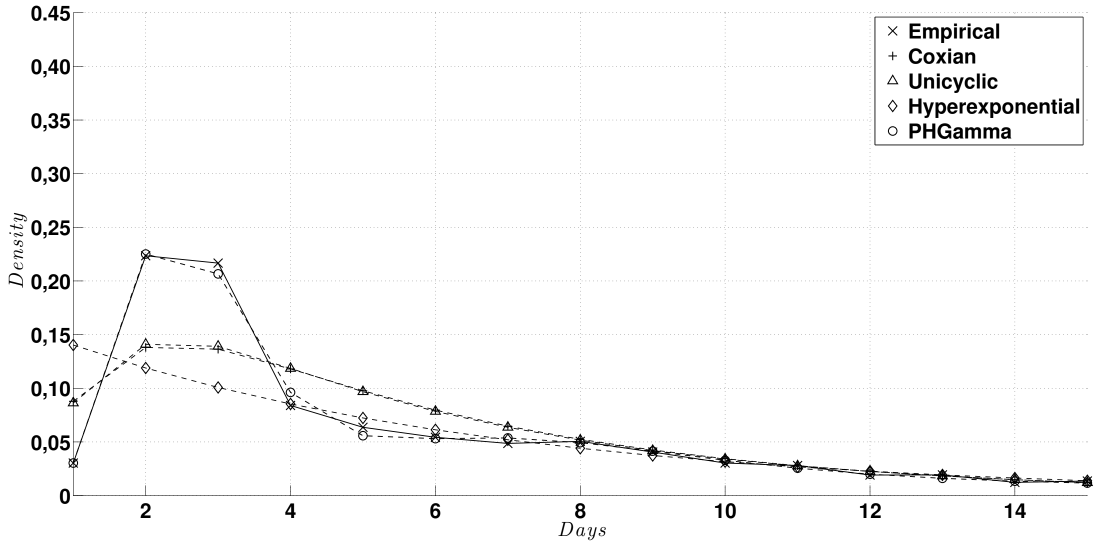


Figure 5.2: LoS estimated density functions of Campus Bio-Medico General Surgery department - Comparison between 3-states phase-type distributions

5.1 shows that the most of patients stay in hospital less then 15 days; in this interval PHGamma model perform a better fit of the distribution. Longer stay patients are very rare therefore is difficult to estimated the probability distribution. Moreover all the models exhibit same tail behaviour. Therefore, in next section, we introduce a new

cumulative measure to adequately evaluate the tail performance of models.
The results show that the proposed model provides a better fit of the data for all the departments. PHGamma distributions can describe data that have different behaviours: Figure 5.2 shows how the model can fit a density function that has 2 different peaks, instead Figure 5.3 shows a case in which to data that exhibit a unique high peak. To

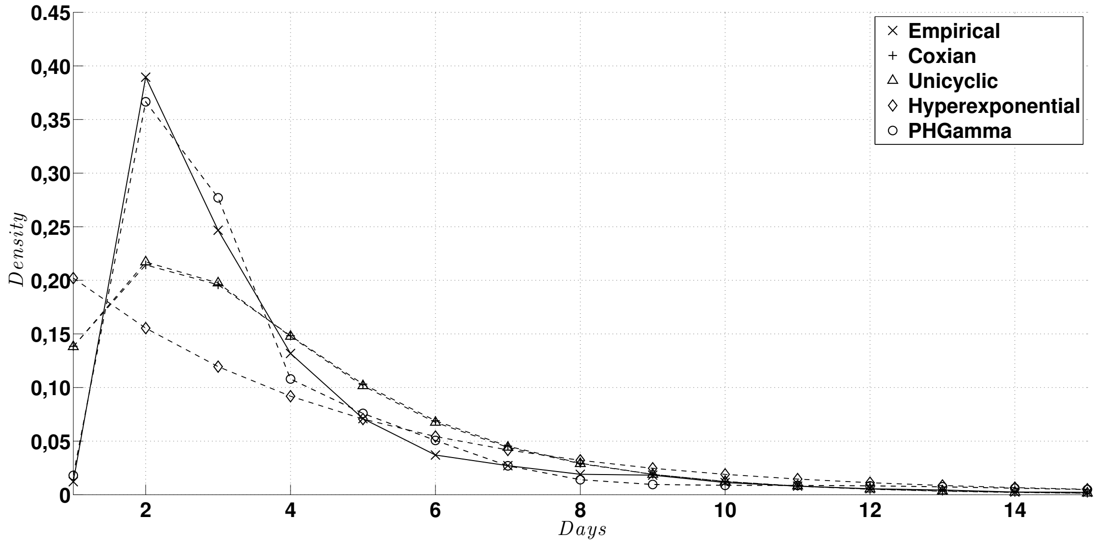


Figure 5.3: LoS estimated density functions of Campus Bio-Medico Endoscopy department - Comparison between 3-states phase-type distributions

evaluate the performance of those models, we also use the following four discrepancy measures: the average prediction error (APE), the average absolute error (AAE), the root mean-square error (RMSE), and the average relative prediction error (ARPE). Let F be the cumulative density function and t_k for $k = 1, 2, \ldots, K$ the evaluation time. Then the four error estimators are defined as follows:

$$\begin{aligned} AAE &= \sum_{k=1}^{K} \frac{|F(t_k)-\hat{F}(t_k)|}{K} \\ APE &= \sum_{k=1}^{K} \frac{|F(t_k)-\hat{F}(t_k)|}{F(t_k)} \\ ARPE &= \frac{1}{K}\sum_{k=1}^{K} \frac{|F(t_k)-\hat{F}(t_k)|}{F(t_k)} \\ RMSE &= \sqrt{\sum_{k=1}^{K} \frac{(F(t_k)-\hat{F}(t_k))^2}{K}} \end{aligned} \tag{2.37}$$

These measures are shown in Table 5.12. In particular we observe that the PHGamma distribution achieves the smallest errors for all these indicators, at least one order of magnitude better then the other models.
The PHGamma n-phases model requires the estimations of $2 \times n + 2$ parameters that

			PHGamma	
	AAE	**APE**	**ARPE**	**RMSE**
General Surgery	0.0020	0.0855	0.0028	0.0027
Geriatrics	0.0039	0.4032	0.0134	0.0054
Clinical Medicine	0.0031	1.3340	0.0445	0.0037
Endoscopy	0.0052	0.6688	0.0223	0.0064
			Coxian	
	AAE	**APE**	**ARPE**	**RMSE**
General Surgery	0.0164	2.6441	0.0881	0.0287
Geriatrics	0.0185	2.7924	0.0931	0.0251
Clinical Medicine	0.0120	10.0621	0.3354	0.0168
Endoscopy	0.0183	10.9755	0.3659	0.0361
			Unicyclic	
	AAE	**APE**	**ARPE**	**RMSE**
General Surgery	0.0157	2.5602	0.0853	0.0274
Geriatrics	0.0184	2.8347	0.0945	0.0251
Clinical Medicine	0.0111	9.8163	0.3272	0.0157
Endoscopy	0.0177	10.9211	0.3640	0.0352
			Hyperexponential	
	AAE	**APE**	**ARPE**	**RMSE**
General Surgery	0.0892	6.7190	0.2240	0.0909
Geriatrics	0.1121	23.3102	0.7770	0.1175
Clinical Medicine	0.1032	38.1336	1.2711	0.1071
Endoscopy	0.1418	20.0347	0.6678	0.1455

Table 5.3: Summary of four discrepancy measures for the fitting of department LoS distribution of Campus Bio-Medico.

are reported in Table 5.5. The value of the moments, calculated using the moment generating function, are reported in Table 5.6. The difference between the real value, reported in Table 5.1 and the estimated is negligible hence the density estimated supply a correct estimation of the first moments. Furthermore the use of this model provides a good description of the empirical probability density and it supplies the following interpretation: an admitted patient has α_i probability to stay μ_i days in average.
Considering parameters reported in Table 5.5, the following diagram describes the

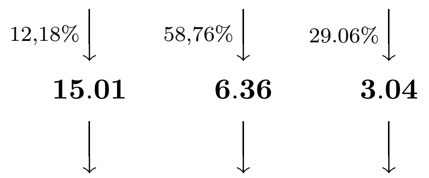
	PHGamma	Coxian	Unicyclic	Hyperexponential
General Surgery	-2.5659	-2.6558	-2.6498	-2.7987
Geriatrics	-2.8484	-2.8794	-2.8794	-3.1812
Clinical Medicine	-2.5785	-2.6156	-2.6129	-2.8645
Endoscopy	-1.8708	-2.0528	-2.0456	-2.3360

Table 5.4: Objective log-likelihood function values of 3-states models

	α_1	α_2	α_3	λ_1	λ_2	λ_3	k	θ
General Surgery	0.1267	0.3247	0.5487	0.9097	1.9450	5.5350	8.3690	0.5614
Geriatrics	0.2037	0.7458	0.0505	0.7945	1.6841	5.3533	7.4938	0.5996
Clinical Medicine	0.1218	0.5876	0.2906	1.1484	2.7095	5.6575	8.3797	0.4861
Endoscopy	0.0695	0.2543	0.6761	1.3972	3.3472	6.8216	12.7696	0.7549

Table 5.5: Estimated parameters of 3-states PHGamma distributions for each department of Campus Bio-Medico.

department of Clinical Medicine:



5.2.5 Forecasting ability

Longer stay patients are important, in fact although the majority of patients is discharged after a short period, beds are mostly occupied by patients that stay for a longer period. Proposed models have same distribution tail behaviour therefore we introduce cumulative measure to underline the differences. In order to provide an operational guideline in bed planning, it is important to estimate the probability that a patient is discharged before a given time.

Inspired by the theory of risk measures [86], [87], and [88], we propose two indices based on the notion of *value-at-risk* and the *expected shortfall* related to the LoS probability distribution. Precisely, given a confidence level $\beta \in (0,1)$, since τ is assumed to have a continuous distribution function, we can consider the following amounts:

$$\mathbb{P}\left(\tau > \mathrm{VLoS}_\beta(\tau)\right) = \beta, \tag{2.38}$$

	Mean	Variance	Skewness	Kurtosis
General Surgery	6.04	27.38	1.91	7.24
Geriatrics	8,86	25.55	1.57	6.50
Clinical Medicine	6.45	18.73	1.99	8.35
Endoscopy	3.80	8.01	2.70	18.83

Table 5.6: Estimated moments of 3-states PHGamma distributions for each department of Campus Bio-Medico.

and

$$\mathrm{ELoS}_{\beta}\left(\tau\right) = \frac{1}{\beta}\int_{1-\beta}^{1} \mathrm{VLoS}_{\mathrm{b}}(\tau)\,\mathrm{db} = \mathbb{E}\left[\tau|\tau > \mathrm{VLoS}_{\beta}(\tau)\right]. \tag{2.39}$$

Thus equation (2.38) implicitly defines $\mathrm{VLoS}_{\beta}(\tau)$ as the smallest time t such that the probability of a demission is greater than β. In other words $\mathrm{VLoS}_{0.05}(\tau)$ is the threshold value of τ such that leaves on its left side the 95% of the probability associated with the probability distribution of τ. Therefore is proposed a procedure such that, starting from a calibration data set, it estimates the model parameters and it provides a forecast of future data. We consider a calibration data set consisting of 2000 patients ordered by time and the $\mathrm{VLoS}_{\beta}(\tau)$ is evaluated for each model, for $\beta = 0.05$ and $\beta = 0.03$. Then a new sample is taken where the oldest 20 elements are replaced by a new data set. On this updated sample the $\mathrm{VLoS}_{\beta}(\tau)$ is evaluated using the empirical distribution. These values are compared with the values previously estimated using the model. The updated data set becomes the calibration data set and the procedure is repeated twenty times using as the starting point for the calibration the previously estimated parameters (for the first step of this procedure a Halton quasi random sequence of $10 \times n$ initial points is used). In Figure 5.4 is reported the result of VLoS estimation for the General Surgery department. This figure shows that, using PHGamma distribution, a better estimation of VLoS is achieved; in fact this approach allows to describe better than other models the tail of the density. The difference between the real values and those estimated by the PHGamma is less than 1 day, hence the estimation of VLoS is validated.

The same procedure, used for the VLoS, is also adopted to evaluate the ELoS. In Figure 5.5 the estimation of ELoS for the Clinical Medicine department is showed. Using the PHGamma distribution is also provided a good estimation of this risk measure since the different between the values is negligible. We remark that the empirical VLoS

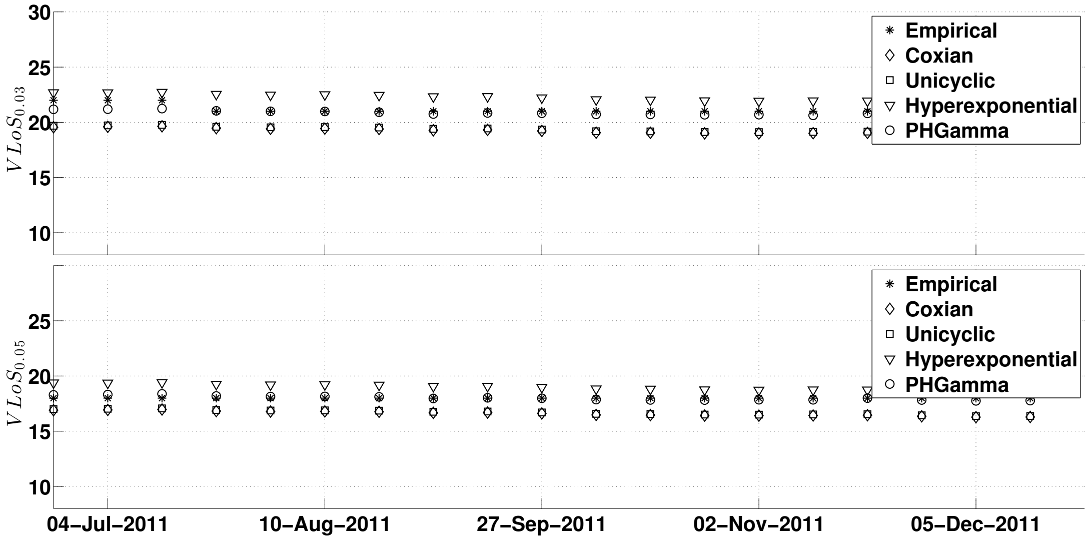


Figure 5.4: VLoS analysis of Campus Bio-Medico General Surgery department - Comparison between 3-states phase-type distributions

and ELoS represented in Figure 5.4 and 5.5 are computed on data sets where newest hospitalizations are not used for the calibration of the models and for the calculation of the corresponding theoretical VLoS (2.38) and ELoS (2.39) associated. The stationary behaviour over a time window of the theoretical VLoS and ELoS, based on the PHGamma model, shows the ability of such model to forecast the higher LoS which yields a relevant consumption of resources for the hospital. Conversely, other models exhibit a poor performance underestimating or overestimating this crucial amount.

5.3 Tor Vergata case of study

In this section is demonstrated how standard density function can not describe the Tor Vergata cardioligy department LoS distribution. In order to capture all the statistical features observed in real situations the use of PHGamma is proposed [29].
Also an analysis of the discharge rate (DR) and the admission rate (AR). Those proxies provide information on the efficiency of departments and give operational guidelines for medical staff is performed. In fact the distribution of discharge rate supplies the probability that a given number of patients is discharged in one day; likewise the admission rate provides the probability that a given number of patients is admitted in one day.

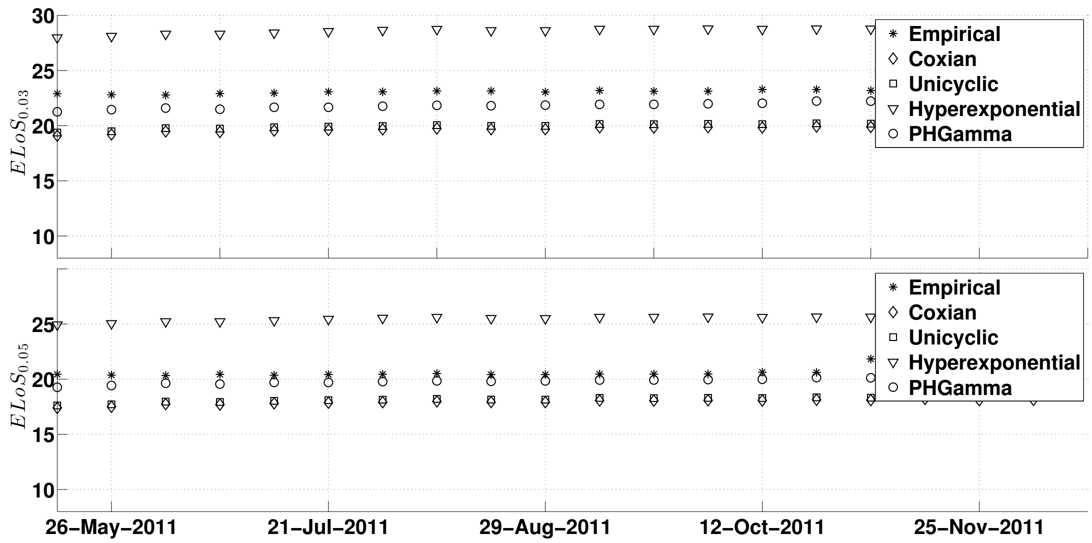


Figure 5.5: ELoS analysis of Campus Bio-Medico Clinical Medicine department - Comparison between 3-states phase-type distributions

5.3.1 Data

Tor Vergata Hospital is a big size structure within the context of Italian healthcare system. Moreover this hospital offers a high quality service compared to the national health system. This work focuses on the flow of patients within the department of Cardiology. All the data are related to a period of 5 years going from 1^{st} January 2008 to 16^{th} May 2013. There have been considered 11510 admissions.

The basic descriptive statistics of the patients LoS are reported in Table 5.7. The values

	Mean	Variance	Skewness	Kurtosis	N
LoS	3.33	4.86	3.46	24.40	11510

Table 5.7: LoS descriptive statistics of Tor Vergata Cardiology department

show that the LoS density function is asymmetric, it can only take positive values, and it exhibits a leptokurtic form (execute acute peak around the mean and have fatter tails). The long gradual tail to the right of the distribution is given by a very small number of patients that stay in hospital for a considerable amount of time.

We remark the importance of longer stay patients, in fact although the majority of patients is discharged after a short period, beds are mostly occupied by patients that stay for a longer period.

The discharge rate is the probability that in one day are discharged n patients; it is estimated from real data as the number of day in which are discharge n patients divided by the total number of considered days. Likewise the admission rate is defined as the number of days in which are admitted n patients divided by the total number of considered days. In Table 5.8 are reported the main characteristics of the admission rate and the discharge rate.

	Mean	**Variance**	**Skewness**	**Kurtosis**	**Max**
Admission	5.83	6.62	0.06	2.59	14
Discharge	5.83	6.96	0.17	2.93	16

Table 5.8: Tor Vergata Cardiology department statistics of admission and discharge rate.

5.3.2 AR/DR model

To evaluate the discharge and the admission distributions it is used an approach based on *single factor model* [102]. In general this model is used in financial mathematics to estimate credit risk of a portfolio asset [103]. The basic idea behind this model consists into creating an analogy between the default of a financial institution and discharge of a patient. The aim is to come up with a simple and closed-form formula for the distribution function of the discharge rate. Deriving a closed form solution requires making a set of simplifying assumptions. We will progressively introduce these assumptions and their implications for the model.

To derive the DR and the AR, knowing the individual probabilities of the patients is not enough; we also need to know their correlation structure. We assume that each patient status is describe by a random variable. Therefore we introduce the real value random variable $X_1, ..., X_N$, where N is the number of inpatients defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which drive the patients discharge/admission probabilities. We also assume that the correlation coefficient of each pair of random variables X_n and X_m is $\rho_{n,m}$. The correlation coefficient $\rho_{n,m}$ between each pair of random variables X_n and X_m is the same for any two patients:

$$corr(X_n, X_m) = \rho_{n,m} = \rho. \tag{3.40}$$

There exists a source of uncertainty affecting all patients. Moreover, the random variables $X_1, ..., X_N$ are described by (3.41)

$$X_n = \sqrt{\rho}Y + \sqrt{1-\rho}\varepsilon_n, \tag{3.41}$$

for all $n = 1, ..., N$ where $Y, \varepsilon_1, ..., \varepsilon_N$ are independent random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. We can interpret (3.41) as follows: each random variable X_n represents the disease state of patient, whose realization determines if the patient n is discharged. In fact, if the disease state is less than a threshold value K the patient is discharged. X_n can be expressed as the sum of two factors: one common or systematic factor Y and an idiosyncratic factor ε_n that are i.i.d. and defined on the same probability space. Y affects all patients in the same way and has density function f with mean $\mathbb{E}[Y] = 0$ and variance $\mathbb{V}[Y] = \sigma^2$, the idiosyncratic factors are independent across patients and have a common density function g such $\mathbb{E}[\varepsilon_n] = 0$ and $\mathbb{V}[\varepsilon_n] = \nu^2$. The covariance between X_n and X_m is

$$\begin{aligned} cov(X_n, X_m) &= \mathbb{E}\Big[X_n X_m\Big] \\ &= \mathbb{E}\Big[(\sqrt{\rho}Y + \sqrt{1-\rho}\varepsilon_n)(\sqrt{\rho}Y + \sqrt{1-\rho}\varepsilon_m)\Big] \\ &= \rho\sigma^2 + (1-\rho)\nu^2\delta_{n,m} \end{aligned} \tag{3.42}$$

where $\delta_{n,m}$ is the Kronecker delta.

Conditioning on the factor Y, the discharge probability of each patient, denoted by $p(Y)$, is easily computable:

$$\begin{aligned} p(Y) &= \mathbb{P}[X_n < K|Y] \\ &= \mathbb{P}\Big[\sqrt{\rho}Y + \sqrt{1-\rho}\varepsilon_n < K|Y\Big] \\ &= \mathbb{P}\Big[\varepsilon_n < \frac{K-\sqrt{\rho}Y}{\sqrt{1-\rho}}|Y\Big] \\ &= G\Big(\frac{K-\sqrt{\rho}Y}{\sqrt{1-\rho}}\Big) \end{aligned} \tag{3.43}$$

where G is the cumulative density function related to g. Hence the random variables $X_1, ..., X_N$ are conditionally independent given the random systematic factor Y.

Consider, for each patient n, the random variable L_n takes value 0 if the patient n has not discharged from the department under consideration and is 1 otherwise. Then define $L(N)$ as the number of discharged patients:

$$L(N) = \sum_{n=1}^{N} \mathbb{1}_{X_n < K_n}. \tag{3.44}$$

In general N cannot be considered as deterministic. Actually, the number of patient hospitalized in each department is modelled as a discrete random variable N_p with probability mass function h. We also assume that N_p is defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ ans is independent of ε_n and Y.
In our empirical study, the maximum number of patients that can be received in the department N_{max} is estimated from historical data and is reported in Table 5.8. Therefore the density function of $L(N_p)$ is formulated in light of the fact that $\mathbb{P}(L = m|N_p = n, Y = y)$ is binomial distributed with parameters n and $p(Y)$, This the probability that $L = m$ is given by

$$\begin{aligned}
\mathbb{P}(L = m) &= \sum_{n=m}^{N_{max}} h(n)\mathbb{P}(L = m|N_p = n) \\
&= \sum_{n=m}^{N_{max}} h(n) \int_{-\infty}^{+\infty} \mathbb{P}(L = m|N_p = n, Y = y)f(y)dy \qquad (3.45) \\
&= \sum_{n=m}^{N_{max}} h(n) \int_{-\infty}^{+\infty} \binom{n}{m} p(y)^n[1 - p(y)]^{n-m} f(y)dy
\end{aligned}$$

In a future work in preparation we will consider the dependence between the efficiency of the department and the number of patients.

5.3.3 Optimization method

According to [95], [96], and [97] maximum log-likelihood estimation is used in order to fit the parameters of LoS distributions. The patients number N is reported in Table 5.7 The log-likelihood objective function is given by

$$LogL(\mathbf{\Theta}) = \arg\max_{\Theta \in \mathcal{D}} \sum_{i=1}^{N} \log(f(t_i; \mathbf{\Theta})), \qquad (3.46)$$

where $f(\cdot; \mathbf{\Theta})$ is the density function of the considered model depending on the set of parameters described by a vector $\mathbf{\Theta}$.
The optimization routine is implemented in the MATLAB involvement using existing algorithm based on the interior-point method. Parameters calibration starts from a Halton quasi-random sequence of $10 \times n$ initial points [98], [99]. In order to choose the optimal PHGamma, a sequential procedure is adopted whereby increasing numbers of phases n are tried starting with $n = 1$ until little improvements in the fit of the data

can be obtained by adding a new phase. The number of the phases that allows the best compromise between model complexity and goodness of the fit is chosen. In Table 5.9 are reported the values of the objective function related to the PHGamma model. The

	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$
$LogL(\Theta)$	-1.9038	-1.8026	-1.7698	-1.7605	-1.6528	-1.5607	-1.5210

Table 5.9: Log-likelihood function values obtaind in LoS density estimation using a n-states PHGamma model on Tor Vergata Cardiology department.

analysis of the data shows that the use of a model with more than 3 states a negligible in the performance. Hence we fix the number of the phases to 3.

5.3.4 Estimation Results

5.3.4.1 LoS model

The PHGamma distribution provides a very versatile tool to describe a wide class of density functions. Using a few parameters is possible to describe distributions that exhibit different characteristics. The performance of the model are compared with different standard density. In Figure 5.6 Normal, Lognormal, Weibull, Gamma and

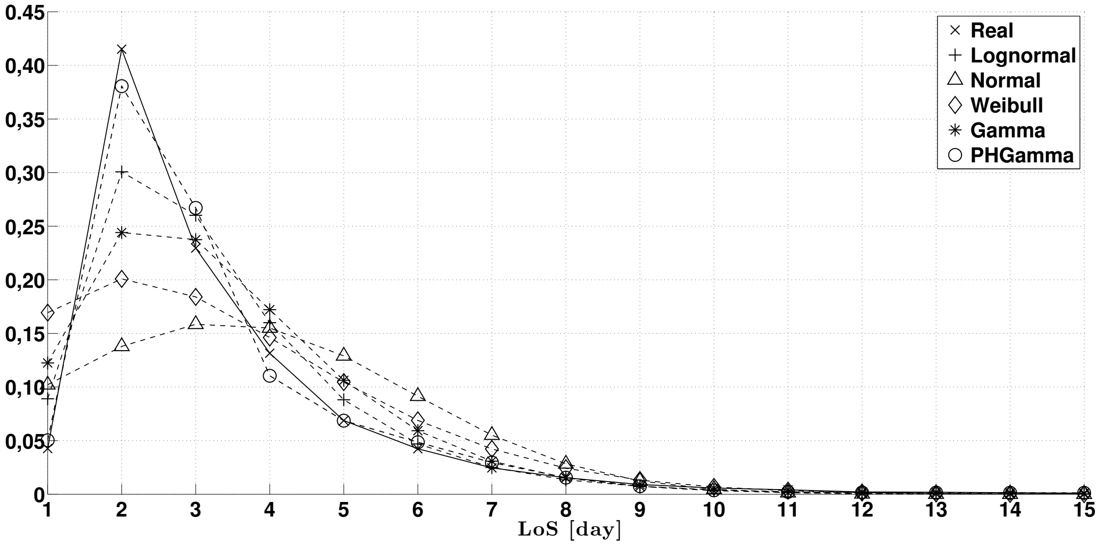


Figure 5.6: LoS estimated density functions of Tor Vergata Cardiology department - Comparison between 3-states PHGamma and standard distributions.

3-states PHGamma is estimated. The results show that the proposed model provides a better fit of the data. In Table 5.10 are reported the parameters estimated. In order

α_1	α_2	α_3	λ_1	λ_2	λ_3	k	θ
0.0161	0.2538	0.7302	0.9794	2.8180	6.1106	9.5710	0.6430

Table 5.10: 3-states PHGamma estimated parameters of Tor Vergata Cardiology department LoS distribution.

to take measure of the goodness of the fit in Table 5.11 are reported the log-likelihood function for each model. 3-states PHGamma model performs the biggest log-likelihood value. To evaluate the performance of those models, we also use four discrepancy

	Lognormal	**Normal**	**Weibull**	**Gamma**	**PHGamma**
$LogL(\Theta)$	-1.8082	-2.3131	-2.0131	-1.9038	-1.7698

Table 5.11: Tor Vergata Cardiology department log-likelihood function value of LoS estimation for each model.

measures introduced in 2.37. These measures are shown in Table 5.12. In particular we observe that the PHGamma distribution achieves the smallest errors for all these measures.

	Lognormal	**Normal**	**Weibull**	**Gamma**	**PHGamma**
AAE	0.0050	0.1286	0.0385	0.0077	0.0039
APE	1.4761	10.3361	5.6001	2.4620	0.4702
ARPE	0.0224	0.1566	0.0849	0.0373	0.0071
RMSE	0.0120	0.1325	0.0441	0.0196	0.0051

Table 5.12: Summary of four discrepancy measures of Tor Vergata Cardiology department LoS distribution.

Forecast ability In order to provide an operational guideline in bed planning, it is important to estimate the probability that a patient is discharged before a given time. To validate the performance of the model and its stability over time is used the following procedure: we consider a calibration data set consisting of 1000 patients ordered by time to estimate the parameters of the model. Then a new sample is taken where the oldest 100 elements are replaced by new data. On this updated sample the empirical distribution is evaluated and is compared with the density previously estimated using the PHGamma model. The updated data set becomes the calibration data set and the procedure is repeated hundred times using as the starting point for the calibration the previously estimated parameters (for the first step of this procedure a Halton quasi random sequence of $10 \times n$ initial points is used). There have been considered more then 11000 different admissions from April 3^{rd} 2008 to April 30^{th} 2013. To evaluate the stability of the model it is considered the time analysis of the parameters. In Figure 5.7 is shown the time trend of estimated parameters that exhibit a regular pattern with a limited variation.

In order to evaluate the model performances in each scenario has been considered the following error estimator

$$\varepsilon = \frac{\sum_k t_k[f_r(t_k) - f_e(t_k)]}{\sum_k t_k f_r(t_k)} \tag{3.47}$$

where f_r is the real distribution and f_e is the estimated distribution. In the first plot of Figure 5.8 is reported the time trend of the error. Each scenario has 1000 patients and covers an average period of 17 days with a use of 3500 bed days in average. The error value, that is bounded in the range $\pm 3\%$, shows that the model has an average bed days error estimation lower then 15. In the second plot is reported the absolute error for the bed days and patients. The bed days error formula is

$$|\varepsilon_b| = \frac{\sum_k t_k|f_r(t_k) - f_e(t_k)|}{\sum_k t_k f_r(t_k)} \tag{3.48}$$

and patient LoS formula is

$$|\varepsilon_l| = \text{APE} = \frac{\sum_k |f_r(t_k) - f_e(t_k)|}{\sum_k f_r(t_k)} \tag{3.49}$$

These indicators evaluate the distance between the distributions and show underestimation and overestimation of bed days consumption and patient LoS in each scenario. In particular, the plot shows that an average error of 17% in the bed days evaluation consumption is made, and an average error of 10% for the patients LoS realized. The absolute error allows to highlight errors of overestimation and underestimation in contrast of using the error based on the simple difference that takes into account the cumulative consumption of resources.

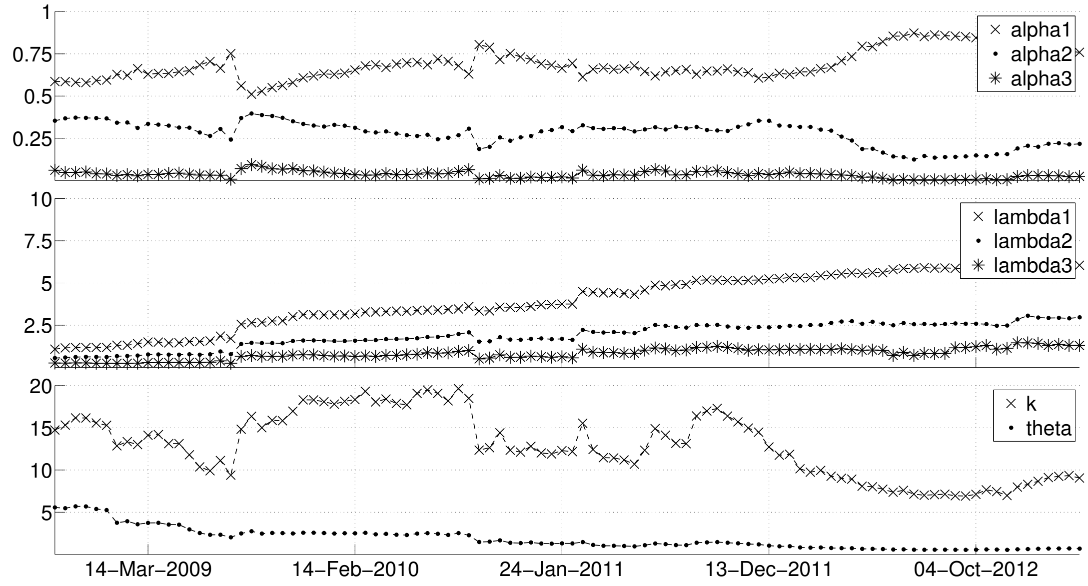


Figure 5.7: Parameters time trends of Tor Vergata Cardiology department - Forecast ability analysis, y-axes represent the estimated value parameters.

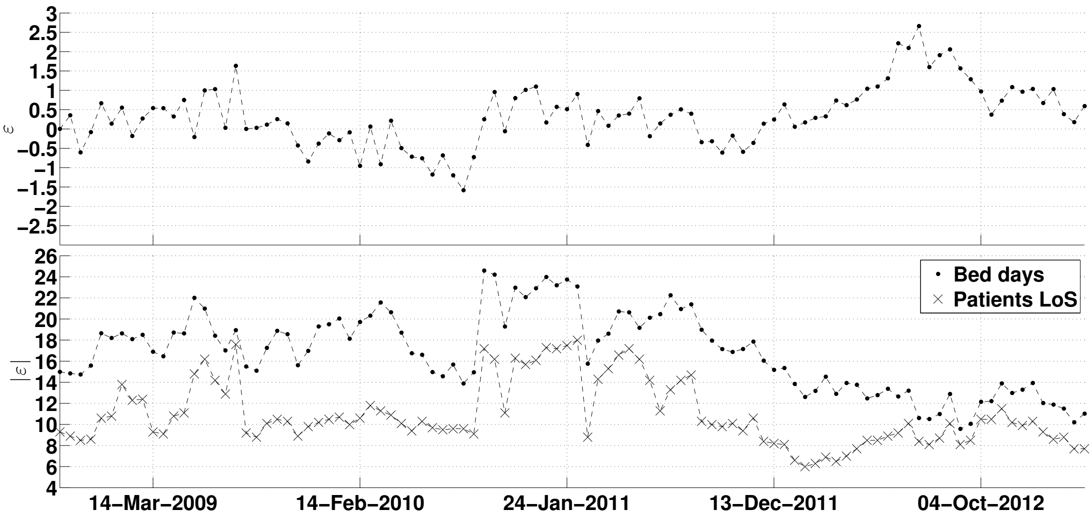


Figure 5.8: Error time trends of Tor Vergata Cardiology department - Forecast ability analysis, in first plot is reporter the estimated density fitting error. In second plot are reported the time trend analysis of error metrics reported in 3.48 (point-mark) and 3.49 (x-mark).

5.3.4.2 AR/DR model

An important proxies to evaluate the efficiency of a department are the admission rate and the discharge rate. Furthermore using the single factor model is provided a phenomenological interpretation. In our empirical study we shall consider the following models for the distribution densities involved in 3.45:

$$\begin{cases} f(t) = & \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2}} \\ g(t) = & \frac{1}{\nu\sqrt{2\pi}} e^{-\frac{t^2}{2\nu^2}} \\ h(t) = & \frac{z(t|\mu,\xi)}{\xi} e^{-z(t|\mu,\xi)}, \qquad z(t|\mu,\xi\,) = e^{\frac{t-\mu}{\xi}} \end{cases} \tag{3.50}$$

Here the density h is selected according with a separated estimation procedure where we observe that the number of patients in the department is well described by the extreme value distribution, see Figure 5.9.

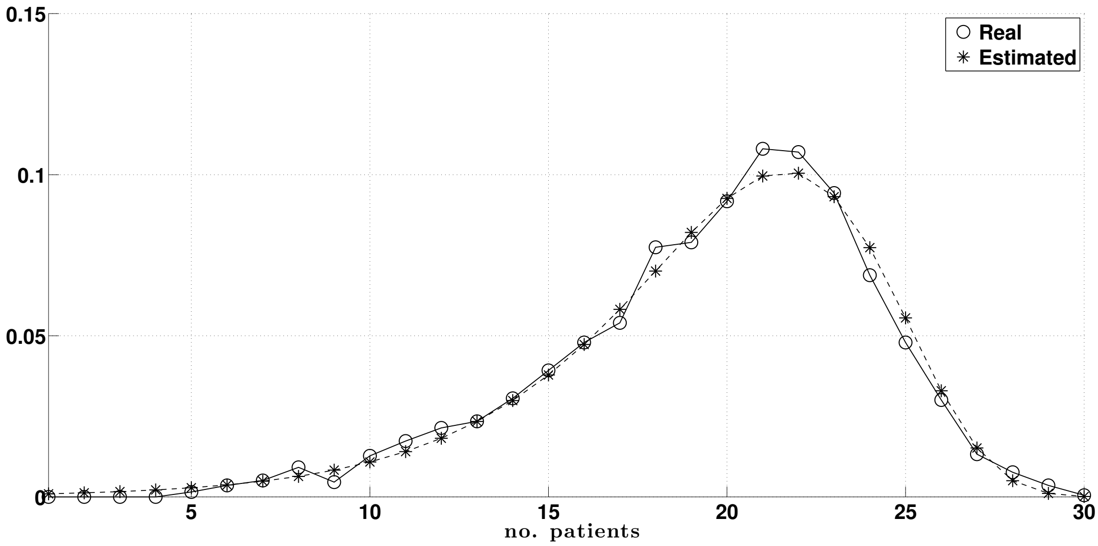


Figure 5.9: Number of patients estimated density function of Tor Vergata Cardiology department - Inpatients data analysis

The estimated parameters of the model are reported in Table 5.13 In Figures 5.10 and 5.11 are reported the probability functions estimated. The proposed model provide a good description of the empirical distribution.

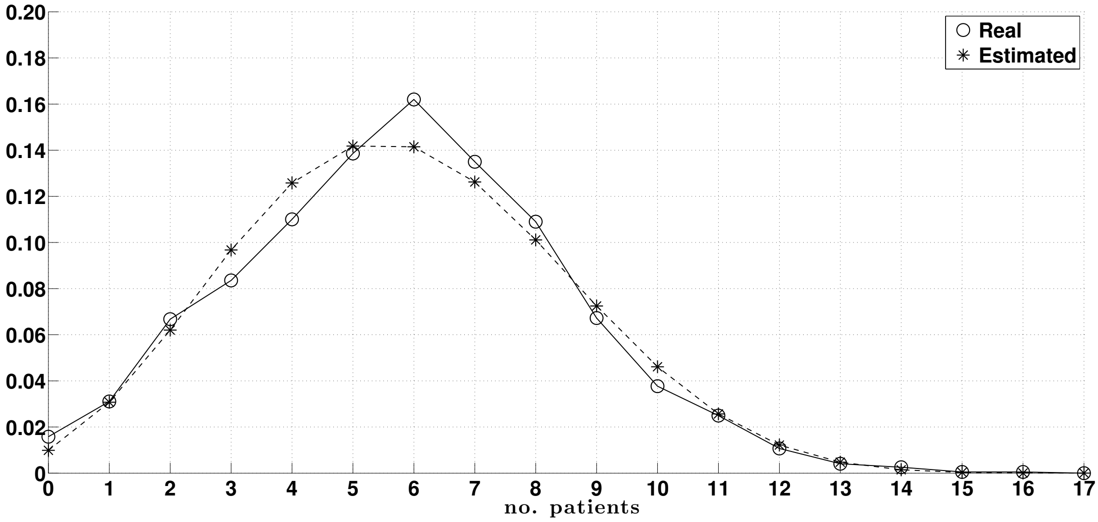


Figure 5.10: Discharge rate estimated density function of Tor Vergata Cardiology department - Inpatients data analysis

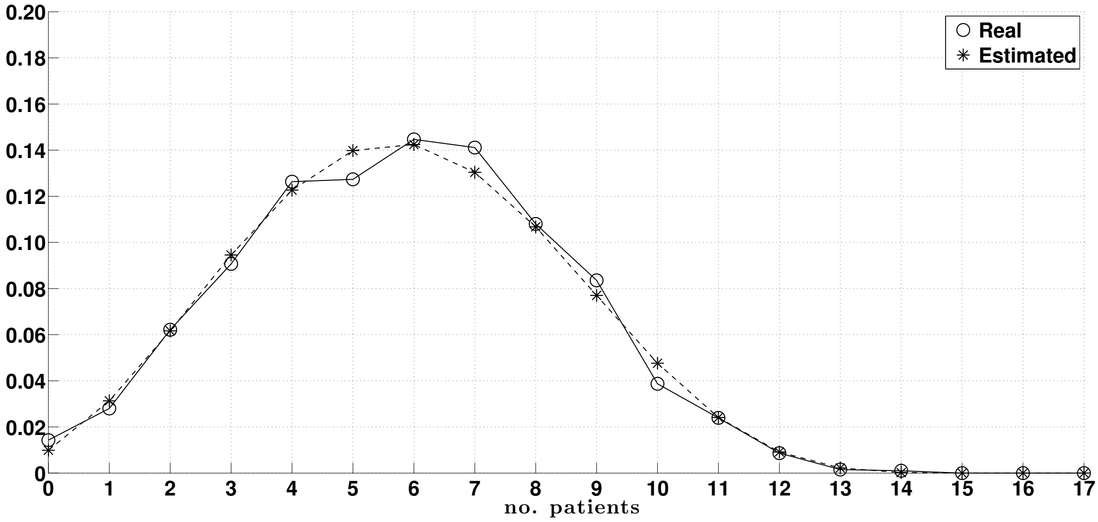


Figure 5.11: Admission rate estimated density function of Tor Vergata Cardiology department - Inpatients data analysis

	$\boldsymbol{K}$	$\boldsymbol{\rho}$	$\boldsymbol{\nu}$	$\boldsymbol{\mu}$	$\boldsymbol{\xi}$	$\boldsymbol{\sigma}$
AR	1.0813	0.8233	22.3929	11.6209	1.3980	4.5974
DR	−0.0386	0.9861	29.7045	12.9137	1.9190	1.1737

Table 5.13: Estimated parameters for admission rate and discharge rate of Tor Vergata Cardiology department.

5.4 Summary

In this chapter we investigated a new type of PH distribution, we gave a brief overview of some properties and we presented a particular application to hospital LoS data. The generalization proposed uses a few number of parameters and it supplies a phenomenological interpretation of data. We introduced measures, used in financial risk modelling, in order to provide an operational guideline for planning and managing hospital beds. We also introduced a new approach (single factor model) in order to describe the admission rate and the discharge rate of the department.
The following conclusions can be drawn:

1. PHGamma distribution provides a good fit of the real data in fact it performs the smallest discrepancy measures (AAE, APE, ARPE, and RMSE).

2. The moment generating function of PHGamma is known in closed-form and the log-likelihood estimation supplies a good estimation of the first four observed moments.

3. The forecast ability is validated using the introduced risk measures (VLoS, ELoS). Differently from other models proposed in the dedicated literature [90], [100], and [101], the PHGamma model estimation error is less than one day.

4. The single factor model provides a good description of the admission rate and the discharge rate and supplies an adequate representation of the hospital operations department.

5. The stability of the estimated parameters is shown using a sequential procedure based on the analysis of some error indicators

6

Hospital resources optimization

This Chapter introduces some quality indicators as a tool to evaluate health systems performance and quality. A dynamic stochastic optimization model is then proposed to optimize hospital bed occupancy. Three different models to describe patient discharge probabilities are also proposed and then used to evaluate the optimal policies.

The costs associated with the healthcare system have risen dramatically in recent years, and the increased public scrutiny to which the system has been subjected has been accompanied by increased attention from operations researchers and systems engineers [104]. Research in this area has touched on nearly all aspects of the healthcare system, with particular emphasis being given to problems in hospital operations management [13], [32].

In recent decades different numeric indicators and methods have been used to qualify structured index that express health service components quality performance. The first step of this work is the definition of a model of the system under examination in order to define its measurable constituent entities (dimensions). Among the common measured clinical parameters, a robust metrics, characterizing the constituent entities and the best opportunity tools for the characterization of the results, have been identified. The last step is multi-modal analysis of the results. Using this approach is provided an application in the evolution of the cardiological intensive care unit (ICU) of San Camillo hospital. Particularly, the attention is focused on the evolving of the medical team.

Moreover we provide also a methodological approach to optimize the hospital resource allocation based on stochastic dynamic programming. Results shows that the optimal

policy proposed performs better than an empirical policy. Empirical discharge frequency models are also provided.
In this chapter are introduced some methods to support the decision process for the department **resource capacity planning** and **financial planning**. The approaches mainly focus on the **tactical level** and **operational level**; using quality indicators to evaluate the health service performance provide decision process insight for tactical level while the use of a stochastic dynamic approach contributes to online and the offline operational planning procedures. Moreover the analysis of some measured clinical parameters produce indications for the **medical planning** process.

6.1 Evaluation of indices for the measurement of quality in health systems

In recent decades the term *quality* has been introduced in to the common language as an essential part of clinical management. The intents of quality approach in healthcare are mainly focused towards the identification of needs, and the evaluation of the service through measurements and indices suitable to assess the performance of the system [105], [106]. As in all sectors the use of suitable quality metrics is crucial. Then must be defined tools to evaluate and combine indices to have structured methodologies that expresses a global evaluation of the quality performance [107].
Quality indicators can be classified in several different ways, in addition they can be compared and analyzed using different approaches. Some indicators can take contributions from assessed classifications, or more indicators can be used simultaneously according to the rules of data stratification [108]. Indicators can describe events and rates of service's utilization by the population [109]. Higher rates of several indicators may be interpreted as efficiency in functioning, whereas an indicator such as the death rate, which is usually referred to specific clinical situations, may serve to assess the risk of death after medical intervention. A widespread approach relies on the definition and use of benchmarks, which allows a comparison between indicators related to homogeneous systems and processes [110]. Indicators can be divided into sub-groups to analyse the system, even through simplified models offer the possibility to identify the areas of the health system in a conceptual framework that link these areas to correlating indicators for an overall analysis. The goal is a first characterization of the individual

components and the successive study of inter-relationship between them to improve the outcome in the individual patient. For this the model proposed by [111] integrated by [112] is widely accepted and proposed in many studies [105] for the characterization and the measure of the quality in healthcare.
An accepted fact in managing metrics for quality is to put in evidence that, in any process, there is a background noise due to cumulative effects of many small, essentially unavoidable causes i.e stable system changes. Thus the estimation of useful indicators requires a phase of analysis in which statistical techniques can be deployed. Between more common approaches to evaluate systems dynamic of Statistical Quality Control (SQC) it is useful to detect the occurrence of assignable causes of process shift. Therefore causes of improvement can be assigned and high-lighted with the final aim of monitoring improvement causes/actions. Key concepts of SQC include the idea of a rational sub process analysis and the use of tools such as histograms, run charts, control chart signals, etc. The tools of SQC helps to create an environment in which the stakeholders of each area seek continuous improvement for their activities. Elsewhere SQC techniques are applied to predicting activities [113] and to understanding whether there are changes in the care and in the management of altered, patient outcome [114] in this clinical field. An example of the application of statistical methods to health service to quantify new paradigms and procedures in health service is provided [115]. Some practical issues in the implementation of SQC applied to evaluate the ICU of a cardiac surgery department, are shown. The final aim is to provide a methodology in order to study how change in healthcare alters patient outcome [114], [116].

6.1.1 Quality measurements in health systems

As stated quality measurements in healthcare need the identification of indicators that allow an objective assessment of the service under examination. This is particularly relevant for new developments in domains, including those concerning diagnostic/treatment procedures, inter-operability, and the use of new medical devices [117]. Besides quality metrics it is necessary to synthesize data in a conceptual model contemplated for global quality assessment and health system monitoring [112]. The conceptual system, here considered to measure the quality of the health system, is derived from the measurement approach proposed by Donabedian and widely accepted in the literature [118]. The approach by Donabedian combines several issues, related to the organizational structure, processes, and outcomes, in a comprehensive clinical model. In particular, the health system is split into four conceptual components: mission, structural capacity, processes and outcomes, all interacting with the macro context [106], [110], [112], [119]. From this division in different areas the evaluation of reference indicators and/or parameters, is a technique widely used in healthcare. As a mater of fact the conceptual system here adopted simplifies this approach by considering the simplest model derived from the classical process approach and performance improvement of ISO 9000 standard and highlights as indicators of quality a set of parameters related to the structure, process and outcomes [105], [108]. The structure (organizational data) is characterized by

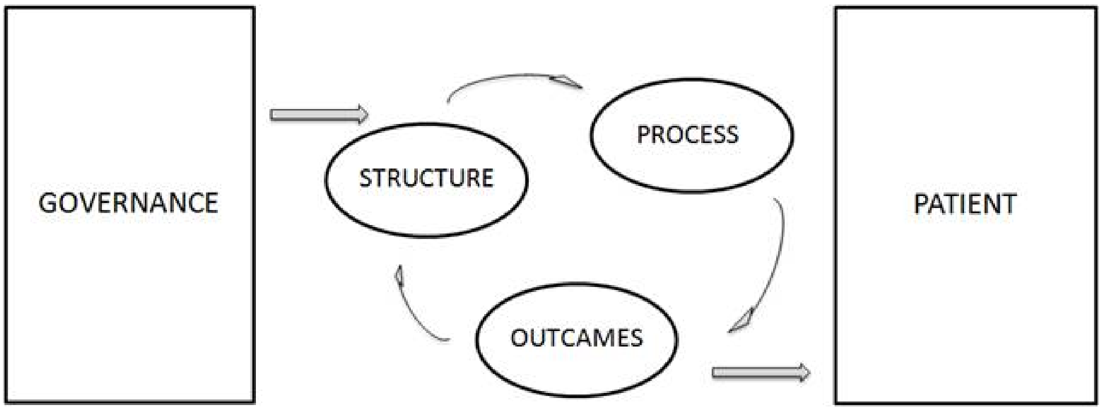


Figure 6.1: Schematic of a model of ICU department - Quality indicators a set of parameters related to the structure, process, and outcomes are highlighted [105]

the features of the hospital, operators' skills, the characteristics of equipment and human resources, and so on. The process consists in the encounter between healthcare physicians and patients (professional quality). It starts from reservation and admission, through anamnesis, screening planning and execution, diagnosis, therapy and administration of drugs. Elsewhere [108] the indicators characterizing the process can be widely recognized clinical standards and are extracted from clinical data. The clinical outcomes (professional quality) are according to the type of clinical service. They give information on the state of health of the patient during and after treatment by the health service including the results such as: mortality, morbidity, perception and satisfaction of the patients at discharge, quality of life after treatment [120]. The evaluation can be performed on each dimension (structure, process and outcome), using different methods of measurement based on as many different indicators .

Controlling and improving quality has become an important business strategy for healthcare providers and control chart is one of the primary techniques to be adopted. This charts plots the average of measurements of a quality characteristic in samples taken from the process versus time. The charts highlight where and when the process fails in order to recognize unusual sources of variability present. When unusual sources of variability are present, sample average will plot outside the control limits. This is a signal that some investigation of the process should be made and corrective action to remove these variability must be taken [115].
Another hospital quality evaluation is the study of internal dynamics of a hospital which represent a complex non-linear structure. The planning and managing of these resources requires a good understanding of the hospital system, therefore an advanced data analysis is proposed. Hospital length of stay of in-patients has been employed as a proxy for measuring the consumption of hospital resources. Simple averages are widely used to describe LoS but parameters can be incorporated in models of resource utilization and patient flow. Thus a model that forecasts the LoS starting from the EuroSCORE (ES), an index of health status of the admitted patients, is proposed. Furthermore in order to explore the complex relationship between bed occupancy and ES values and optimize resource utilization, a model to describe the discharge frequency is used.

6.1.2 Data analysis

Quality measurement in health care is the process of using data to evaluate the performance of health plans and health care providers against recognized quality standards. Quality measures can take many forms, and these measures evaluate care across the full range of health care settings, from doctors offices to imaging facilities to hospital systems. We discuss each of these measures below. However, it is important to note that no single type of measure can give a complete picture of the quality of care that is provided and received. Rather, each type of measure addresses a key component of care. Hundreds of different quality measures are used in health care. These measures generally fall into three broad categories:

1. **Structure measures** evaluate the infrastructure of health care settings, such as hospitals or doctor offices, and whether those health care settings are able

to deliver care. These measures include staffing of facilities and the capabilities of these staff, the policy environment in which care is delivered, and the availability of resources within an institution. Structure measures are often used by insurance companies and regulators to determine whether a provider has certain capacities needed to deliver high quality care, such as whether a hospital has a system in place to order prescription drugs electronically. These measures are also commonly used in the certification or accreditation of health plans and providers. Two key reasons for using structure measures are that characteristics of health care settings can significantly affect the quality of care, and care settings that meet certain standards have an advantage when it comes to providing high-quality care.

Although structure measures provide essential information about a providers capacity, it is important to note the limitations of these measures. In particular, structure measures provide just one piece of the full picture of care. For example, the fact that a hospital has the ability to perform certain functions does not capture whether or not these functions actually occur, nor does it capture whether those functions improve patient health.

2. **Process measures** are used to determine the extent to which providers consistently give patients specific services that are consistent with recommended guidelines for care. These measures are generally linked to procedures or treatments that are known to improve health status or prevent future complications or health conditions.

   Process measures are useful in that they give providers clear, actionable feedback and a straightforward way to improve their performance. However, overreliance on process measures to track performance and administer provider incentives can be problematic, for several reasons.

   - Process measures are not available for many key areas of care, such as whether the care provided was appropriate, or whether a provider coordinated treatment for patients with physical and mental illnesses, for example.
   - Process measures that do exist tend to focus on preventive care and the management of chronic conditions, which may distract from other important

quality areas that are more difficult to measure. Areas where measuring quality is harder include teamwork and organizational culture.

- Process measures may also not capture the true quality of the care provided. For example, a measure that looks at what percentage of patients who smoke received smoking cessation advice will yield the same results whether the advice provided was a brief admonition to quit or a conversation with the patient about barriers he or she faces when trying to quit and the availability of smoking cessation supports.

3. **Outcome measures** evaluate patients health as a result of the care they have received. More specifically, these measures look at the effects, either intended or unintended, that care has had on patients health, health status, and function. They also assess whether or not the goals of care have been accomplished. Outcome measures are where the rubber meets the road: Patients are interested in surviving illness and improving their health, not the clinical processes that support these outcomes.
Outcome measures frequently include traditional measures of survival (mortality), incidence of disease (morbidity), and health-related quality of life issues. And while these measures often incorporate patient-reported information on how satisfied patients are with the health care services theyve received, these measures do not assess the full extent of the patient experience. Although outcome measures are important to patients and providers, their usefulness is limited by the fact that developing outcome measures that are truly meaningful can be quite hard. Key challenges to developing meaningful outcome measures include:

   - Measuring outcomes often requires detailed information that is available only in medical records, and this information is difficult and expensive to obtain.
   - Gathering enough data to provide useful information about a particular outcome can also be a challenge.
   - Although social determinants of health (such as access to safe housing, social support, and economic opportunity) can have a profound impact on health outcomes, there is little agreement on whether or not providers can be held accountable for the confounding effects of social determinants.

- Differences in patient population can make certain outcomes more difficult to achieve. For example, ensuring that a certain percentage of a provider diabetic patients have controlled blood sugar levels may be more difficult for a provider with a patient population that is sicker or that has multiple chronic conditions.

Following we apply this framework to the data of the unit of Anesthesiology and Reanimation intensive care unit of *San Camillo-Forlanini* Hospital in Rome which is part of the Department of Cardioscience. The team of the Unit was formed in 1999, and even started to systematically collect some important data about patients admitted. The data analyzed refer only to the patients coming from the Cardiac Surgery Unit. In particular, the database consists of 11,770 patients (records) collected from 1999 to 2012, and the corresponding entries were: Name, Intervention Date, Age, Pathology, Sex, Urgency, Intervention, Ejection Fraction, Pulmonary Hypertension, Diabetes, Obesity, Chronic Obstructive Bronchopathy, Chronic Kidney Failure, Liver Failure, Neuropathy, Reintervention, Extracorporeal Circulation Time, Orotracheal Intubation Time, Length of Stay in Intensive Care Unit, ES, Post-Operative Complications, Exitus (dead patient).

Structure The Unit counts up to 12 beds, and the team is made up of one head physician, 19 staff physician, one head nurse, and 36 staff nurses. The activity of the Unit is aimed at supporting the whole Department (Cardiac Surgery Unit and the Vascular Surgical Unit). In addition to the well-established transplant activity, the Cardiac Surgery Unit ordinarily performs coronary artery bypass graft surgery (CABG), beating, replacements and repair of the aortic and mitral heart valve (Valv), surgical treatment of heart failure (AO), reconstructive surgery of the aortic root and of ascending aorta, and surgery interesting aortic arch. While, Vascular Surgical Unit ordinarily performs thoracic and abdominal surgical interventions for peripheral arterial occlusive disease, for carotid artery disease, for the lower limbs varicose veins disease, and also performs video-laparoscopic procedures, angioplasty and carotid artery stents, angioplasty and peripheral stent, and aortic endoprosthesis. As stated, the data refer only to cardiac surgery.

Process The first process indicator considered is the number of interventions, in general and divided into different surgeries. Figure 6.2 shows the trend of the number of interventions in each of the above considered categories. The yearly average number

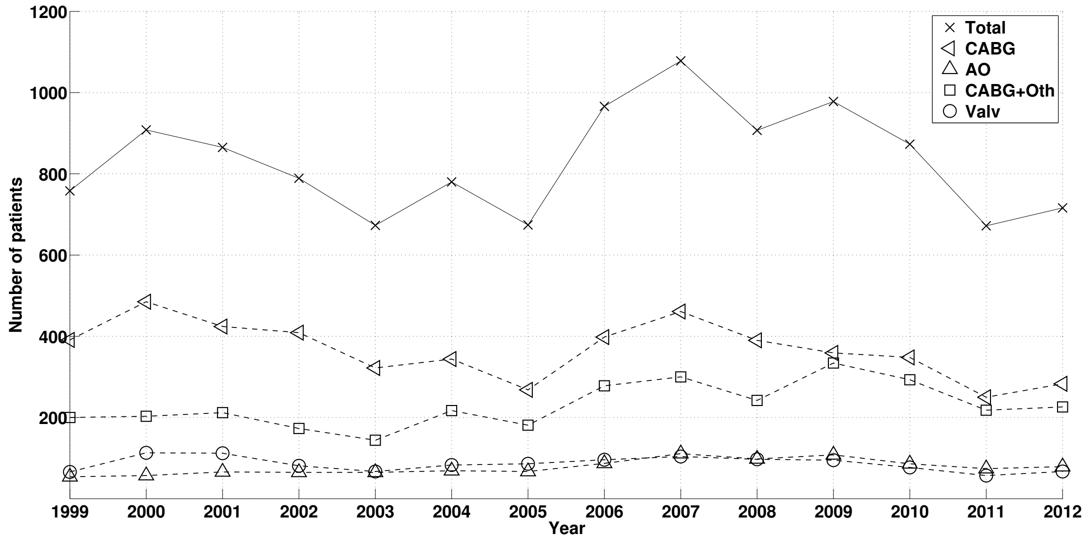


Figure 6.2: Number of patients for different intervention in San Camillo ICU - Process indicator

of interventions is 838. If the percentage of isolated CABG interventions every year is considered, it can be observed that it becomes less and less (passing from the 48% on average of the first seven years, to the 40% of the last seven). This is a logical consequence of the fact that the Cardiac Surgery Unit is trying to treat this kind of patient not with a surgical approach, but with stenting applications, reducing the admissions to the Reanimation Unit.
The second indicator analyzed was the use of mechanical supports for cardiocirculatory function assistance divided by different types. There are various types of device, considers the main one used:

- Mechanical Assistance
  - > VAD (Ventricular Assist Device)
  - > BIVAD (BIVentricular Assist Device, and Artificial Heart)
- LEVITRONIX (Magnetic Centrifugal Pump)

- ECMO (Extra Corporeal Membrane Oxygenation)
- IABP (Intra Aortic Balloon Pump).

Figure 6.3 shows the number of devices used every year, and Figure 6.4 shows the trends for each type of device (note that the variable Mechanical Assistance includes VAD and BIVAD). In line with technological development both Figures show an

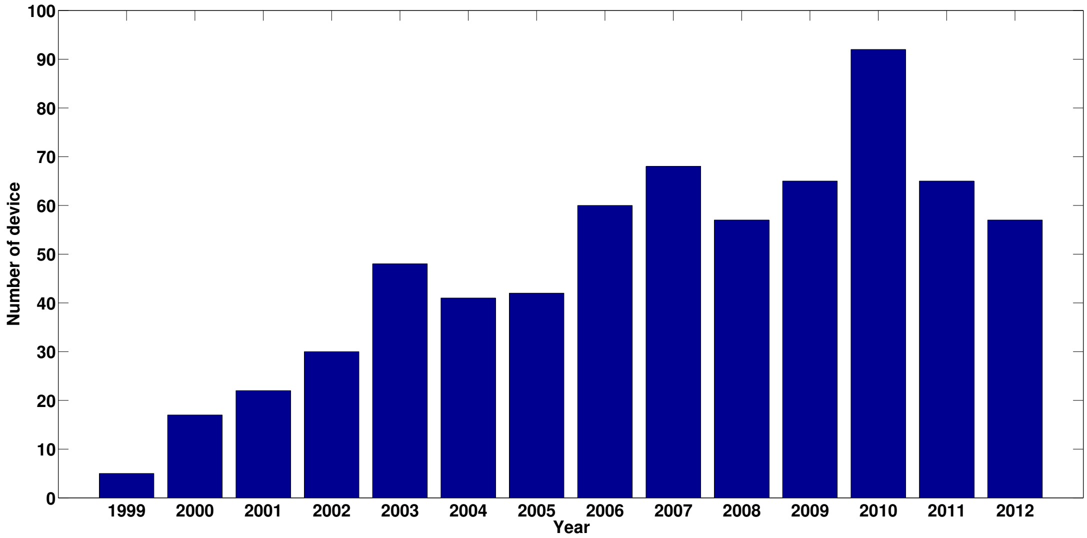


Figure 6.3: Overall number of device used each year in San Camillo ICU - Process indicator

increase in the use of the mechanical support.
The third indicator is the length of stay (LoS): the term Length of Stay indicates the measure (generally days) of the duration of a stay in a department. The importance of LoS as an indicator in the ICU process is twofold [28]: first of all, it gives the measure of the efficiency of the process and it is used as the measure of costs of the ICU and of the global hospital (the ICU while using between 5% and 10% of the beds can consume up to 20% of hospital budgets). Secondly, the LoS can be used as an indirect indicator of the quality of care [121]. Figure 6.5 shows the yearly average LoS, in general and in different types of patients. The reference solid line ($Y = 3.02$) is the average of all admitted patients (aggregated from 1999 to 2012) and the dashed line shows the trend of average LoS for different types of patients. In Table 6.1 the values of the averages for each type of patients aggregated from 1999 to 2012 are reported.

Figure 6.4: Time trends of devices number in San Camillo ICU - Process indicator

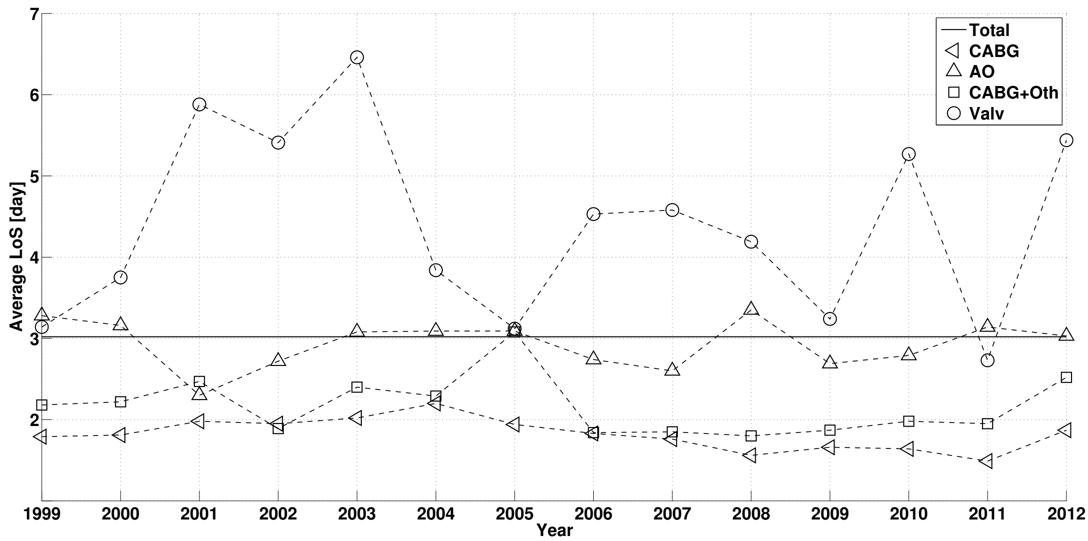


Figure 6.5: Yearly average LoS for different type of patients in San Camillo ICU - Process indicator

Average LoS				
Total	**CABG**	**AO**	**Valv**	**CABG+Other**
3.02	2.08	3.78	2.76	4.48

Table 6.1: Average LoS for each type of patients in the ICU of San Camillo.

Among process indicators patient risk parameters are also included that describe the clinical condition of the population admitted into the ICU. The reason for this choice is that these parameters give a description of those which are the inputs to the activities that represent processes. The risk parameters analyzed are:

- Age of patients
- ES of patients.

Table 6.2 shows the percentage of patients being part of the represented age range. The

< 50	$50-60$	$60-70$	$70-80$	> 80
11.22%	17.70%	33.39%	32.86%	4.83%

Table 6.2: Percentage of patients being part of the represented age range in the ICU of San Camillo.

majority of patients admitted in ICU were more than 60 years old (about 70% of the whole population). Instead, in Figure 6.6 the percentages of patients are represented in every considered age range, divided in every year. The most important thing to observe is the increase of the percentage of patients > 80 *years*, from an average of 3.2% in the first seven years to 6.5% in the last seven, thus showing the fact that the Unit has faced a remarkable changing of the population due to the increased age of the population and the capacity of clinical staff to face more critical patients.
The last indicator considered in this section is EuroSCORE (ES). EuroSCORE (European System for Cardiac Operative Risk Evaluation) is a risk indicator for patients undergoing cardiac surgery, which has been widely used since it was introduced in 1999. For simplifying the use of the system and promoting the evaluation of the risk even in a time where information technologies were not so important, the ES was first published and largely clinically accepted as an additive system in which it was given to each risk factor a *weight* or a number of points that, when combined, provided an estimate of the expected percentage mortality for a patient undergoing a cardiosurgery procedure.

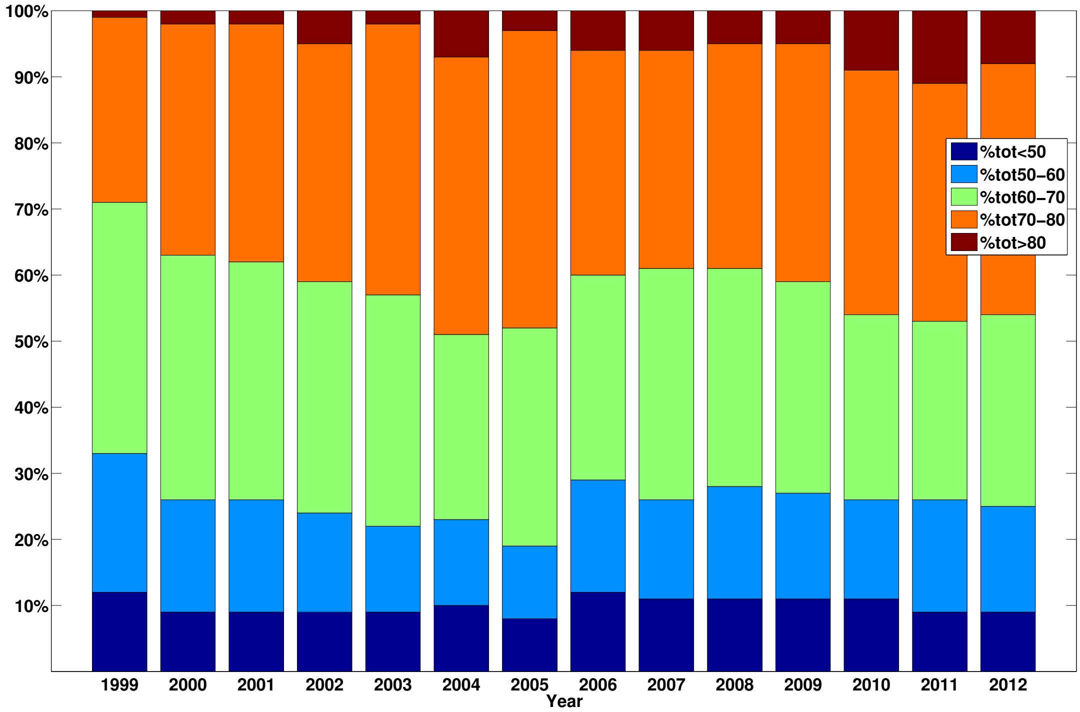


Figure 6.6: Percentages of patients in every considered age range recorded in San Camillo ICU - Process indicator

However, because of its addictive nature, the additive ES underestimates the risk in some groups of patients. At the same time, there has been an exponential growth in the availability of information technologies for the cardiac and surgical units of hospitals, which explains why the use of a risk model based on the equation complete logistics was gradually established [122]. Analysis included concerns only logistic ES. Approximating ES values to integer numbers, Figure 6.7 represents the number of patients per every specific logistic ES value. Figure 6.7 shows that the majority of patients (about

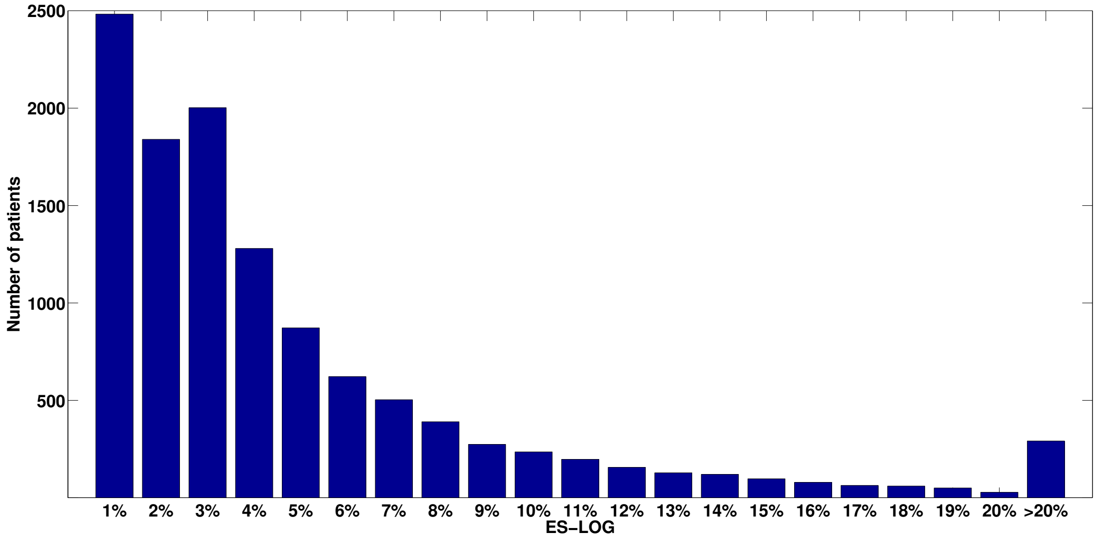


Figure 6.7: Number of patients for every specific logistic ES values in San Camillo ICU - Process indicator

75% of the whole population) falls into the first six ranges of ES.

Outcome The outcome indicators analyzed are the exitus and the complications of the patients. Starting from the first one, Figure 6.8 shows the yearly mortality rate for two cluster of patients based on the different types of intervention:

- No distinction between different types of intervention (Total)
- Patients with isolated CABG intervention (CABG)

Patients with isolated CABG intervention were considered, because it was well accepted as proxy of common cardiac surgery performance indicator. Figure 6.8 shows a decreasing trend of the mortality rate: the average mortality rate considering the whole population is 5.2%, while the other one considering only patients with CABG intervention is 2.2%. Analogue analysis of the mortality rate has been done considering only

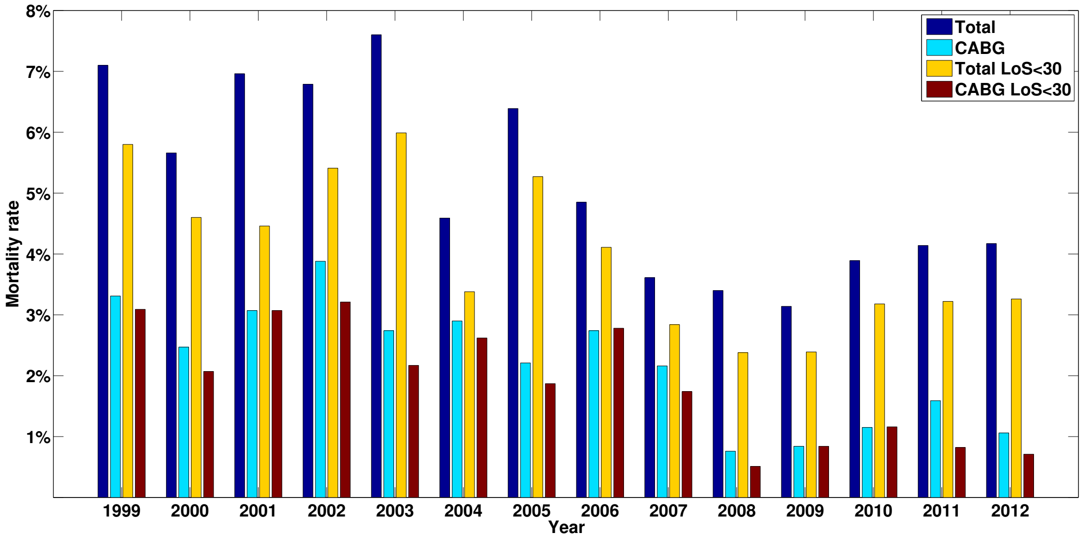


Figure 6.8: Yearly analysis of mortality rate for CABG intervention in San Camillo ICU - Outcome indicator

patients with LoS< 30 days, because this kind of indicator allowed making a comparison between the ICU here considered and some other national or foreign ICUs. Data shows that the overall average mortality rate is 4.02%, instead, patients with LoS< 30 and with only isolated CABG intervention have an average mortality rate, of 1.90%. The analysis of the most important critical care quality indicators makes it necessary

	All	***LoS* < 30**
CABG	2.2%	1.9%
Other surgery	5.2%	4.2%

Table 6.3: Comparison of CABG and other surgery mortality rate in the ICU of San Camillo.

to discuss the these results. In particular a comparison between the expected mortality rate (following only the logistic ES) and the real mortality rate should be done. According to clinical practice it was necessary to exclude, in this type of analysis, very critical groups of patients that are not involved in the ES calculation: these patients present specific interventions (VAD, BIVAD, Levitronix, ECMO, Biopump, Artificial

Heart, Transplantation) or a very long Length of Stay (LoS> 30 days). First of all, in Table 6.4 the global difference between the real and the expected mortality is analyzed considering the whole population of patients. The observed mortality is lower than the

	Mortality rate
Expected [ES]	4.94%
Observed	3.80%

Table 6.4: Real and expected mortality rate of ICU of San Camillo.

expected of 1.14% suggesting the fact that the Unit has very successful performances. Secondly, in Table 6.5 the difference between the expected and observed mortality rate was analysed, but now considering different ranges of ES in order to check the unit's performances in homogeneous groups of patients as much as possible. The Table shows

	ES Range			
Expected [ES]	2.12%	4.91%	8.32%	18.94%
Observed	1.20%	3.76%	7.81%	14.85%

Table 6.5: Real and expected mortality for different ES in ICU of San Camillo.

that the Unit mortality rate is better than the expected one in each range, and the unit's performances are good. Lastly, in Figure 6.9 the difference between the mortality rates is analyzed considering the observed mortality rate values for each unit of logistic ES (the real ES values of patients were approximated to the numeric unit in order to have 1%, 2%, 3%, etc.). The Figure 6.9 shows that the performances of the unit are better than the expected until the group of patients with logistic ES 12%. After this value, the curve of the observed mortality fluctuates around the ideal curve. This is probably because the number of patients falling into these groups is too small to have a significance, and it is therefore difficult to explain the corresponding results.
The second outcome indicator concerns clinical complications of patients (i.e. bleeding, redo, and general neurological episode). In Table 6.6 is shown the number of patients having a complication (including isolated exitus).

6.1.3 Statistical Quality Control in Intensive Care Unit

Another analysis of exitus has been made using the instruments of SQC, in particular the Control Chart which are usually used to study the stability of a process. Even

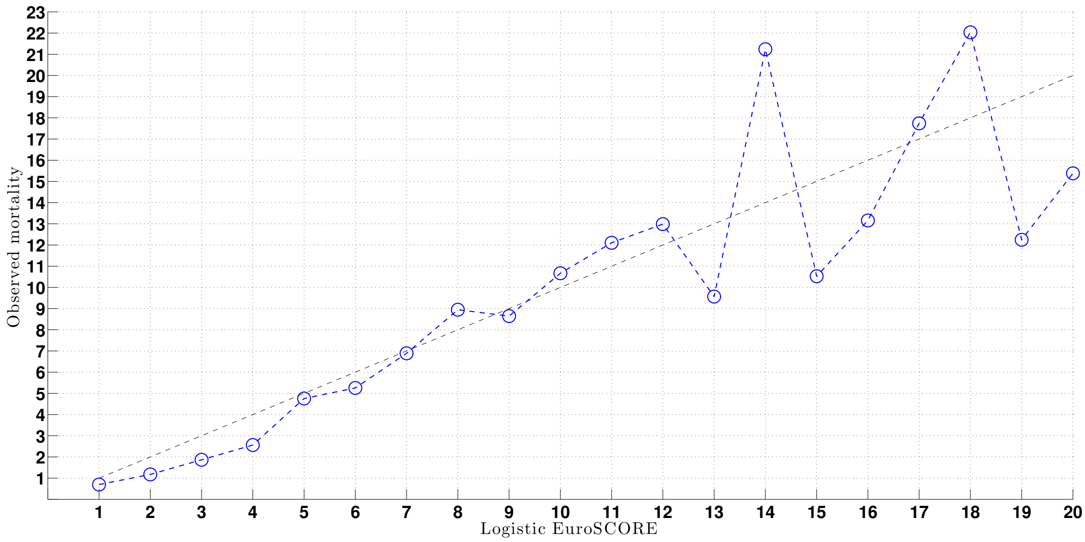


Figure 6.9: Logistic ES forecast ability of observed mortality in San Camillo ICU - Outcome indicator

	With complication	Without complication	Total
Number	9379	2391	11770
Percentage	79.7%	20.3%	

Table 6.6: Number patients having a complication in ICU of San Camillo.

in the study of health services the SQC techniques are useful to monitor the stability of performance indicators (as LoS and exitus) but also to evaluate their variability due to external factors and to process changes (e.g. introduction of new surgical or clinical protocols) [105], [106], [109], and [108]. It is well know that the Control Chart is a statistical tool used to distinguish between variations in a process resulting from common causes and variation resulting from special causes. One goal of using a Control Chart is to achieve and maintain process stability. Process stability is defined as a state in which a process has displayed a certain degree of consistency in the past and is expected to continue to do so in the future. This consistency is characterized by a stream of data falling within control limits. The most important use of a control chart is to improve the process.

- Most processes do not operate in a state of statistical control
- Routine and attentive use of control charts will identify assignable causes. If

these causes can be eliminated from the process, variability will be reduced and the process will be improved.

- Control charts will only detect assignable causes. Management, physicians, nurses and engineering action will usually be necessary to eliminate the assignable causes.

Generally SQC is a largely used method to investigate, by statistical instruments, the variability of a process. The variability of mortality rates among patients is studied.

P-CHART This attribute control chart is widely used for monitoring a process quality by the knowledge of the fraction (proportion) of non conformity. This fraction is defined as the ratio between the number of non conformity compared with the total number of the population. The analyzed set of patients may have several quality characteristics that are examined simultaneously. If the patient does not conform to standard on one or more of the characteristics, it is classified as non conforming. The characteristic analyzed is the exitus, the patient is non conforming when his exitus is positive. The statistical principles underlying the control chart for a non conforming fraction are based on the binomial distribution [115]. Therefore, considering a $3-\sigma$ interval the Control Limit and the Central Line are given by

$$\begin{aligned} \text{Upper Control Limit (UCL)} &= p_0 + 3\sqrt{\tfrac{p(1-p)}{n}} \\ \text{Central Line (CL)} &= p_0 \\ \text{Upper Control Limit (LCL)} &= p_0 - 3\sqrt{\tfrac{p(1-p)}{n}} \end{aligned} \tag{1.1}$$

In the following exitus over the observed years are reported. For each year all the patients are taken and are computed into the sample fraction of non conforming $\hat{p}$, and the statistic of $\hat{p}$ is plotted in the chart. In this case of study it is assumed that $p_0 = 3.8\%$. As long $\hat{p}$ remains within the control limits and the sequence of plotted points does not exhibit any systematic non random pattern, it is possible to conclude that the process in control at the level of p. Moreover Figure 6.10 P-chart shows that the first year represents a statistically out-control process for exitus and successively the process is established is in-control (excepted for 2009 data). Out-control process means that a special cause of variation has occurred. The decreasing trend itself is an indicator of the improvement of the Unit performances since the formation of the team.

Figure 6.10: P Chart of Exitus of San Camillo ICU - Statistical Quality Control

U-CHART In other cases, the measure of quality does not consist in the monitoring of non-conforming products in output from the process, but in the calculation of the average number of non-conformities per unit of reference, therefore this attribute control chart (U-chart) assumes the possibility that each unit may present more than one event of non-conformity as opposed to the P-Chart. This second approach involves setting up a control chart based on the average number of nonconformities per inspection unit. If an x total non conforming are found in a sample of n inspection units, then the average number of non conformities per inspection unit is:

$$u = \frac{x}{n} \tag{1.2}$$

x is a Poisson random variable; consequently, the parameters of the control chart for the average number of non conformities per unit are as follows:

$$\begin{array}{rl} \text{UCL} & = u_0 + 3\sqrt{\frac{u_0}{n}} \\ \text{CL} & = u_0 \\ \text{LCL} & = u_0 - 3\sqrt{\frac{u_0}{n}} \end{array} \tag{1.3}$$

where $\bar{u}$ is the observed average number of non conformities per unit in a preliminary set of data. In this case the non-conformity is always the exitus, but the unit of reference is the number of days stay in hospital for all patients every year. In our case of study $u_0 = 1.5\%$. U-charts 6.11 shows that the exitus are out-control only in the first period

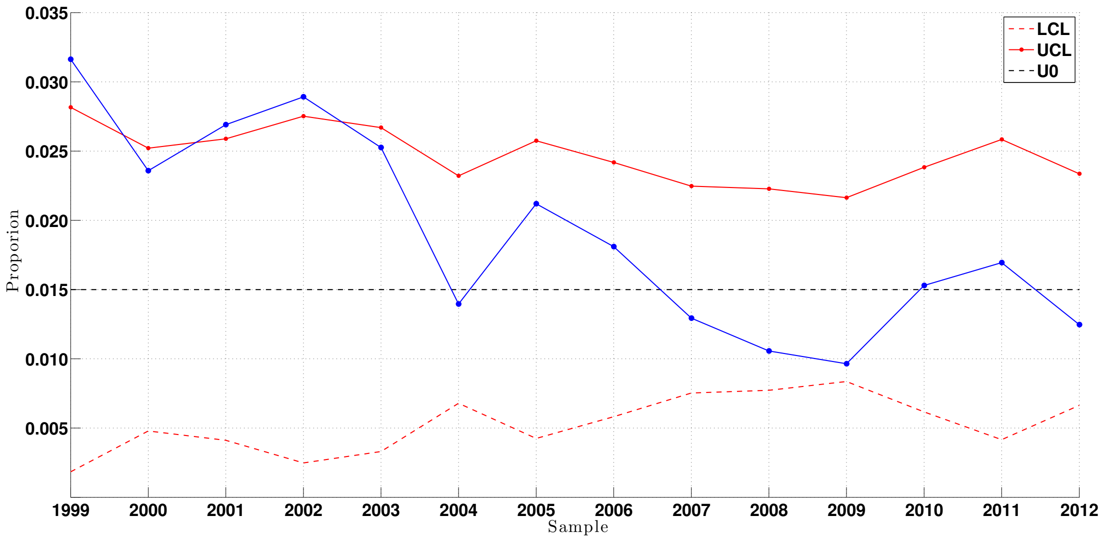


Figure 6.11: U Chart of No. Exitus of San Camillo ICU - Statistical Quality Control

that the team settled, and successively are not, showing again the improvement of Unit performance thanks to stabilized team skills.

6.2 Stochastic dynamic programming in hospital resource optimization

Healthcare decision-makers, especially in areas of hospital management, is rarely fortunate enough to have all necessary information made available to them at once. As a result, their decisions occur sequentially as information becomes available and situations around them change. This problems can be understood as sequential decision making processes and can be modelled using stochastic dynamic programming (SDP) [123]. Stochastic programming is an approach for modelling optimization problems that involve uncertainty. When the parameters are uncertain, but assumed to lie in some given set of possible values, one might seek a solution that is feasible for all possible parameter choices and optimizes a given objective function. Stochastic programming models take advantage of the fact that probability distributions governing the data are known or can be estimated. These models apply to settings in which decisions are made repeatedly in essentially the same circumstances, and the objective is to come up with a decision that will perform well on average [124].

In healthcare resource allocation problem an under-provision of hospital beds leads to patients in need of hospital care being turned away. When insufficient medical beds are provided to meet demand, emergency medical patients spill over into surgical beds; consequently, waiting lists increase as planned admissions are postponed. On the other hand, the goal of the hospital is to assign beds in order to provide the best level of service possible without wasting resources. We here address this dilemma by minimising the number of empty beds subject to maintaining the delay probability at a sufficiently low level [10], [11].

Stochastic model are successfully used in many aspects of healthcare resource optimization; appointment scheduling optimization [125], capacity planning [126], [127], nurse assignment [128], medical care use [129], discharge policies [130], and operating room schedule [131]. Much research has aimed at maximizing operating room utilization, due to its high operational cost ([132], [133], [134]). However, [135] showed that, at hospitals with fixed or nearly fixed annual budgets, allocating operating room time based on utilization can adversely affect the hospital financially, and suggested considering not only operating room time but also the resulting use of hospital beds. There have recently been some studies on the impact of surgery schedules on the use of the other resources in hospitals. For example, [136] developed analytical models and solution heuristics for building cyclic master surgery schedules to minimize the expected total bed shortage. [137] presented a multi-objective non-linear optimization model for surgery scheduling, taking surgical priorities, surgery length, demand for equipment and conflicts in the schedule into account.

SDP approach is well-positioned to model these types of problems because of the explicitly sequential nature of the decision policies they produce. Our aim is to reduce the probability of having a number of patients different from fixed level over a define interval of time in the Intensive Care Unit (ICU) department of San Camillo. Intensive care units are a costly resource [138], [139]. They are oversubscribed and the treatments they offer are expensive and labour intensive. Resources within the hospital are finite; therefore, it is important that intensive care clinicians are aware of how the costs of an individual unit are incurred and how they relate to its therapeutic activity, case mix and clinical outcome. The process of cost analysis helps to allocate resources efficiently, thereby improving both quality and quantity of ICU provision. An example

is the analysis of the discharge strategies in order to optimize the resource consumption [130].
To control the number of patients in the department physician can not force the discharge rate because it strictly depend on patients health state and on patients number. In the other hand physician can decide the number of patients that could be admitted. SDP formulations of this problem generate meaningful insights that lead to high quality and intuitive heuristic procedures.
In this work a stochastic optimization model is proposed in order to provide operational guideline in ICU management. To optimize the resource the hospital ward the number of beds used should be kept under control. Therefore our aim is to present a admission policy that stabilize the number of the hospital bed used considering a stochastic perturbation represented by the patient discharges. Considering fixed and a variable number of target patients different empirical policies are compared. An innovative method to describe discharge probability distribution it is also presented.

6.2.1 Data description

The Unit of Anesthesiology and Reanimation of *San Camillo-Forlanini* Hospital in Rome is part of the Department of Cardioscience. The team of the Unit has settled in 1999, and even started to collect some important data about patients admitted. The data analyzed are referred only to the patients coming from Cardiac Surgery Unit, which accurate studies consider an efficient intensive care department in the Italian health care system. In particular, the database is made of 11770 patients (records) collected from 1999 to 2012, and the corresponding items were: Intervention Date, length of stay in Intensive Care Unit, and EuroSCORE (ES). The unit counts 12 beds and the team is made up of one head physician, 19 staff physicians, one head nurse, and 36 staff nurses. The activity of the unit is aimed at supporting the whole department, especially the Cardiac Surgery Unit and the Vascular Surgical Unit.
According to [28] LoS is considered to be a reliable and valid proxy for measuring the consumption of hospital resources. We consider ES as a synthetic proxy for the admission health status of the patient and we assume that exist a relation between the ES and the LoS. Consider, for example, two patients that have similar treatments. We expect that the patient with a measure of risk lower - thus risk and complication less likely - will spend less time in hospital. In fact, ES is a method of predicting the

chances of dying during or shortly after undergoing heart surgery [140]. ES is a tool that was designed for doctors to work out how risky heart operation is. All operations and other treatments have benefits and risks. It is important to underlying that ES can only tell about the risks and not the benefits. Although patients exact scores can only be determined by someone who has access to all the detailed medical history [141]. In order to investigate this relation between LoS and ES, the patients are

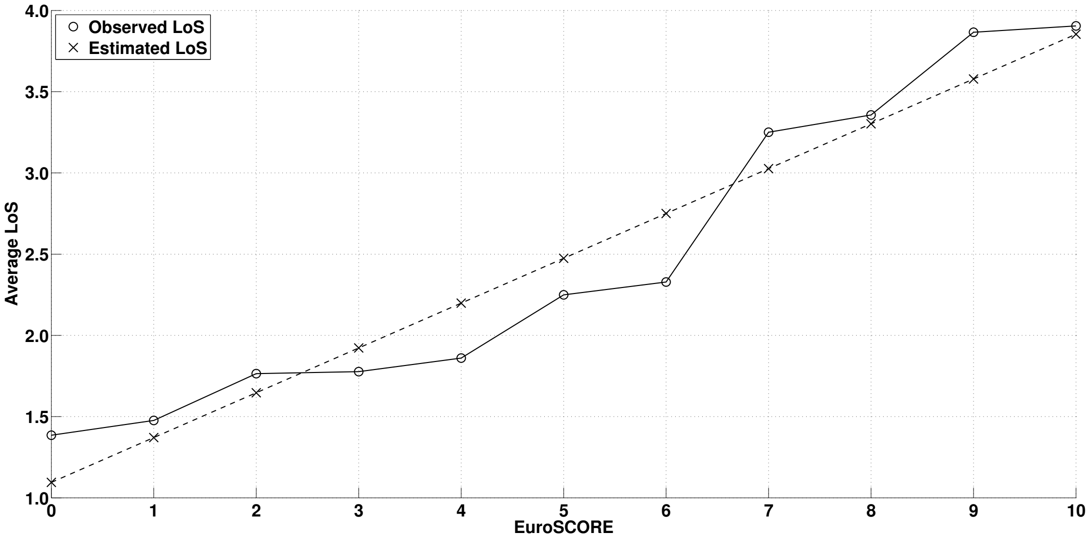


Figure 6.12: Relation between average LoS and ES of San Camillo ICU - Data recorded from 1999 to 2012

divided considering the ES and for each cluster is evaluated the average LoS. In Figure 6.12 x-axes represents the value of the ES and y-axes the average LoS, circle marks represent the average LoS estimated in the cluster, and x marks define the average LoS estimated using a linear regression. The ES value is recorded when the patient is admitted, therefore the expected LoS ($L\hat{o}S$) is estimated by the following equation

$$L\hat{o}S = \beta_0 + \beta_1 \times ES. \tag{2.4}$$

The coefficients estimated from the linear regression are $\beta_0 = 1.11$ and $\beta_1 = 0.28$, and the Average Absolute Error is 0.2045. The value of AAE suggests that exist a linear relation between LoS and ES. The ES value is recorded when the patient is admitted therefore it is possible to estimate the expected LoS. In Table 6.7 are reported the details of the cluster with more then 200 patients. We observe that for bigger value

ES	No. patients	$\bar{LoS}$	$\hat{LoS}\sigma$	
0	825	2.39	2.10	1.40
1	1216	2.48	2.37	1.54
2	1049	2.76	2.65	2.21
3	1677	2.78	2.92	2.21
4	1624	2.86	3.20	2.38
5	1500	3.25	3.47	3.00
6	1236	3.33	3.75	3.09
7	887	4.25	4.03	4.67
8	586	4.36	4.30	4.73
9	403	4.87	4.58	5.26
10	210	4.91	4.85	5.68

Table 6.7: Clusters details: $\bar{LoS}$ is the observed average LoS, $\hat{LoS}$ is the estimated average LoS by the linear model, and σ is the standard deviations of the LoS values in each cluster of San Camillo ICU.

of ES - riskier patients - standard deviations of the LoS increase. This behaviour can be explained considering that the ES is a proxy of the admission health status. If a patients has bigger ES he has a lower health status therefore his hospitalization will be more long and uncertain [142].
Hospital ward is a system characterized by a finite number of resources, number of bed is limited therefore if new patient requires urgent treatment is reasonable to expect that less severe patients are discharged or transferred to other departments. Hence, we assume that the average LoS also depends on the number of patient hospitalized in the department. The physician, knowing the health status of his patients, can decide to admit a new patient even if the department is almost full. He will expect a discharge, after a short period, of the pre hospitalized patients. In order to validate this assumption, for each patient are counted the average number of patients that stay with him during the hospitalization. Considering a patient P that stay in hospital for 5 days. During his hospitalization in the department he is treated with other patients. For example the number of patients is reported in Table 6.8

Day	**1**	**2**	**3**	**4**	**5**
Np	5	3	5	3	4

Table 6.8: Number of patients in the department during the hospitalization of patients P.

Given this scenario it is reasonable to assume that patient P in average stay in hospital with other 3 patients. In this way, we associate each patient hospitalization with a proxy of the *load factor*[1] of the ward. Then data are clustered considering the average number of patients and, for each cluster, the average LoS is evaluated. A similar procedure is performed in order to evaluate the average ES of clusters. First plot of Figure 6.13

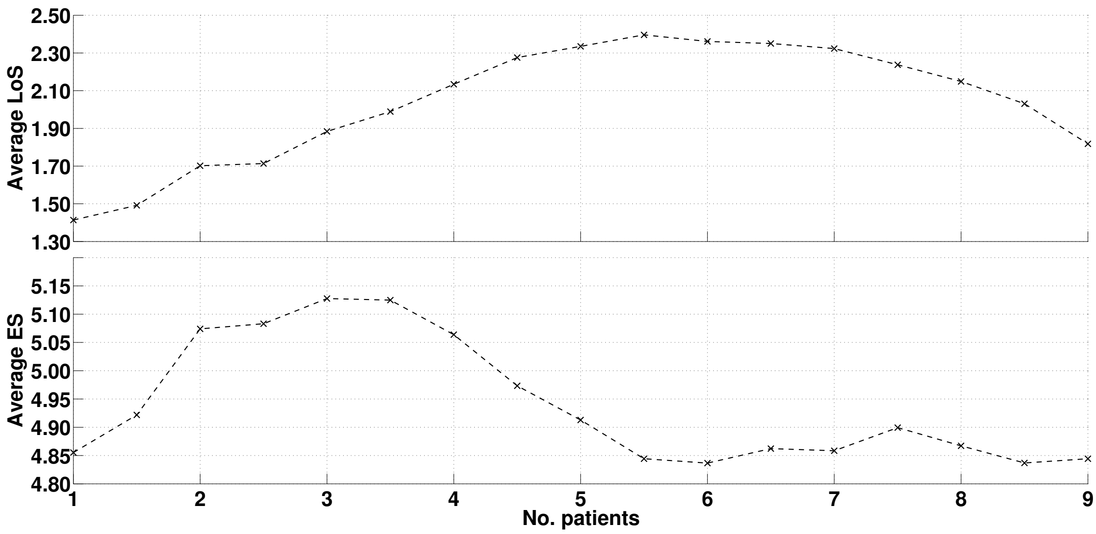


Figure 6.13: Average LoS and ES for different patient cluster of San Camillo ICU - Data are clustered considering the average number of patients and, for each cluster, the average LoS is evaluated

shows the relation between the average number of patients in the department (x-axes) and the average LoS (y-axes). Data shows that exist a number of patient that leads to the maximum LoS. As could be expected, an increase of the patient number, leads to an increase of the LoS. However after this fixed amount the LoS starts to decrease. This feature can be explained considering that the ward is almost full and the admission demand forces the physician to discharge early patients. Considering only the number of patients could be misleading: the number of resources required could depend on the health status of the patients. In second plot of Figure 6.13 are reported the relation between the average number of patients in the department (x-axes) and the average ES (y-axes). When the ward is almost full, the average ES decreases, conversely sicker patients are treated when in the ward are hospitalized less patients.

[1]The ratio of actual number of hospitalized patients and the maximum number possible

6.2.2 Model

A number of researchers have investigated patient demand and bed capacity planning at a specific department within a hospital. McClain has developed a stochastic model to forecast the allocation of non-obstetric patient-days to the obstetric unit and to predict the effect of such allocations on demand for obstetric beds [143]. Dexter and Macario have modelled the distribution of patients at an obstetrical unit as a Poisson distribution and minimized the number of staffed beds subject to remaining below a specified probability of patient overflow [144]. Harris has developed a simulation model to aid decision making in the area of operating theatre time tables and the resultant hospital bed requirements [145]. Furthermore, many authors have created models for the entire hospital, while capturing the inherent variability in patient arrival and length-of-stay [12], [146]. They have demonstrated that managing capacities based on simple deterministic spreadsheet calculations typically do not provide the appropriate information, and result in underestimating true bed requirements. However, they ignore the patient demand for beds at each department within a hospital, such as emergency rooms, intensive care units, and acute care units. To calculate the patient demand at each department, Gorunescu et al. and Harrison have used compartment models, in which a facility is subdivided into categories of patients with different transition rates to model patient flow through wards [10], [147].
Healthcare providers recognize the importance of implementing simulation to support quality learning outcomes [148]. It has been applied to practically every topic in healthcare, such as space considerations, physiology, crisis management, critical care, and general surgery [149]. In comparison to analytical models, more procedural details can be included in a computer simulation model [150]. Linear or nonlinear programming models, queuing models and Markov chains often rely on closed-form mathematical solutions [151]. They are more sensitive to the size, complexity and level-of-detail required by the system under study. Simulation models, on the other hand, are much less sensitive to these parameters [151]. However, simulation may be more difficult to use for several reasons [143]. First, the added complexity of constructing a more realistic model requires considerable institution-specific data that may be costly to collect. Second, computer programming is usually expensive and time-consuming. Third, forecasts of parameters used in such models are often subject to significant error, which

may negate gains in accuracy achieved through simulation. Sinreich and Marmor incorporate three principles to minimize the short-comings of simulation, and to increase managements involvement and confidence in their model [151]:

1. The simulation tool has to be general and flexible enough to model different possible hospital settings.

2. The simulation tool has to be intuitive and simple to use. This way, managers, hospital engineers, and other nonprofessional simulation modelers can run the simulation tool with very little effort.

3. The simulation tool has to include reasonable default values for many of the system parameters. This will reduce the need for comprehensive, costly, and time-consuming time and motion studies, which are usually among the first steps taken when building any simulation model.

Sinreich and Marmor satisfy the first principle by testing their model against five hospital data sets. They address the second principle by designing a user-friendly interface that mirrors a unified patient process chart, which managers are familiar with. To comply with the third principle, default values are used in the simulation and can be easily accessed through the models interface. In this thesis, our approach follows these three principles closely to offset the difficulties of developing simulation models.
One of the most promising techniques for multiple objective decision analysis is Goal Programming. Goal Programming is a powerful tool which draws upon the highly developed and tested techniques of Linear Programming, but provides a simultaneous solution to a complex system of competing objectives [152]. Goal Programming can handle decision problems having a single goal with multiple sub goals. In Goal Programming, instead of attempting to maximize or minimize the objective function directly as in the Linear Programming, the deviations between goals and what can be achieved within the given set of constraints are minimized. Three steps of goal programming model include define the decision variables, define the goals, define the deviational variables [153].
A linear program is an optimization problem where we want to minimize a linear function subject to some linear constraints. In a dynamic program, we want to make a series of decisions in such a way as to maximize some function. The key is that the set

of decisions available to us at any given time depends on what we have already decided to do previously. The key difference between the two is that in dynamic programming assumes a continuous decision process over the time, whereas in linear programming all the decisions are taken are taken up front. Stochastic dynamic programming is a methodology for sequential decision making over periods which was originally used in the 1940s by Richard Bellman to describe the process of solving problems where one needs to find the best decisions one after another [154]. At each decision epoch, a decision maker observes the state of a system and chooses an action, and then gets an immediate reward and moves to a new state according to the probability distribution determined by the action choice. Because of the dynamical structure of the problem, this process continues, but now the system may be in a different state and there may be a different set of actions. The value of each state at a given time can be found by working backwards, using a recursive relationship called the Bellman equation. Bellman equation is useful because it reduces the choice of a sequence of decision rules to a sequence of choices for the control variable. There are different algorithms to solve stochastic dynamic programming problems suitability of which depend on the planning horizon. Dynamic stochastic programming models take advantage of the fact that the probability distribution governing the data are known in explicit form or can be estimated [123]. In our framework the number of discharge is the random parameters that influences the state variable. In order to describe the number of discharge probability distribution we propose 3 different model.

Binomial model The number of patients in the department is given by the number of new patients admitted and on the number of patients discharged. Considering the number of patient hospitalized and the external demand, we assume that the physician decides the patient admission number [155]. Since the number of discharge can not be settled exactly because depends on the health status of the patients hospitalized. We also assume that it is possible to describe the probability density function of the discharge number u_{k-1} given the hospitalized number of patient n_k. In Table 6.9 are reported the average number of discharge $\mu_u^{(n)}$ and the standard deviation $\sigma_u^{(n)}$ with respect to the number of patients hospitalized. We describe the conditional distribution of the discharge number u_k using a binomial model. In the rest of the define we define all the random variable on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathcal{F}$ σ-algebra on Ω and

n_{k-1}	No. days	$\mu_u^{(n)}$	$\sigma_u^{(n)}$
1	193	0.30	0.46
2	405	0.86	0.78
3	658	1.39	0.95
4	795	1.77	1.20
5	974	2.19	1.37
6	836	2.72	1.47
7	605	3.21	1.60
8	336	3.55	1.61
9	168	3.84	1.65
10	64	4.67	1.52

Table 6.9: Average number of discharge $\mu_u^{(n)}$, deviation standard $\sigma_u^{(n)}$, and number of patient in each cluster of San Camillo ICU.

$\mathbb{P}$ probability measure. The binomial distribution is the discrete probability distribution of the number of successes in a sequence of n independent Bernoulli experiments, each of which yields success with probability $p_b \in [0,1]$. In our framework, we consider the patient discharge process as a Bernoulli experiment. Hence the probability mass function is given by

$$\mathbb{P}[u_k = u | n_{k-1} = n] = \binom{n}{u} p_b^u (1-p_b)^{n-u}, \tag{2.5}$$

where the probability parameter of the conditional distribution are estimated using the Minimum Mean Squared Error (MMSE) method and the binomial number of trial are equal to the number of hospitalized patients. In Table 6.10 the main results of the estimation are reported, the small values of AAE prove that the Binomial model provides a good description of the empirical frequency.

Conditional Binomial model The previous approach assumes that the number of discharge patient depends on the number of hospitalization without take in consideration the health status of patients. Each patient has a unique medical history and react differently to the treatment, it is impossible to quantify uniquely the health status after the same operation. Even if we consider patients with the same ES we can not exactly forecast which will be the health status [156]. Hence we have to introduce a proxy of the health status that provides some information on the probability distribution of the discharge number. When a patient is admitted, using his ES value, we evaluate

$n_{k-1} = n$	$p_b^{(n)}$	AAE
1	0.30	0.00 $\times 10^{-3}$
2	0.40	20.16 $\times 10^{-3}$
3	0.49	16.77 $\times 10^{-3}$
4	0.47	29.08 $\times 10^{-3}$
5	0.48	27.75 $\times 10^{-3}$
6	0.50	22.24 $\times 10^{-3}$
7	0.50	20.73 $\times 10^{-3}$
8	0.48	17.30 $\times 10^{-3}$
9	0.46	26.64 $\times 10^{-3}$
10	0.47	26.60 $\times 10^{-3}$

Table 6.10: Discharge distribution parameters and estimated AAE using a binomial probability mass function to fit the data of San Camillo ICU.

the expected LoS through 2.4. Let $v_{i,k}$ be the expected LoS of patients i at time k. After 1 day of hospitalization, if the patient has not been discharged his expected LoS decreases by 1, therefore we can write:

$$v_{i,k+1} = v_{i,k} - 1 \tag{2.6}$$

If department is occupied mainly by patients that has lower expected LoS we expect an increase in the number of discharges. The number value $v_{i,k}$ represents the remaining expected LoS of patient i at time k, nevertheless during some hospitalization, the value $v_{i,k}$ could be less then zero. It means that, given the linear model 2.4, the patient should be already discharged.
Let A defined as

$$A_k = \sum_{j=1}^{n_k} \mathbb{1}_{v_{j,k}<0} \tag{2.7}$$

where $\mathbb{1}$ stands for the indicator function. The value of A_k counts the number of patients that we expect to discharge and depends on n_k. Assuming that patients health status are independent, then we describe the probability distribution of A_k conditional to n_k using a Binomial model. Hence conditional probability mass function is

$$\mathbb{P}[A_k = a | n_k = n] = \binom{n}{a} p_A^a (1 - p_A)^{n-a}. \tag{2.8}$$

For each value of n_k probability parameters of the distribution are estimated using the MMSE method and the binomial number of trial are equal to the number of hospitalized

$n_k = n$	p_A	AAE
1	0.46	0.00 $\times 10^{-3}$
2	0.28	6.08 $\times 10^{-3}$
3	0.26	4.98 $\times 10^{-3}$
4	0.28	5.93 $\times 10^{-3}$
5	0.29	4.86 $\times 10^{-3}$
6	0.27	6.97 $\times 10^{-3}$
7	0.28	4.08 $\times 10^{-3}$
8	0.29	8.55 $\times 10^{-3}$
9	0.27	7.99 $\times 10^{-3}$
10	0.24	15.26 $\times 10^{-3}$

Table 6.11: A probability distribution parameters and estimated AAE using binomial probability mass function.

patients. In Table 6.11 the parameters of the binomial distribution and the AAE are reported. In Figure 6.14 the empirical and estimated distribution of A_k for $n_k = 6$ are reported. AAE values in Table 6.11 and plots in Figure 6.14 show that Binomial model provides a good description of the empirical distribution. Once we have described the probability distribution of A we characterize discharge frequency using the value of A. Number of patients that has an expected LoS less or equal to zero, could be an observable proxy of the discharge number. In this case, we assume that the value of A_k will provide information on the discharge number of patient at time k. We cluster the data considering the value of A at time $k-1$, and for each cluster, we evaluate the average number of discharges. Table 6.12 reports the average number of discharges

A_{k-1}	No. days	$\mu_u^{(A)}$	$\sigma_u^{(A)}$
0	1175	2.00	1.54
1	1767	2.20	1.61
2	1211	2.26	1.60
3	589	2.47	1.69
4	230	2.69	1.60
5	66	3.27	1.65

Table 6.12: Average number of discharge $\mu_u^{(A)}$, deviation standard $\sigma_u^{(A)}$, and number of day in each cluster.

($\mu_u^{(A)}$) and standard deviation ($\sigma_u^{(A)}$) conditional to A. For each cluster we model the empirical distribution of u_k using a weighted average of a Binomial distribution and a

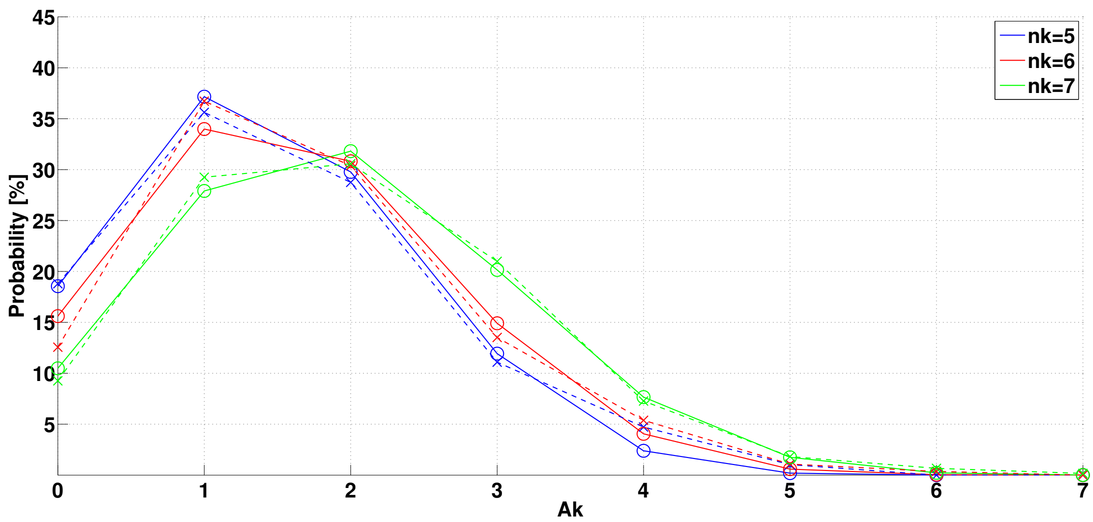


Figure 6.14: A_k **density function of San Camillo ICU for** $n_k = 6$ - Dashed lines with x markers reppresent the observed frequencies, solid lines with circle markers reppresent the estiamted distributions.

Negative Binomial distribution. We set the Binomial number of trials as the maximum number of patients that have been hospitalized in the department ($n_{max} = 13$) and the negative binomial number of failures as the maximum value of A recorded ($A_{max} = 5$). The estimation of parameters is performed using the MMSE method. In Table 6.13

A_{k-1}	w	p_b	p_n	AAE
0	0.83	0.18	0.97	4.86×10^{-3}
1	0.75	0.22	0.91	2.59×10^{-3}
2	0.80	0.21	0.92	5.38×10^{-3}
3	0.72	0.24	0.84	4.34×10^{-3}
4	0.82	0.24	0.82	11.69×10^{-3}
5	0.94	0.28	0.88	22.67×10^{-3}

Table 6.13: Discharge distribution parameters and estimated AAE using a binomial and a negative binomial weighted average probability density function.

p_b is the binomial success probability, p_n is the negative binomial success probability, n_b is binomial number of success, and n_n is the negative binomial number of failure. Data show that, even considering clusters with more patients, the discrepancy indicator AAE is reduced.

The Poisson model It is widely acknowledged that the number of operations, other procedure and diagnostic tests, vary from day to day. Hospital ward works 24/7 but there are peaks and troughs in activity throughout the week. Discharges as other procedure may vary during the week [157]. In Table 6.14 the average and the standard deviation values of the number of admission are reported, number of discharge, and number of patients by day of week. Values show that the average number of discharges depends strongly on the day of the week. According to the previously approach, dis-

	No. Admissions		No. Discharges		No. Patients	
	μ	σ	μ	σ	μ	σ
Mon	3.11	1.53	1.77	1.33	5.09	2.07
Tue	2.98	1.36	2.82	1.59	5.25	2.05
Wed	3.10	1.40	2.91	1.44	5.44	2.12
Thu	2.84	1.36	3.04	1.54	5.25	2.15
Fri	2.95	1.44	2.78	1.46	5.42	2.21
Sat	0.43	0.63	1.55	1.33	4.30	1.96
Sun	0.11	0.37	0.65	0.91	3.76	1.93

Table 6.14: Average and standard deviation values of number of admissions, number of discharges, and number of patients by day of week. Data are estimated in the San Camillo ICU.

charge number depends on the number of patients hospitalized. We also assume that the discharge number depends by day of week. Hence we describe the conditional distribution of u_k using a Poisson model:

$$\mathbb{P}[u_k = u | n_{k-1} = n] = \frac{\lambda_k(n)^u e^{-\lambda_k(n)}}{u!} \tag{2.9}$$

where $\lambda_k(n)$ is the cluster's average number of discharge. In Table 6.15 the average number of discharge (λ) estimated are reported. It is reasonable to assume that the increase in the number of patients corresponds to an increase of the average number of discharge. If the ward occupancy rate raises, in order to control the number of patients, the discharge probability raises too. Data endorse that exists a linear relationship between n_{k-1} and λ_k. Hence the following linear model is assumed:

$$\lambda_k(n_{k-1}) = m(k) \times n_{k-1} + q(k) \tag{2.10}$$

where $m(k) = m(k+7)$ and $q(k) = q(k+7)$. In Table 6.16 are reported the parameters and the estimated AAE between the observed average and the estimated average

n_{k-1}	Mon	Tue	Wed	Thu	Fri	Sat	Sun
1	1.46	1.24	1.31	2.05	1.78	1.08	0.59
2	1.08	1.71	2.06	2.36	2.27	1.17	0.56
3	1.37	2.12	2.49	2.56	2.34	1.55	0.63
4	1.38	2.68	2.55	2.83	2.28	1.72	0.80
5	2.01	2.92	3.02	2.96	2.66	1.69	0.52
6	1.87	3.23	2.97	3.30	3.03	1.48	0.82
7	2.26	3.19	3.38	3.43	3.11	1.50	0.66
8	1.96	3.21	3.30	3.35	3.27	2.04	0.82
9	2.13	3.73	3.42	3.90	3.77	1.83	0.80
10	1.83	3.00	3.71	4.46	3.50	2.00	0.63

Table 6.15: Average number of discharge λ_k for each cluster in the San Camillo ICU.

number of discharges. In Figure 6.15 are reported the values of λ_k estimated from the

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
$\boldsymbol{m}$	0.10	0.23	0.23	0.23	0.20	0.09	0.01
$\boldsymbol{q}$	1.20	1.46	1.56	1.84	1.69	1.12	0.72
AAE	0.20	0.29	0.18	0.09	0.12	0.15	0.18

Table 6.16: λ_k linear model parameters estimated and AAE for the San Camillo ICU.

linear model 2.10. In Table 6.17 the AAE of the Poisson probability mass function fitting obtained through the linear model are reported. Value show that the proposed approach provides a good description of the empirical frequencies.

6.2.3 Stochastic dynamic optimization of bed occupancy

The hospital system that we considered can result in a patient being turned away because all beds are occupied; such a patient may not receive the necessary care. On the other hand, the goal of the hospital is to assign beds in order to provide the best level of service possible. We here address this dilemma by minimizing the number of empty beds subject to maintaining the delay probability at a sufficiently low level. The most common goal in the above setting is to find a *policy* that specifies the action to take in each time in order to reduce the cost. A key idea is that an optimization over time can often be regarded as *optimization in stages*. We trade off our desire to obtain the lowest possible cost at the present stage against the implication this would have for the cost at future stages. Therefore the best policy minimizes the sum of the cost incurred

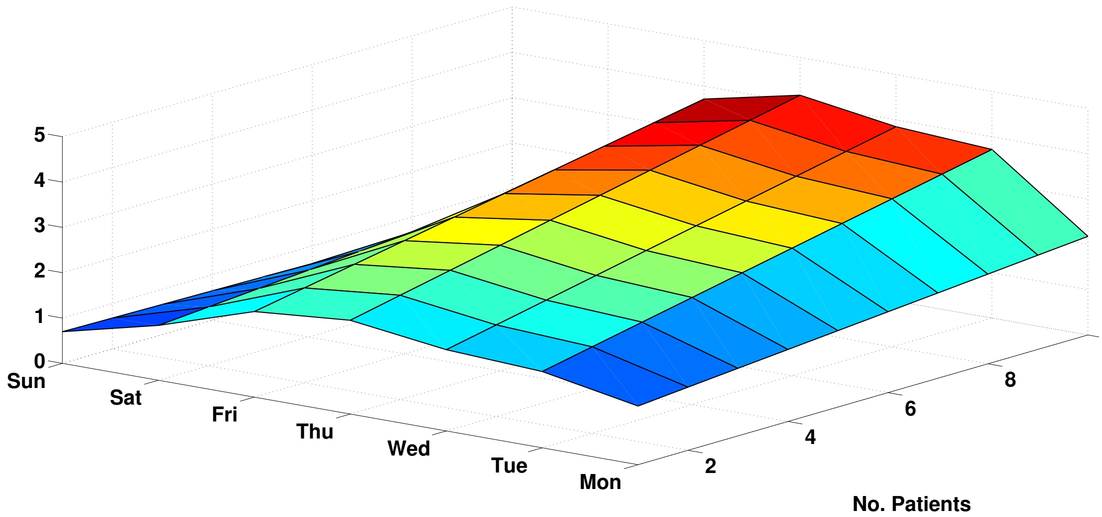


Figure 6.15: Poisson model estimated λ_k of San Camillo ICU - Discharge probability model.

at the current stage and the least total cost that can be incurred from all subsequent stages, consequent on this decision.
A dynamic programming approach over a finite time horizon allows to fine this policy. This control theory accounts for the fact that a dynamic system may evolve stochastically and that key variables may be unknown or imperfectly observed. Our model has two principal features:

- An ***underlying discrete dynamic system***
- A ***cost function*** that is additive over the time

The underlying discrete dynamic system describes the number of patients that are hospitalized in the department at time k. The evolution of n_k, the system's state, under the influence of decision made at discrete instances of time is give by the equation

$$n_{k+1} = n_k + e_k - u_{k+1} \tag{2.11}$$

where n_k is the number of patients hospitalized, e_k is the number of parameters admitted, and u_k is the discharges number. In our model we consider the discharges number as a random parameter that influence the system's state. According to the results obtain in Section 6.2.2 we assume that the conditional distribution of u_k, with

n_{k-1}	Mon	Tue	Wed	Thu	Fri	Sat	Sun
1	0.03	0.05	0.05	0.03	0.04	0.04	0.00
2	0.03	0.02	0.03	0.03	0.02	0.02	0.01
3	0.03	0.02	0.02	0.02	0.02	0.01	0.02
4	0.01	0.03	0.03	0.01	0.02	0.01	0.01
5	0.02	0.02	0.03	0.02	0.02	0.02	0.01
6	0.01	0.03	0.03	0.03	0.03	0.01	0.02
7	0.02	0.02	0.03	0.03	0.04	0.02	0.01
8	0.03	0.04	0.04	0.05	0.03	0.04	0.01
9	0.03	0.03	0.04	0.03	0.04	0.09	0.02
10	0.04	0.06	0.04	0.07	0.04	0.13	0.00

Table 6.17: AAE of the Poisson probability mass function fitting on San Camillo ICU data.

respect to n_{k-1}, is binomial with the estimated parameters of Table 6.10. We also assume that, in order to plan the activities of the department, is possible to control the number of patients that are admitted. Therefore e_k is the control decision variable to be selected at time k. In discrete time k takes integer values, say $k = 0, 1, \ldots, K$. Let $E_k = (e_0, \ldots, e_{k-1})$ denote the partial sequence of control taken over the first k stages. A ward with all beds assigned can result in a patient being turned away, on the other hand empty beds produce a waste of resources. We here address this feature trying to reduce the randomness of the number of beds occupied. We assume that the waste of resources is described by the monotonically increasing cost function $g : \mathbb{R} \to \mathbb{R}^+$:

$$g(n_k) = (n_k - \bar{n}_k)^\eta \tag{2.12}$$

where $\bar{n}_k$ is the desired number of patients by day of the week and η is a even positive integer number. According to results in Table 6.14, the department works during the week with different load capacity. The cost is additive in the sense that the cost incurred at time k accumulates over time. The total cost is therefore given by:

$$g_N(n_N) + \sum_{k=0}^{N-1} g(n_k) \tag{2.13}$$

where $g_N(n_N)$ is the terminal cost incurred at the end of the process. However, because the presence of u_k the cost is generally a random variable and cannot be meaningfully optimized. We therefore formulate the problem as an optimization of the expected total cost

$$\mathbb{E}\Big[g_N(n_N) + \sum_{k=0}^{N-1} g(n_k, e_k, u_k)\Big], \tag{2.14}$$

where the expectation is with respect to the joint distribution of the random variable involved. The optimization is over the control variable E_k; each control e_k is selected with some knowledge of the current state n_k. In general the constraints set will depend on n_k and the time index k. In our case, the state variable represents the number of patients that are hospitalized at time k, therefore we want ensure that this value is always positive and less then the ward capacity. Thus the following constraints conditionally to $n_{k-1} \in [0, n_{max}]$ must hold:

$$\begin{cases} n_k \leq n_{max} & \Rightarrow \quad e_{k-1} \leq n_{max} - n_{k-1} + u_k \\ n_k \geq 0 & \Rightarrow \quad e_{k-1} \geq \max(0, u_k - n_{k-1}) \end{cases} \tag{2.15}$$

Given two probability levels α,β $\in (0, 1)$, knowing the conditional distribution we define a threshold of u_k given n_k as:

$$u_k^\alpha \doteq \sup\Big\{\bar{u} : \mathbb{P}(u_k \leq \bar{u} | n_{k-1}) \leq \alpha\Big\} \tag{2.16}$$

hence u_k^α is the threshold value such that the probability that the discharge number falls behind this value is less than α. In the same way we define u_k^β such that

$$u_k^\beta \doteq \inf\Big\{\bar{u} : \mathbb{P}(u_k \geq \bar{u} | n_{k-1}) \leq \beta\Big\} \tag{2.17}$$

where u_k^β is the threshold value such that the probability that the discharge number exceeds this value is less than β. Considering the Poisson model the cumulative distribution function conditionally to n_{k-1} is

$$\frac{\Gamma\big(\lfloor u+1 \rfloor, \lambda_k(n_{k-1})\big)}{\lfloor u \rfloor !} \tag{2.18}$$

where $\lfloor u \rfloor$ is the floor function and$\Gamma(\ x, y)$ is the incomplete gamma function defined as $\Gamma(x, y) = \int_y^\infty t^{x-1} e^{-t} dt$. Therefore our constrain set$\Theta\ _{k-1}(n_{k-1})$ given $n_{k-1} \in [0, n_{max}]$ becomes

$$\begin{cases} e_{k-1}^u \leq n_{max} - n_{k-1} + u_k^\alpha \\ e_{k-1}^l \geq u_k^\beta - n_{k-1} \end{cases} \tag{2.19}$$

where $\alpha = 5\%$, $\beta = 5\%$ and e_{k-1}^u is the upper bound at time k and e_{k-1}^l is the lower bound at time k, otherwise$\Theta\ _{k-1}(n_{k-1}) = \{0\}$.

Let $\pi = \{\mu_0, ..., \mu_{N-1}\}$ be the policy where μ_k maps states n_k into control $e_k = \mu_k(n_k)$

and such that $\mu_k(n_k) \in \Theta_k(n_k)$ for all n_k, $k = 0, \dots, N-1$. Given that the expected cost of π starting at n_0 is

$$J_\pi(n_0) = \mathbb{E}\Big[g_N(n_N) + \sum_{j=0}^{N-1} g(n_j)\Big] \tag{2.20}$$

and the optimal cost function is

$$J^*(n_0) = \min_\pi J_\pi(n_0), \tag{2.21}$$

the optimal policy π^* is one the satisfies $J_{\pi^*}(n_0) = J^*(n_0)$.

6.2.3.1 Stochastic dynamic programming algorithm

The stochastic dynamic programming technique rests on a very simple idea, the *principle of optimality*. The name is due to Bellman, who contributed a great deal to the popularization of stochastic dynamic programming and to its transformation into a systematic tool. Roughly, the principle of optimality states the following rather obvious fact.

Principal of Optimality Let $\pi^* = \{e_0^*, e_1^*, \dots, e_{N-1}^*\}$ be the optimal policy for the basic problem, and assume that, when using π^*, a give state n_k occurs at time k with positive probability. Consider the subproblem where we are at n_k at time k and wish to minimize the *cost-to-go* form time k to time N:

$$\mathbb{E}\Big[g_N(n_N) + \sum_{j=k}^{N-1} g(n_j, e_j(n_j), u_j)\Big] \tag{2.22}$$

Then the truncated policy $\{e_k^*, e_{k+1}^*, \dots, e_{N-1}^*\}$ is optimal for this subproblem. The principle of optimality suggests that an optimal policy can be constructed in piecemeal fashion, first constructing an optimal policy for the tail subproblem involving the last stage, then extending the optimal policy to the tail subproblem involving the last two stages, and continuing in this manner until an optimal policy for the entire problem is constructed.
Our aim is to reduce the probability of having a number of patients different from $\bar{n}$. To control the number of patients in the department, we can not force the discharge rate because it strictly depends on patients health status and on patients number. On

the other hand, physicians can decide the number of patients that could be admitted. In order to find the optimal policy, we start with

$$J_N(n_N) = g_N(n_N) = 0, \tag{2.23}$$

and we go backward, by applying the one-step optimization procedure:

$$\begin{aligned} &\text{for} \quad k = N-1, \dots, 0, \\ &J_k(n_k) = \min_{e_k \in \Theta_k(n_k)} \mathbb{E}_k \Big[g_k(n_k) + J_{k+1}\Big(g_{k+1}(n_{k+1}) \Big) \Big] \end{aligned} \tag{2.24}$$

where $\mathbb{E}_k$ is the expected value conditional to n_k. $J_0(n_0)$, generated at the last step, is equal to the optimal cost $J^*(n_0)$ and also the collection of the policies that minimize the functional in Equation 2.24, $\pi^* = \{\mu_0^*(n_k), ..., \mu_{N-1}^*(n_k)\}$ represents the optimal policy for the whole minimization problem.

6.2.4 Estimation result

In order to validate our approach we evaluate the stochastic dynamic algorithm in different cases. Preliminary we adopt the binomial model fixing the same number of patients for all days of the week. Afterwards a Poisson model is adopted, changing the required number of patients as a function of the day of the week. We set the α parameters of the cost function 2.12 equal to 2. Stochastic dynamic programming results are compared with the current number of beds managed in practice. Medical practice commonly applied in the department can be described by a linear model. Therefore as a proxy of the empirical policy, we consider a policy estimated from the average number of discharges (LP) according to the following relation:

$$\tilde{e}_k = \Big[\bar{n}_k - n_k + \mathbb{E}[u_{k+1} | n_k] \Big]^+ \tag{2.25}$$

where $\bar{n}_k$ is the target number of hospitalizations and $\mathbb{E}[u_{k+1}|n_k]$ is the conditional expectation of the discharges number given n_k.

We evaluate the stochastic dynamic programming for different values of $\bar{n}$ and we consider 10^4 simulated scenarios of 28 days (4 weeks) in which the number of discharges are simulated according to the models proposed. In the case in which the number of the required patients changes during the day of the week the value of $\bar{n}$ represents the weekly average number of patients. The daily number of patients is given by

$$\bar{n}_k = \begin{cases} \lfloor \bar{n} \times w_k \rfloor, & \text{if} \quad \{\bar{n} \times w_k\} \leq 0.5 \\ \lfloor \bar{n} \times w_k \rfloor + 1, & \text{if} \quad \{\bar{n} \times w_k\} > 0.5 \end{cases} \tag{2.26}$$

where $w = [1.04 \quad 1.06 \quad 1.10 \quad 1.06 \quad 1.09 \quad 0.87 \quad 0.77]$. Vector w is estimated from historical data. Otherwise if we consider a fixed value during the days of the week $\bar{n}_k = \bar{n}$. For each scenario the absolute distance between the required number of patients $\bar{n}$ and the hospitalized number (n_k) is evaluated in the following way:

$$\epsilon = \sum_{k=1}^{28} |n_k - \bar{n}_k|. \tag{2.27}$$

In Tables 6.18 and 6.19 the average absolute distance for the stochastic dynamic programming method (SDP) and for the empirical policy are reported. Values show that, despite the proposed model is relatively simple, it allows to obtain better results than using a linear policy. Using the cost function proposed a better management of hospital bed is obtained. Figure 6.16 shows an example of two scenario in which red line

	$\bar{n}$			
	4	**5**	**6**	**7**
SDP	23.80	23.31	26.49	29.83
LP	20.68	23.26	26.03	28.40
	$\bar{n}$			
	8	**9**	**10**	**11**
SDP	33.56	33.48	41.50	41.78
LP	30.10	31.68	33.07	33.90

Table 6.18: Average absolute distance between the number of patients $\bar{n}$ and the hospitalized number n_k. The required number of patients is the same for all the day of the week, discharges number are simulated using the Binomial model.

describes the evolution of the number of patients based on the application of the empirical policy and the blue line describes the number of patients under the optimal policy. The first plot in Figure 6.16 shows a scenario obtained by considering a fixed number patients for all the days, where the Binomial model is applied in order to describe the discharge frequency. The second plot in Figure 6.16 shows a scenario obtained by considering a time varying number of patients over the days of the week and the Poisson model is applied in order to describe the discharge frequency. Both cases consider a weekly average number of patients equal to 5. In Table 6.20 is reported the optimal policy estimated considering a weekly average equal to 4, given the periodicity of the problem in Table 6.20 is reported only the weekly strategy.

	$\bar{n}$			
	4	**5**	**6**	**7**
SDP	32.33	35.23	36.32	41.02
LP	32.22	37.36	37.99	42.18
	8	**9**	**10**	**11**
SDP	42.05	43.77	45.61	53.18
LP	42.67	43.36	43.96	48.36

Table 6.19: Average absolute distance between the number of patients $\bar{n}$ and the hospitalized number n_k. The required number of patients changes daily during the day of the week, discharges number are simulated using the Poisson model.

6.3 Summary

The conceptual framework proposed allows to analyse different aspects of public health system performance.
Distinctive quality indicators and their representation are proposed in order to monitoring the health system. The reduction of CABG interventions (passing from 48% to 40%), the increase of mechanical support, and the decrement of the average LoS for CABG intervention show structural evolution of the department.
The average age of patients treated and the decreasing mortality rate for each type of intervention confirms the overall structural improvement of the department.
The control chart is an instrument for describing, in a precise manner, exactly what is meant by statistical control. In many applications, it is used for on-line process surveillance. The use of two different Control Charts is proposed. This tool puts in evidence, that after a starting period in which the team is settled, also in this case, it is observed an improvement of the performance due to the standardization of the procedure. The charts show a time trend that could be explained by the evolution of the department. First year data highlights that the process was out of control and analysis confirm that the process improves in the following years. The tools allowed a reasoned characterization of the system.
A linear relation between LoS and ES is investigated and it is proven that for small value of ES the mortality is better than the expected one. Moreover LoS analysis suggest that the optimal number of hospitalized patient is 5.

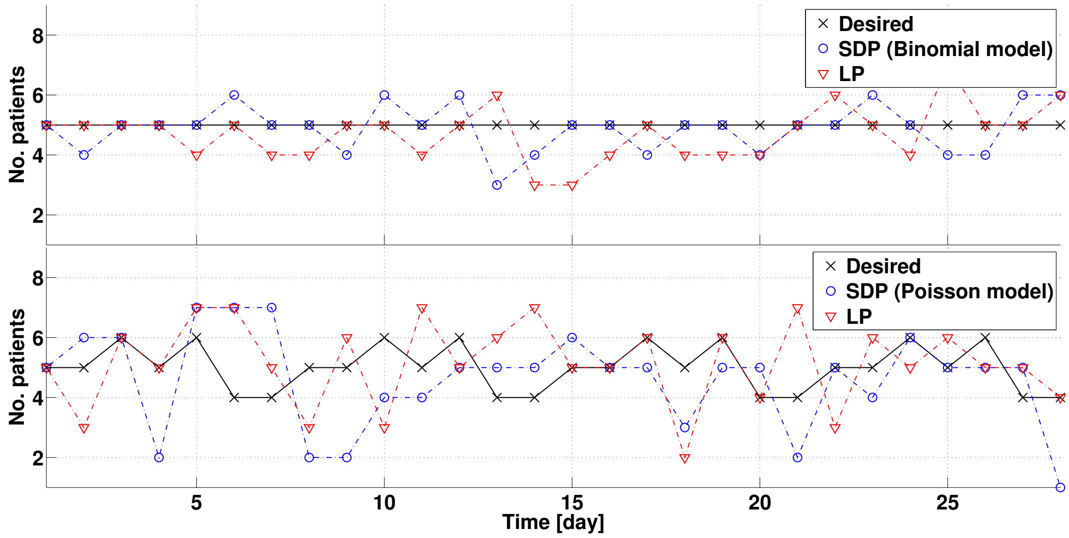


Figure 6.16: Admission policies in San Camillo ICU - Comparison between the SDP and LP

Hospital management takes place in an increasingly competitive environment and it is therefore essential to focus on delivering high quality care to patients. The aim of the hospital bed management is to allocate beds to patients while taking into account capacity constraints. Therefore is proposed SDP approach to describe the healthcare resource allocation. In order to evaluate the optimal policy we propose different model to describe the discharge frequency of the department. According to the analysis of the empirical data the Binomial model and the Poisson model are proposed. In particular we describe the conditional distribution of the empirical discharge frequency given:

- the number of hospitalized patients
- the synthetic proxy that describe the health status of the hospitalized patients
- the number of hospitalized patients and the day of the week

Results show that the models proposed provide a good description on the empirical data. SDP methodology allows the estimation of the optimal policy in order to control the average number of patient hospitalized in the department. In this way the hospital manager could balance the cost of empty beds against the cost of turning patients away, thus facilitating a good choice of bed provision in order to have a low cost and high access to service. Results of the empirical policy compared with the results of the

n_{k-1}	Mon	Tue	Wed	Thu	Fri	Sat	Sun
1	4	4	5	5	4	4	3
2	3	3	4	4	3	3	2
3	2	2	3	3	2	2	1
4	1	2	2	2	2	1	0
5	0	1	1	1	1	0	0
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0

Table 6.20: SDP policy obtained considering a variable number of patient during the day of the week. Average number of required patient is equal to 4 and discharge distribution is described using a Poisson distribution.

optimal policy show that the proposed model contribute to describe the department management strategy. Moreover the SDP approach provides a methodological tool that allows to introduce further improvements. In future work the impact of emergency demand could be investigated in order to understand which is the best strategy to adopt to face these situations. An evolution during the years of the optimal strategy could be also analysed in order to find out the impact of the new technology and procedures.

7

Financial modelling in healthcare optimization

This Chapter introduces three financial-like models to describe the variable costs associated with patient hospitalization: a Nelson-Siegel model; a Black-Scholes model and a Cox-Ingersoll-Ross model.

In modern societies, the costs of healthcare are increasing year by year. One of the reasons for this growth is the fact that the population is ageing. Elderly patients (over 70 years of age) have a high incidence of chronic diseases such as hypertension, diabetes and chronic lung disease. Another factor contributing to increasing costs is the use of expensive technology such as coronary artery bypass graft (CABG) surgery. The third party (state, communities, insurance companies and health maintenance organizations) has taken an active part in controlling hospital costs. The requirement is to cut costs without diminishing the quality of care. One solution is to increase efficiency; hospitals need to plan their operations to use available resources in an optimal fashion.

An approach is to study the relation between admission risk score, Length of Stay (LoS), and hospitalization costs. Risk score can be used to predict the hospital LoS and total hospitalization costs [141], [158], [159]. It have been proved that ICU costs are grater on the first ICU day and by day 2 and become stable after day 3 [159]. It also been shown that exist a relation between risk score and total cost [158], [141]. According to [141] in Table 7.1 are reported the mean values of cost at different level of risk score and LoS. When a patient is admitted the expected cost and LoS depends on the risk score observed. Data are reported as an example of relation between risk score,

Risk score	LoS	Empirical Cost [C]	Estimated Cost [$\hat{C}$]
0	8.3	7856	7500
1	8.9	8031	8154
2	9.7	9036	8968
3	10.3	9336	9663
5	10.0	10205	10220
11	11.3	14995	14943

Table 7.1: Mean values of cost for different level risk scores and LoS.

LoS and cost; it is assumed that this relation can be observed also in other department. EuroSCORE (ES) is a method of calculating predicted operative mortality for patients undergoing cardiac surgery, it is assumed that EuroSCORE or its linear transformation are reliable and valid proxy of risk score.
The analysis is based on the Unit of Anaesthesiology and Reanimation of S. Camillo-Forlanini Hospital data, a part of the Cardiac Sciences Department. The data analyzed are referred to the patients receiving cardiac surgery. Particularly, the database considers 11,770 patients (records) who received cardiac surgery from 1999 to 2012 and the corresponding items were:

- Name, Intervention Date, Age, Pathology, Sex, Urgency, Intervention, Ejection Fraction, Pulmonary Hypertension, Diabetes, Obesity, Chronic Obstructive Bronchopathy, Chronic Kidney Failure, Liver Failure, Neuropathy, Reintervention, Extracorporeal Circulation Time, Orotracheal Intubation Time, Length of Stay in Intensive Care Unit, EuroSCORE, Post-Operative Complications, Exitus (dead patient).

- Two main indicators were considered: 1) ICU Length of Stay (LoS) and 2) European System for Cardiac Operative Risk Evaluation-ES (EuroSCORE). The term Length of Stay indicates the measure of the duration (in days) of a stay in a hospital department. The relevance of LoS as performance indicator of an ICU is reported elsewhere (Clemente et al. 2014) as it measures the efficiency and the quality of the process. As described in (Millard, 1994), the use of the LoS in analyzing ICU processes is twofold: it is the measure of the efficiency of the process, and it is the measure of costs of the ICU and of the global hospital. In fact, the ICU while using between 5% and 10% of the beds can consume up to

20% of hospital budgets. Moreover, the LoS can be used as an indirect indicator of the quality of care. Figure 6.5 shows the average LoS per year, in general and for different types of patients. The reference solid line is the average of all admitted patients (aggregated from 1999 to 2012) and the dashed line shows the trend of average LoS for different type of patients. The EuroSCORE is a risk indicator for patients undergoing cardiac surgery, which was widely used since it was introduced in 1999. It represents one of the main quality measurement that allows an objective assessment of the service under examination, predictive assessment of the safety, efficacy and quality of health (see Handler at al. 2001). EuroSCORE gives to each clinical risk factor a weight or a number of points that, when combined, provide an estimate of the expected percentage mortality for a patient undergoing cardiac surgery. Even if different EuroSCORE models have been proposed, the additive EuroSCORE underestimates the risk in some groups of patients. At the same time, there has been an exponential growth in the availability of information technologies for the cardiac and surgical units of hospitals, thats why the use of a risk model based on the equation complete logistics gradually established. Analysis faced into this paper concerned only with logistic EuroSCORE (ES in the following), approximating the values to integer numbers. Figure 6.7 shows the number of patients per every specific logistic EuroSCORE value: the majority of patients (about 75% of the whole population) falls in the first six ranges of ES).

7.1 Cost function

Starting from data of Table 7.1 it is provided a methodological approach that allow to describe the cost function with respect to the ES and the LoS. The aim is to introduce a cost function that describe the empirical data and furnish a phenomenological interpretation. In according to Nelson-Siegel model [160], used to estimate the term structure of interest rates from observable data, a method to interpolate healthcare costs is provided. Nelson-Siegel model is a function-based approach that approximates empirical data to create yield curves. The model of Nelson and Siegel and its extension by Svensson [161] are used by central banks and other market participants as a model for the term structure of interest rates.

In order to interpolate healthcare data and provide a cost surface, starting from the couples of data observed, a two variables function is calibrated. It is assumed that the overall hospitalization costs mainly depend by three factors:

- **Admission health state** it is evaluated through different risk model. For example several risk indexes have been developed for the prediction of postoperative mortality and morbidity in coronary bypass surgery. EuroSCORE is one of the most widely used in Europe [140], [162]. It is calculated by adding the points assigned to several variables. A minimum value indicates the absence of risk variables, and therefore should correspond to minimum mortality. The risk indexes can only give a rough estimate of the risk for an individual patient, they can be used for planning purposes at the population level. The results of [141] and [158] demonstrate that there is a close relationship between the preoperative risk scores and total cost. Admission health state costs should be described using a function that increases with EuroSCORE and decreases with time, severe patients have higher costs and the care provided by hospital reduce the impact of initial health state costs.

- **Fixed costs** represents all the fix expediencies arising from the standard department resource consumptions. This cost compressive elements such us bed occupancy, nurses, mechanical ventilation. It is assumed that this costs are fixed and do not depend on time and health state.

- **Clinical complications** represents all the cost arising from the extra activities that must be provided in order to face complication during the hospitalization of the patients. Longer hospitalizations have higher probabilities of complication, however complications are more likely during the first day of hospitalization. Hence it is assumed that the complication costs are described by a decreasing function of the time.

Let X and Y be exponential independent random variables of parameters $\beta_3 > 0$ and $\beta_4 > 0$ defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Specifically X describes the probability that the admission health state affects the cost and Y describes the probability that during the hospitalization a complication is found. During the hospitalization the treatments received by the patients reduce the probability that the admission health state is reflected on cost and the probability that complication arise. In order to justify the use of an exponential distribution, for the variable X we consider the following assumptions:

1. The health state impact on cost is described by a Poisson process. Hence X is the time between events.

2. The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently.

3. The rate at which events occur is constant. The rate cannot be higher in some intervals and lower in other intervals.

4. Two events cannot occur at exactly the same time.

Regarding assumption 1 is reasonable to assume that the health state cost impact is described by a Poisson process; patients with the same admission health state can have different cost give by the *jump* of their health evolution. Regarding assumption 4, it is unlikely that two events that change the health state occur at the same time. Assumption 2 and 3 may be unrealistic but they are needed in order to simplify the mathematical framework of the problem. The same assumptions can be made for variable Y.

The cumulative hospitalization cost up to time t, considering an initial EuroSCORE of ES, can be expressed by

$$\hat{c}(ES, t) = \zeta[\beta_1 t + \beta_2 ES \mathbb{1}_{X \le t} + \mathbb{1}_{Y \le t}] \tag{1.1}$$

where $\mathbb{1}$ stands for the indicator function, $\beta_1, \beta_2 \geq 0$, and $\zeta = 10,000$ is a normalizing factor chose in according to value of Table 7.1. Time is expressed in days. Evaluating the expected value the cost for a hospitalization up to time t is

$$\hat{C}(ES, t) = \mathbb{E}[\hat{c}(ES, t)] = \beta_1 LoS + \beta_2 ES(1 - e^{-\beta_3 t}) + (1 - e^{-\beta_4 t}). \tag{1.2}$$

Considering the derivative with respect to the time, we obtain the daily expected cost of hospitalization is

$$\frac{\partial \hat{C}(ES, t)}{\partial t} = \underbrace{\beta_1}_{\text{fixed cost}} + \underbrace{\beta_2 \beta_3 ES \times e^{-\beta_3 t} + \beta_4 e^{-\beta_4 t}}_{\text{variable cost}} \tag{1.3}$$

where β_1 is the fix costs, $\beta_2\beta_3 ES \times e^{-\beta_3 t}$ is the variable cost depending admission ES, and $\beta_4 e^{-\beta_4 t}$ is the variable cost arising from medical complication that can occur during the hospitalization.

Starting from observed data, model parameters are estimated through the Minimum Mean Squared Error method (MMSE). Last column of Table 7.1 shows the cost estimated and in Table 7.2 are reported the parameters of the model. In order to evaluate

$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$	$\boldsymbol{\beta_4}$
10.0014	21.3399	0.0463	32.5151

Table 7.2: Surface cost function estimated parameters.

the performance of this approach, four discrepancy measures are applied: the average prediction error (APE), the average absolute error (AAE), the root mean-square error (RMSE), and the average relative prediction error (ARPE). Let $C(ES_i, LoS_i)$ be the observed data of Table 7.1 and $\hat{C}(ES_i, LoS_i)$ be the estimated data, the four error estimators are defined as follows:

$$APE = \sum_{i=1}^{N} \frac{|C(ES_i, LoS_i) - \hat{C}(ES_i, LoS_i)|}{C(ES_i, LoS_i)}, \tag{1.4}$$

$$AAE = \sum_{i=1}^{N} \frac{|C(ES_i, LoS_i) - \hat{C}(ES_i, LoS_i)|}{N}, \tag{1.5}$$

$$ARPE = \frac{1}{N} \sum_{i=1}^{N} \frac{|C(ES_i, LoS_i) - \hat{C}(ES_i, LoS_i)|}{C(ES_i, LoS_i)}, \tag{1.6}$$

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(C(ES_i, LoS_i) - \hat{C}(ES_i, LoS_i))^2}{N}}. \tag{1.7}$$

In Table 7.3 are reported the discrepancy measures of the cost function fitting. Values

APE	AAE	ARPE	RMSE
0.11	156.83	0.02	206.64

Table 7.3: Error metrics for cost function fitting. Values are estimated considering data of Table 7.1.

in Table 7.1 and 7.3 show that the proposed cost model provides a good description of empirical data. Starting from the observed data, the proposed approach provides an estimation on the cost for every couple of ES and LoS. Figure 7.1 is estimated a cost surface that shows the increasing of cost as a function of LoS and EuroSCORE.

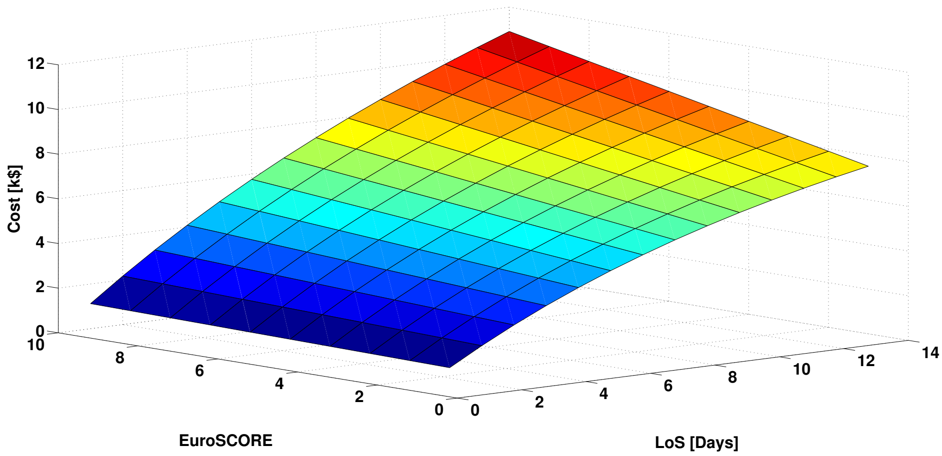


Figure 7.1: Hospitalization costs estimated surface of San Camillo ICU department. - Relation between ES, LoS and costs.

7.2 Black-Scholes approach

Real options analysis is a mathematical approach that calculates the value of options associated with a decision, it determines optimal investment scope and timing taking future decisions and flexibility into account. It originated from options theory, which determines the value of financial options that give option holders the right to buy or sell stocks at a previously set price. The beginning of options theory starts with the works of [57] and [58] on the pricing of financial options and the development of closed-form solutions for the value of call and put options. Alternatives to closed form solutions are partial differential equation models, simulations, or portfolio optimization techniques [59]. The holder of a call option has the right, but not the obligation, to buy the underlying asset of the option often a stock within a specified period at a given price, called the strike price. Similarly, a put option allows the selling of the underlying asset at the strike price [60]. The parallels between financial options and real options were first discussed by [61]. Early literature on real options focused on determining the value of one specific type of option at a time, such as the option to delay or modify the operating scale of a project [62], [63]. Later research focused on using these options to model more complex settings [65], [66]. In this framework the aim is to fill this gap by providing a mathematical and accessible model to measure the cost of hospitalization jointly with a commonly observed health risk indicator, the EuroSCORE. Hence, it is given also a contribution the real option technique in healthcare. Through a case study, it is provided clinicians and executives with insights on the advantages of integrating effectively this quantitative modelling technique into the organizational decision process. It is also provided details for hospital analysts concerning with the data required as input, and the information generated as output.

A key responsibility of hospital executives is to make important decisions concerning budgeting, investment and resource allocation. Hence it is described the hospitalization flexibility as a call option. It is assumed that during the patient hospitalization physicians have to choose between two different opportunities; continue to treat the patient or discharge him. Physicians decide to continue the treatment if it considers that the benefits exceeded the hospitalization cost. The benefits are greater the more severe the patients state of health is. The hospitalization value is given by the likely benefits that the patients health state can obtain, less the cost associated with the consumption of the bed place. It is assumed that the bed cost is not fixed and depends on the department situation. An overused department has higher management costs and has also a higher probability of refuse a new patient.

Calling S the value associated to the investment in patient care and K the bed cost, the patient stays in hospital since $S > K$ otherwise the patient is discharged. The overall value of the hospitalization is

$$(S-K)^+ = \begin{cases} S-K & \text{if} \quad S \geq K \\ 0 & \text{if} \quad S < K \end{cases} \tag{2.8}$$

where Equation 2.8 represents the payoff of an European Call option. If S is less than K, the patient is discharged because investing in his care is no longer profitable: the amount of treatment administered have brought an improvement of the health and further action would no bring additional benefits. The patient hospitalization could be represented as an option in which the physician has the right, but not the obligation, to abandon the investment in the patients care. The underlying value of this option is a proxy of the health state and the strike price is the cost of the bed. Since that every day physicians evaluate the patient health state, and decide to discharge or not, it is assumed that the value of a multi-day stay can be described by adding the single day option value.

7.2.1 Parameters estimation

The patient hospitalization cost is described using an option approach under the following assumptions:

1. Let LoS and ES be random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. LoS describes the hospitalization time and ES the health state of the patent. Then exists a function $g : \mathbb{R}^+ \to \mathbb{R}^+$ such that, the underlying $S = g(ES|\Lambda)$ is log-normally distributed, where S describes the value associated to the patient hospitalization (bigger values of S entail a worse health state and thus further benefits that can be obtained in case of treatment) and Λ is the function parameters.

2. The strike K depends on the number of patient hospitalized according the following formula

$$K(t) = \gamma_1 n(t) \tag{2.9}$$

   where $\gamma_1 > 0$ is a parameter and $n(t)$ is the number of patient hospitalized at time t.

According to Chapter 6, Section 6.2.1, Equation 2.4, exist a linear relation between LoS and ES. Given that S is a proxy of the health state, all patient at discharge have the same small value of S, hence considering an initial EuroSCORE of ES, after t days of hospitalization the expected ES is

$$ES(t) = \max(ES + \mu t, 1). \tag{2.10}$$

where $\mu < 0$ is an estimated parameter and the max is applied in order to avoid negative value for ES. To estimate parameter μ the following objective function is considered

$$\min_{\mu} \sum_{i=1}^{N} \Big[ES_D - \max(ES_i + \mu LoS_i, 1)\Big]^2. \tag{2.11}$$

where N is the number of patient considered, ES_i is the admission ES of patient i, and LoS_i is the time spent in ward by patient i. According the previous assumption the hospitalization cost can be estimated by the Black-Scholes formula (see Chapter 4 Section 4.3, [57])

$$\hat{C}_{BS}(ES, LoS|\Theta) = \gamma_2 \sum_{t=1}^{LoS} \Pi(g(\mu t + ES), \gamma_1 n(t), r, \sigma, \frac{1}{365}) \tag{2.12}$$

where Π is the Black-Scholes price of a Call option, $\gamma_2 > 0$ is a estimated scale parameter, r is the risk free rate, and σ is the volatility. In this framework it is assumed that $g(ES, t|\Lambda) = \gamma_3 + \gamma_4 ES(t)$.

Starting from the cost surface estimated, a calibration procedure is performed in order to estimate the parameters of the pricing model. The following objective function is considered:

$$\min_{\Theta} \sum_{i=1}^{N_{ES}} \sum_{k=1}^{LoS} \Big[\hat{C}(ES_i, k) - \hat{C}_{BS}(ES_i, k|\Theta)\Big]^2 \tag{2.13}$$

where N_{ES} is the number of ES and $\Theta = [\gamma_1, \gamma_2, \gamma_3, \gamma_4, \mu, r, \sigma]$. Parameters calibration starts from a Halton quasi-random sequence of 20×7 initial points [98], [99]. Grid is generated according the following constraints:

$$\begin{cases} \gamma_i \geq 0 \quad \forall i \in [1, 4] \\ r, \sigma > \; 0 \\ \mu < 0 \end{cases} \tag{2.14}$$

In Table 7.4 are reported the estimated parameters for the Black-Scholes model. In

$\boldsymbol{\gamma_1}$	$\boldsymbol{\gamma_2}$	$\boldsymbol{\gamma_3}$	$\boldsymbol{\gamma_4}$
2.50	307.45	0.21	3.33
$\boldsymbol{\mu}$	$\boldsymbol{r}$	$\boldsymbol{\sigma}$	
-0.78	0.50	0.05	

Table 7.4: Black-Scholes cost function estimated parameters.

order to validate the performance of the model, the following metrics are evaluated

$$APE = \sum_{i=1}^{N_{ES}} \sum_{k=1}^{LoS} \frac{|\hat{C}(ES_i, k) - \hat{C}_{BS}(ES_i, k)|}{\hat{C}(ES_i, k)}, \tag{2.15}$$

$$AAE = \sum_{i=1}^{N_{ES}} \sum_{k=1}^{LoS} \frac{|\hat{C}(ES_i, k) - \hat{C}_{BS}(ES_i, k)|}{N_{ES} \times LoS}, \tag{2.16}$$

$$ARPE = \frac{1}{N_{ES} \times LoS} \sum_{i=1}^{N_{ES}} \sum_{k=1}^{LoS} \frac{|\hat{C}(ES_i, k) - \hat{C}_{BS}(ES_i, k)|}{\hat{C}(ES_i, k)}, \tag{2.17}$$

$$RMSE = \sqrt{\sum_{i=1}^{N_{ES}} \sum_{k=1}^{LoS} \frac{(\hat{C}(ES_i, k) - \hat{C}_{BS}(ES_i, k))^2}{N_{ES} \times LoS}}. \qquad (2.18)$$

In Table 7.5 are reported the error metrics, data show that the BS approach pro-

APE	AAE	ARPE	RMSE
3.03	152.47	0.04	239.61

Table 7.5: Error metrics of Black-Scholes approach. Cost function surface fitting.

vides a good description of empirical costs; however Figure 7.2 points out the model performance. The model provides a better estimation of the data for shortest hospitalizations, for longer stays the average cost is estimated well but the model shows an overestimation for less sick patients and an underestimation for sicker patients. Overall the model performances are adequate to analyse the hospitalization costs.

The proposed approach describes data and provides phenomenological interpretation; through the use of the BS formula it is evaluated the optionality linked to the patients hospitalization. The aim is to highlight the components that most influence the hospitalization costs. Key elements turn out to be the patients health state and the number of hospitalized patients. Hospital departments are systems characterized by a finite number of resources, hence allocation policy should maximize the expected return of each investment. Therefore recovering patients in a nearly full department appears to be an investment with low expected value. This methodological approach introduces a new economic evaluation method and allows comparison between different investments in order to provide a decision support tool.

7.3 Cox-Ingersoll-Ross approach

Providing the appropriate medical care involves decision-making in terms of planning and management of healthcare resources. The requirement is to cut costs without diminishing the quality of care. One solution is to increase efficiency; hospital need to plan their operations to use available resources in optimal fashion. For the past 40 years, practitioners and researchers alike have been grappling with the natural shortcomings associated with the net present value approach to strategic decision making and capital budgeting. Work by scholars in option pricing theory has evolved into an

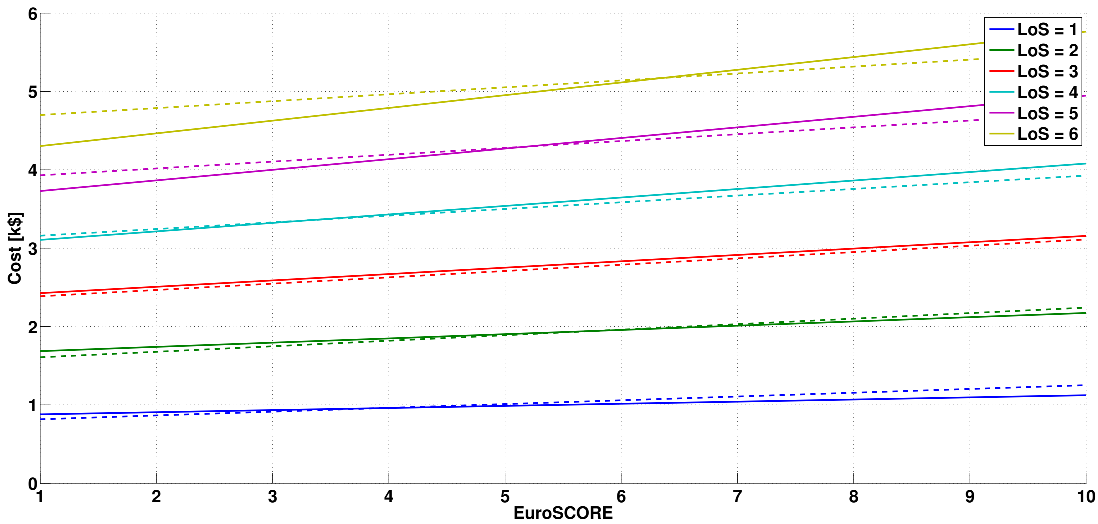


Figure 7.2: Comparison between real hospitalization value and Black-Scholes estimation. - The continuous lines represent the real costs and the dashed lines represent the estimated costs.

alternative perspective on strategic capital investments, called *real options*. Proponents of real options argue that this is a superior way of approaching decision making and capital budgeting, compared with other approaches, as it allows for greater strategic flexibility and encourages exploration, experimentation, and innovation. Within the healthcare literature, articles on real options have been focused on pricing these options [60]. Moreover, the real option analysis also has been used to evaluate vaccination programs. Standard analyses might not capture the full economic value of novel vaccination programs because the cost-effectiveness paradigm fails to take into account the value of active management. Management decisions can be seen as real options, a term used to refer to the application of option pricing theory to the valuation of investments in non-financial assets in which much of the value is attributable to flexibility and learning over time [163]. The uncertainties at the hospital level should be explicitly addressed and accounted for in executive decision-making processes. Furthermore, past decisions affect the range of decisions that can be taken in the future. Therefore options give executives flexibility to respond to future events and manage costs. The purpose of our study is to explore the interplay between two crucial drivers used to evaluate the clinical process, the Length of Stay and the European System for Cardiac Operative Risk Evaluation (EuroSCORE).

7.3.1 Model Specification

In order to find the fair cost of patient hospitalization (in an intensive care department), we analyze the relation between the EuroSCORE and the LoS. We introduce an equilibrium model in order to derive the value for the *premium* that a *representative patient* pays to have the opportunity to be discharged after a finite time. Standard models can not capture all the characteristics of a complex quantity as the disease state of a patient. Moreover the costs associated with the hospitalization of a patient do not depend linearly on the state of the disease, therefore is not simple, starting from a synthetic index as the EuroSCORE, to derive an estimate of the whole cost of the hospitalization. Our main assumption is that it is possible to properly describe the hospitalization of a patients through his disease state behaviour. Moreover patients disease state can be improved spending resources such as doctor working hours, hospital beds, and other. To explain our approach, preliminary we describe a simplified scheme where there are three economic subjects: a representative patient (P), the hospital department (H) that takes care of P, and the healthcare system (HS). The evolution of patient health is the source of uncertainty in the model. Precisely, the health state refers to the disease state of the representative patient. When a patient is admitted his initial disease state is the EuroSCORE and after a given time it assumes two different values. At different values of the health state correspond different values for the LoS. Actually, the LoS of a patient strictly depends on the disease state, the higher is the disease state the bigger is the time the patient will spend in the hospital. Let us consider the following **variables**:

- $h > 0$ is the disease state of P at the time of hospital admission, that may increase to $h_u > h$, with probability $p \in (0, 1)$, in the case the patient develops a serious illness, causing a drop in the health state, which causes the death of the patient (called the *exitus*). Otherwise it may decrease to $0 < h_d < h$, with probability $1 - p$, when the patient is discharged and the health state is improved.

- $K > 0$ represents a threshold for the health state. Until the disease state is greater than K, the patient is considered sufficiently ill to increase the LoS and hence the hospitalization cost. Clinicians do not discharge the patient if the health (read *disease*) state is higher than K.

- The economic position of patient P, at the time of hospital discharge, can be identified through the increasing function $g : \mathbb{R} \to [0, \infty)$ of $h - K$, the difference between the health state and the clinical threshold.
- The unit cost value for a single healthcare intervention is affected by the health state of P. Precisely, it is an increasing function of h, $Q = Q(h)$, where $Q : [0, \infty) \to [0, \infty)$.
- The amount of resources invested (at the time of hospital admission) by H to reduce the disease state of P is proportional to the unit cost value, namely: $\Delta \cdot Q$, for some$\Delta > 0$.
- $B > 0$ denotes the amount of money delivered by HS to H for the hospitalization cost coverage of P, paid after the hospital discharge.
- $r > 0$ is a compounded interest rate per-period of time.

The initial poor health condition leads to the hospitalization of P, whereas he will be discharged only if the disease state goes below K. Therefore it is reasonable to suppose that $h_d < K < h$. This simplified world works under the following **actions**:

- P pays a *premium* Π (a part of his income taxes) to HS for the right to receive the health treatment by H and to own the chance to be discharged after the time of hospitalization (LoS).
- The hospital management has the obligation to allocate resources of H to improve the health state of P.
- HS has the opportunity to invest the amount of money that is not delivered to H in alternative financial assets earning the interest rate r.

At the end of the hospitalization, the financial position of P and HS, jointly with H, are given by

$$P : \quad \begin{cases} g(h_u - K) & \text{with probability } p \\ g(h_d - K) & \text{with probability } 1 - p \end{cases} \tag{3.19}$$

$$HS \text{ and } H : \begin{cases} -\Delta \cdot Q(h_u) + B & \text{with probability } p \\ -\Delta \cdot Q(h_d) + B & \text{with probability } 1 - p \end{cases} \tag{3.20}$$

HS and H can choose Δ and B in order to match the financial position of P:

$$\begin{cases} -\Delta \cdot Q(h_u) + B = g(h_u - K) \\ -\Delta \cdot Q(h_d) + B = g(h_d - K). \end{cases} \tag{3.21}$$

These equations yield the number of care interventions:

$$\Delta = \frac{g(h_u - K) - g(h_d - K)}{Q(h_d) - Q(h_u)}. \tag{3.22}$$

By assuming the perspective of a rational investor for all three agents [164] and [165], the system is in equilibrium only if the premium paid by P corresponds to the initial joint position of HS and H:

$$\Pi = -\Delta \cdot Q(h) + \frac{B}{1+r} = -\Delta \cdot \left[Q(h) - \frac{Q(h_d)}{1+r} \right] + \frac{g(h_d - K)}{1+r}. \tag{3.23}$$

which can also be written as

$$\Pi = \frac{1}{1+r} \cdot \left[p^\star \cdot g(h_u - K) + (1 - p^\star) \cdot g(h_d - K, 0) \right], \tag{3.24}$$

where

$$p^\star = \frac{Q(h)(1+r) - Q(h_d)}{Q(h_u) - Q(h_d)}. \tag{3.25}$$

Note that if $Q(h)(1+r) < Q(h_u)$, then $p^\star \in (0, 1)$. Hence 3.24 represents the discounted expected value of the random payoff of the patient P, that is

$$\Pi = \frac{1}{1+r} \mathbb{E}^\star \left[g(h_T - K) \right], \tag{3.26}$$

where $\mathbb{E}^\star[\cdot]$ is the expected value of the random variable $g(h_T - K)$, where h_T is the health state at the time of discharge T, with two possible outcomes, $g(h_u - K)$ with probability $p^\star$, and $g(h_d - K)$ with probability $1 - p^\star$. We remark that, if the condition on the unit cost function $Q(\cdot)$ is not satisfied, then the cost of a care is not adequate, from an economic viewpoint, to the effective health state.
From a financial perspective, Equation 3.26 represents the price of a European-style contingent claim (*option*) written on the health state variable h_T and it can be also interpreted as the fair value of the premium paid by the representative patient P for his hospitalization cost coverage. In next sections we extend model 3.26 to a continuous-time framework and we propose an estimation method based on a panel of real data.

7.3.2 State Transition Equations

The evolution of health (h) for the representative patient is described by the state of disease over time $t \geq 0$ and it is the main source of uncertainty in our model. This is described formally through a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a filtration $\{\mathcal{F}_t : \ t \geq 0\}$ of sub-σ-algebras of $\mathcal{F}$ satisfying the usual conditions [165]. We consider a one-dimensional Brownian motion $\{W_t\}_{t \geq 0}$ adapted to $\{\mathcal{F}_t : \ t \geq 0\}$ which describes the underlying source of randomness. The health state is a predictable stochastic process, which is, intuitively speaking, a process whose value at any time t depends only on the information in the underlying filtration $\{\mathcal{F}_t : \ t \geq 0\}$ that is available up to, but not including, time t. Specifically, the following stochastic differential equation governs the behaviour of the process:

$$dh_t = a(h_t)dt + b(h_t)dW_t, \qquad h_t = h_0, \tag{3.27}$$

where $h_0 > 0$ is a constant initial parameter. Here the coefficient function a, b : $[0, \infty) \to \mathbb{R}$ are assumed to be continuous, b is strictly positive on $(0, \infty)$, and satisfy sufficient conditions in order to ensure the existence of a unique strong solution for 3.27 (see Chapter 4, Theorem 4.1.7) $\{h_t, \ ; t \geq 0\}$, for any initial datum h_0, such that $h_t > 0$, $\mathbb{P}$ almost surely, for any $t \geq 0$. These conditions are, for instance, well studied in a paper of [166]. We observe that the second term in Equation 3.27 captures the continuous volatility of Brownian motion, the magnitude of the variance being determined by $b(h_t)$. We shall use the EuroSCORE assigned by clinicians to the patient as a proxy for the initial health state value h_0.

Let $T > 0$ be the time of discharge of the representative patient. Thus, we assume the existence of a deterministic function $Q : [0, \infty) \times [0, T] \to [0, \infty)$ such that $Q_t := Q(h_t, t)$ describes the unit cost value for a single healthcare intervention. In particular we suppose that $Q \in C^{2,1}([0, \infty) \times [0, T))$, with $\partial_h Q(h, t) > 0$, for all $h \geq 0$, $t \in [0, T)$. This choice of modelling Q_t implies that the unit cost of intervention is affected by the same source of randomness as the disease state h_t. In the following, the valueΠ_t of the financial position at time t is a deterministic function of h_t, namely$\Pi_t :=\Pi(h_t, t)$, for some smooth enough functionΠ [1].

[1] A sufficient condition for Π is that it is continuous on $[0, \infty) \times [0, T)$ with two continuous derivatives with respect to $h \in (0, \infty)$ and with a continuous derivative with respect to the time variable $t \in (0, T)$. This allows for the application of Ito's formula.

Using equilibrium arguments like those illustrated above, in the continuous-time framework, the following equation must hold:

$$-\Delta_t \cdot dQ_t + rB_t dt = d\Pi_t, \tag{3.28}$$

for any $t \in (0, T)$. Here, B_t stands for the time t value of the amount of money set aside by healthcare system in order to cover the future cost of hospitalization, which is capitalized at continuously compounded (constant) interest rate $r > 0$. Moreover we have assumed that the amount of care provided to the patientΔ_t does not change in the small interval of time dt. By applying the well known Ito formula [165] to$\Pi(h_t, t)$, and by the equation $-\Delta_t Q_t + B_t = \Pi_t$ (in economic equilibrium), we argue that 3.28 is satisfied only if and only if Π solves the partial differential equation

$$\frac{\partial \Pi}{\partial t}(h,t) + \mathcal{L}\Pi(h,t) = r\Pi(h,t) \tag{3.29}$$

where $\mathcal{L}$ stands for the linear differential operator

$$\mathcal{L} = \frac{1}{2} b^2(h) \frac{\partial^2}{\partial h^2} + \mu(h,t)\frac{\partial}{\partial h}, \tag{3.30}$$

$$\mu(h,t) := r\frac{Q(h,t)}{\partial_h Q(h,t)} - \frac{1}{2} b^2(h) \frac{\partial_h^2 Q(h,t)}{\partial_h Q(h,t)} - \frac{\partial_t Q(h,t)}{\partial_h Q(h,t)}, \tag{3.31}$$

for all $h \in (0, \infty)$, $t \in (0, T)$, subject to the terminal condition$\Pi(h, T) = g(h)$, for a given payoff function $g \in C([0, \infty))$. By Equation 3.28, we also deduce that the unit cost value is$\Delta_t = \delta(h_t, t)$, where

$$\delta(h,t) := -\frac{1}{\partial_h Q(h,t)} \frac{\partial \Pi}{\partial h}(h,t), \tag{3.32}$$

for all $h > 0$, $t \in (0, T)$. If 3.29 admits a classical solution, then by the Feynman-Kac formula, this implies the following probabilistic representation for the solution:

$$\Pi(h,t) = e^{-r(T-t)} \mathbb{E}_\star \left[g(h_T) | h_t = h \right]. \tag{3.33}$$

According with a consolidated approach in option pricing theory, the conditional expectation $\mathbb{E}_\star \left[\cdot | \cdot \right]$ is computed under a probability measure $\mathbb{Q}$ equivalent to $\mathbb{P}$. Under such a measure the dynamics of the health state is the following:

$$dh_t = \mu(h_t, t)dt + b(h_t)dW_t^\star, \tag{3.34}$$

where $W_t^\star$ denotes a Brownian motion on the probability space $(\Omega, \mathcal{F}, \mathbb{Q})$, with filtration $\{\mathcal{F}_t : t \geq 0\}$ and it is defined as

$$dW_t^\star = dW_t + \frac{1}{b(h_t)}\left[a(h_t) - \mu(h_t, t)\right] dt. \tag{3.35}$$

We remark that 3.29 is defined on the spatial domain $(0, \infty)$, without the need to impose any condition at $h = 0$. Moreover the equation may degenerate to a first-order partial differential equation at this point, hence a specific existence and uniqueness result is needed. We can establish the following theorem whose proof follows from some reverse conditions introduced in [166].

Theorem 7.3.1. *Let $h \mapsto \mu(h,t) - \frac{1}{2}(b^2)'(h)$ be continuously differentiable on $[0, \infty)$, with bounded derivative, and satisfies*

$$\left[\mu(h.t) - b(h)b'(h)\right]_{|h=0} \geq 0, \tag{3.36}$$

for any $t \in [0, T)$, with $b(0) = 0$. If a and g are Lipschitz continuous, then the final value problem 3.29 has a unique solution $\Pi \in C([0,\infty) \times [0,T)) \cap C^{2,1}((0,\infty) \times (0,T))$.

From Equation 3.33, we argue thatΠ_t corresponds also to the value, at time $t < T$, of a contingent claim (an option) - implicitly owned by the patient - on its health state h, with payoff g. This option is a fair amount, paid by the patient to the healthcare system, which incorporates both the value for the right to obtain the hospital care and the right to be discharged at a given time T. Under a financial perspective [164], the probability measure $\mathbb{Q}$ gives a *risk-neutral* perspective to the relationship between the patient P and HS. This is a concept heavily used in the pricing of financial derivatives due to the fundamental theorem of asset pricing [167]. In our framework, the premium Π depends crucially on the uncertainty of the patient health state h. The model assumes that the healthcare system demands more profit for bearing more uncertainty. Therefore, today's price of patient's claim, which is based on an uncertain health state realised tomorrow, will generally differ from its expected value. The healthcare system is supposed to be risk-averse and today's price is below the expectation. To price the claim, consequently, the calculated expected values need to be adjusted for the risk preferences. In our equilibrium model there is an alternative way to do this calculation: instead of first taking the expectation and then adjusting for the risk preferences, we adjust the probabilities of future outcomes so that they incorporate a risk premium, and then take the expectation under this new probability distribution, exactly the *risk-neutral* measure $\mathbb{Q}$. From the modelling viewpoint, the health state acts as a hidden variable which represents an explanatory factor for the relation between LoS and ES. Precisely we will use the function$\Pi(\ h, t)$ to estimate model parameters on our dataset with the purpose to establish a satisfactory relationship between these to reliable indicators used for measuring the efficiency of the hospital ICU department taken into account.

7.3.3 Specific Disease Model

In order to compare our approach with real data, we consider the following specification for disease state dynamics:

$$dh_t = \kappa(\theta - h_t)dt + \sigma\sqrt{h_t}dW_t, \tag{3.37}$$

with a constant initial value $h_0 > 0$. The drift term is given by $\kappa(\theta - h_t)$, with $\theta,\ \kappa > 0$, whereas the second term captures the continuous volatility of Brownian motion, the variance parameter being $\sigma > 0$. Equation 3.37 defines an elastic random walk around a trend, with a mean reverting characteristic: when h_t goes over (respectively: under θ), the expected variation of h_t becomes negative (respectively: positive) and h_t tends to come back to its average long term level θ at an adjustment speed κ. Equation 3.37 refers to the well known affine model, the so called CIR model [81] which has been extensively studied in the financial literature as a model of the term structure of interest rates. The standard deviation factor, $\sigma\sqrt{h_t}$, avoids the possibility of negative values for health state, for all positive values of θ and κ. An health state of zero is also precluded if the Feller's condition is satisfied, namely

$$2\kappa\theta \geq \sigma^2. \tag{3.38}$$

More generally, when h_t is close to zero, the standard deviation also becomes very small, which dampens the effect of the random shock on the health state. In particular this feature of the model describes an amplitude of fluctuations that is proportional to the disease state. The transition density of h_t, $t > 0$, conditional to h_0, is given by

$$p(h_t|h_0) = e^{-u-v}\left(\frac{u}{v}\right)^{q/2} I_q(2\sqrt{uv}), \tag{3.39}$$

where

$$c = 2\kappa/[\sigma^2(1-\exp(-\kappa t)], \qquad q = \frac{2\kappa\theta}{\sigma^2} - 1, \qquad u = ch_0 e^{-\kappa t}, \qquad v = ch_t,$$

and $I_q(\cdot)$ is the modified Bessel function of the first kind of order q. In our empirical analysis the terminal condition for the pricing function$\Pi(h,t)$ solution to 3.29 is assumed to have the form of a call option payoff:

$$g(h) = \max(h-K,0).$$

The constant parameter $K>0$ represents the threshold described in Section 7.3.1. Moreover, in the following we will use the notation$\Pi(h,t;T,\Theta)$ for the pricing function related to model 3.37.

Thus, we follow an approach often used in affine modelling [168] assuming that, under the measure $\mathbb{Q}$, the health state dynamics 3.34 preserves the same structure as model 3.37. Therefore, we are able to state the following result which allows us to obtain a quasi-closed formula for the numerical computation of the price function.

As an extension of the (standard) Marcum Q-function, originally appeared in [169], we introduce the so called Nuttall Q-function, defined in [170], given by the integral representation

$$\mathcal{Q}_{M,N}(\alpha,\beta) := \int_\beta^\infty x^M \exp\left(-\frac{x^2+\alpha^2}{2}\right) I_N(\alpha x)dx. \tag{3.40}$$

where the order indices are generally reals with values $M \geq 0$, $N > -1$ and α, β are real parameters with $\alpha > 0$, $\beta \geq 0$. Such a special function is involved in the price function representation. From the computation point of view, the Nuttal Q-function can be easily evaluated thanks to available software packages. Moreover, when $M+0.5$ and $N+0.5$ are integers and $M \geq N$ there exists a closed-form series expansion for 3.40, see Corollary 1 in [171].

Theorem 7.3.2. *If $Q(h,t) = e^{rt}\overline{Q}(h,t)$, where $\overline{Q}$ satisfies the partial differential equation*

$$\partial_t\overline{Q} + \frac{1}{2}\sigma^2 h\partial_h^2\overline{Q} + \overline{\kappa}(\overline{\theta}-h)\partial_h\overline{Q} = 0, \tag{3.41}$$

for all $h \in (0,\infty)$, $t \in (0,T)$, for some constant parameters $\overline{\kappa} > 0$, $\overline{\theta} > 0$, satisfying Feller's condition 3.38, then, under the risk-neutral measure the health disease follows

$$dh_t = \overline{\kappa}(\overline{\theta}-h_t)dt + \sigma\sqrt{h_t}dW_t^\star. \tag{3.42}$$

Moreover, let

$$\eta := 2\overline{\kappa}/[\sigma^2(1-\exp(-\overline{\kappa}(T-t))], \qquad \nu := \frac{2\overline{\kappa}\overline{\theta}}{\sigma^2}-1,$$

$$C(h,T-t) := h^{-\nu/2}\exp\left(-\eta h e^{-\overline{\kappa}(T-t)}+\frac{1}{2}\overline{\kappa}(T-t)\nu\right),$$

and

$$A(h,T-t) := \sqrt{2\eta h}\exp\left(-\frac{1}{2}\overline{\kappa}(T-t)\right),$$

then, under model 3.37, the price function takes the following form

$$\Pi(h,t;T,\Theta) = \frac{Ce^{A^2/2-r(T-t)}}{2^{\nu/2}\eta^{2+\nu/2}}\left[\mathcal{Q}_{\nu+3,\nu}(A,\sqrt{2\eta K}) - K\mathcal{Q}_{\nu+1,\nu}(A,\sqrt{2\eta K})\right] \qquad (3.43)$$

Proof. Equation 3.41 easily implies that the drift coefficient in 3.31 is

$$\mu(h,t) = \overline{\kappa}(\overline{\theta}-h), \qquad (3.44)$$

for all $h>0$, $t\in(0,T)$. Hence, by Girsanov's Theorem, under the measure $\mathbb{Q}$, h_t satisfies a CIR dynamics with coefficients $\overline{\kappa}$, $\overline{\theta}$ and σ. From the stochastic representation of the price function 3.33, we can write

$$\begin{aligned}\Pi(h,t;T,\Theta) &= e^{-r(T-t)}\mathbb{E}_\star\left[h_T\mathbf{1}_{h_T>K}|h_t=h\right]\\ &\quad -Ke^{-r(T-t)}\mathbb{Q}\left(h_T>K|h_t=h\right),\end{aligned} \qquad (3.45)$$

where $\mathbf{1}_A$ stands for the indicator function of set A. Thus, from the transition density of h_T, given h_t, under the measure $\mathbb{Q}$, we argue:

$$\mathbb{E}_\star\left[h_T\cdot\mathbf{1}_{h_T>K}|h_t=h\right] = C\int_K^\infty \xi^{1+\nu/2}e^{-\eta\xi}I_\nu(A\sqrt{\xi})d\xi, \qquad (3.46)$$

where, for the sake of simplicity, we have omitted the dependence of C and A on h and $T-t$. Therefore, after applying the change of variable $\xi=\tau^2/(2\eta)$ in the previous integral, and by 3.40, we get

$$\begin{aligned}\mathbb{E}_\star\left[h_T\cdot\mathbf{1}_{h_T>K}|h_t=h\right] &= \frac{C}{2^{\nu/2}\eta^{2+\nu/2}}\int_0^{2\eta K}\tau^{3+\nu}e^{-\tau^2/2}I_\nu(A\tau)d\tau\\ &= \frac{C}{2^{\nu/2}\eta^{2+\nu_2}}e^{A^2/2}\mathcal{Q}_{\nu+3,\nu}(A,\sqrt{2\eta K}).\end{aligned} \qquad (3.47)$$

Similarly, we have

$$\mathbb{Q}\left(h_T>K|h_t=h\right) = \frac{C}{2^{\nu/2}\eta^{2+\nu/2}}e^{A^2/2}\mathcal{Q}_{\nu+1,\nu}(A,\sqrt{2\eta K}). \qquad (3.48)$$

By replacing the expectation and the probability in 3.45 with the relations 3.47-3.48, we prove the 3.43. □

Remark 7.3.3. *We observe that a suitable model for the unit cost of care function Q that satisfies Equation 3.44 is the following exponential affine form*

$$Q(h,t) = \exp(rt + u(t) + v(t)h), \tag{3.49}$$

for some continuously differentiable functions u and v. In order to find these functions, it suffices to replace the partial derivatives of Q in 3.44:

$$u'(t) + v'(t)h + \frac{1}{2}\sigma^2 h v^2(t) + \overline{\kappa}(\overline{\theta} - h)v(t) = 0,$$

for all $h > 0$, $t \in (0,T)$. We derive the system of ordinary differential equations of Riccati type:

$$\begin{cases} u'(t) = -\overline{\kappa}\overline{\theta}v(t), \\ v'(t) + \frac{1}{2}\sigma^2 v^2(t) = \overline{\kappa}v(t). \end{cases}$$

By considering the change of variable $w(t) = \exp(\frac{1}{2}\sigma^2 \int_0^t v(s)ds)$, it is easy to see that w solves the equation $w''(t) = \overline{\kappa}w(t)$. Therefore, we can integrate both the equations in the system, to get

$$\begin{cases} u(t) = -\frac{2\overline{\kappa}\overline{\theta}}{\sigma^2}\log\left[\frac{R}{\overline{\kappa}}\left(e^{\overline{\kappa}t} - 1\right) + 1\right] + u_0, \\ \\ v(t) = \frac{2Re^{\overline{\kappa}t}}{\sigma^2\left[\frac{R}{\overline{\kappa}}(e^{\overline{\kappa}t}-1)+1\right]}, \end{cases}$$

for some costant coefficients $R > 0$, $u_0 \in \mathbb{R}$. We remark that, in particular, the function Q in 3.49 is strictly increasing in h. In order to estimate our model, in the following section, we refer to the exponential affine form.

In order to estimate our model, in the following section, we refer to the exponential affine form for the function Q, where we assume that $u_0 = 0$.

7.3.3.1 Real parameters estimation

Starting from the ES observation under the real probability measure the parameters of the model 3.37 are estimated. ES is considered to be the realization of the same stochastic process with transition density 3.39. Each ES at time t is associated with all the ES at time $t+1$. In order to estimate the parameters of the density function the dataset is divided considering the ES recording in time t. For each cluster the empirical probability distribution is evaluated and, according the Minimum Mean Squared Error (MMSE) method, the following objective function is considered:

$$\min_{\Theta} \sum_{i=1}^{N}\sum_{k=1}^{N}[p(h_t(k)|h_0(i), \Theta) - f(h_t(k)|h_0(i))]^2 \tag{3.50}$$

where $\Theta = [k, \theta, \sigma]$ and N is the maximum ES considered. In Table 7.6 the parameters

κ	θ	σ
2.33	5.46	3.37

Table 7.6: Cox-Ingersoll-Ross real distribution parameters estimation.

of the model are estimated and in Table 7.7 the error metrics are reported.

APE	AAE	ARPE	RMSE
24.48	0.02	0.25	0.02

Table 7.7: Error metrics of Cox-Ingersoll-Ross real parameters estimation.

7.3.4 Parameters estimation

The whole amount of money invested by the healthcare system, at time t, is

$$\Delta_t \cdot Q_t - B_t \tag{3.51}$$

where$\Delta_t \cdot Q_t$ describes the cost attributable to the disease state of the patient. This cost reflects a typical behaviour of resource consumption in a hospital ward. If the manager decides to allocate resources to improve the health state, he will have a lower amount of money to invest in alternative assets. Moreover if a patient has a higher level of disease, he will require more resource than one less sick, even if the hospital assigns the same amount of resources to both patients (they have the sameΔ_t). If the manager chooses the amount of careΔ_t so that to replicate the premiumΠ_t for each time t, then he realizes a perfect match between the cost of hospitalization and the amount paid by patient P. Unfortunately, even if the cost of hospitalization is quantified using the functionΠ , a complete *hedging* strategy cannot be applied in practice since clinicians and executives do not operate continuously in time. In fact, their actions occur at discrete times, implying replication errors that have a negative impact in the hospital cost management. However this point explains the real situation where the healthcare system is not able to match perfectly the costs coverage. In order to validate our approach, we compare the premium Π with the cost surface estimated 1.2. The objective function is

$$\min_{\Theta} \sum_{i=i}^{N_{ES}} \sum_{k=1}^{LoS} \Big[\hat{C}(ES_i, k) - \Pi(ES_i, 0; k, \Theta)\Big]^2. \tag{3.52}$$

where N_{ES} is the number of ES and $\Theta := (K, \bar{\kappa}, \bar{\theta}, \sigma) \in (0 + \infty)^4$. The optimization routine is implemented in the MatLab environment using existing algorithms based on the interior-point method. The optimization starts from a Halton quasi-random sequence of 50 initial points [99]. In Table 7.8 are reported the estimated parameters for the CIR model. The strike parameter K estimated shows that if the patient has a

K	$\bar{\kappa}$	$\bar{\theta}$	σ
7.00	0.02	356.66	7.42

Table 7.8: Cox-Ingersoll-Ross approach cost function parameters estimated.

ES bigger then 7 the patient may decide to exercise the call option. In Table 7.9 are reported the discrepancy measure for the cost fitting procedure. Values show that the proposed model provides a good estimation of the cost surface. Moreover in order to

APE	AAE	ARPE	RMSE
3.34	115.90	0.06	147.09

Table 7.9: Error metrics of Cox-Ingersoll-Ross approach. Cost function surface fitting.

validate the proposed approach, the following **Average Daily Error** (**ADE**) formula is considered:

$$ADE = \frac{1}{N_{ES}} \sum_{i=i}^{N_{ES}} \sum_{k=1}^{LoS} \frac{|\hat{C}(ES_i, k) - \Pi(ES_i, 0; k, \Theta)|}{k}. \tag{3.53}$$

The ADE of the fitting procedure is 53.32. Figure 7.3 points out that the model provides a good description of the empirical data, the slope of the curve shows a slight difference in the input variable dependences. For small LoS the model shows a lower sensitivity to the EuroSCORE with respect to the empirical cost function estimated and for higher LoS the sensitivity to the EuroSCORE of the model is higher.

7.3.4.1 Clustering analysis

In order to investigate the properties of the model a clustering analysis is performed. For each value of ES a different parameters estimation is executed. In Table 7.10 the estimated parameters are reported; values show a relation between the parameters and the value of ES. In particular K and $\bar{\boldsymbol{\theta}}$ could be described by a linear function of ES. In Table 7.11 the error metric are reported and the ADE is 0.30. Using a bigger number of

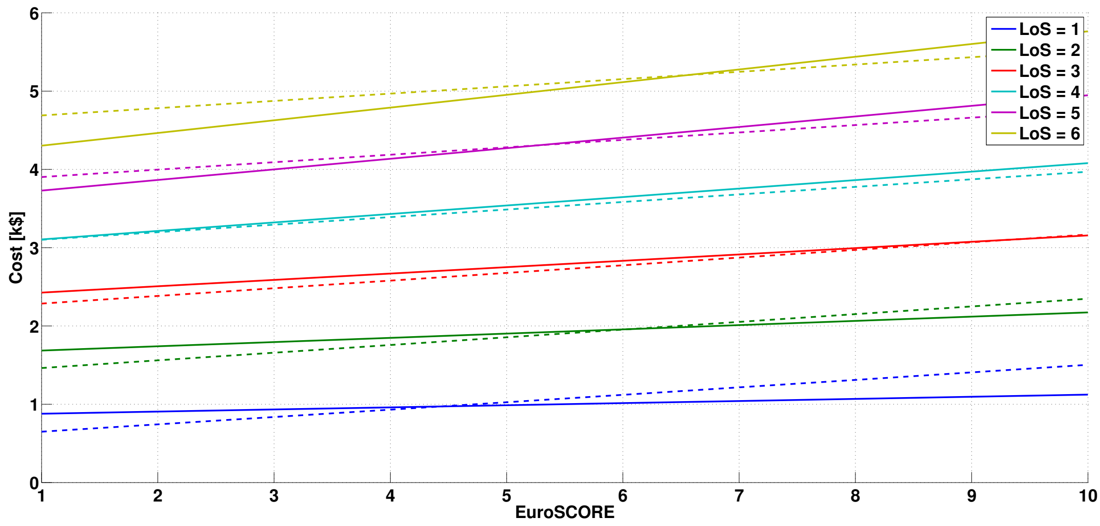


Figure 7.3: Comparison between real hospitalization value and CIR estimation. - The continuous lines represent the real costs and the dashed lines represent the estimated costs.

degree of freedom the proposed approach allows to better describe the empirical costs. According to value of Table 7.10, Figure 7.4 represents the average value of the process evaluated for each cluster.

7.4 Summary

It is proposed a new approach in order to describe the variable cost associated to the patient hospitalization. The framework of the model is described and the main assumptions are introduced. It is proposed the application of the Black-Scholes formula to describe the empirical data. It is also proposed the application of an extension of the Cox-Ingersoll-Ross, a general introduction of the model is provided and the semi-closed pricing formula is in introduced. The model proposed uses a few number of parameters and it supplies a phenomenological interpretation of data. We also introduced risk metrics in order to analysed the fitting error.
The following conclusions can be drawn:

1. The surface cost function presented allows a good description of empirical data. It is assumed that hospitalization cost arising from three main factors:

ES	K	$\bar{\kappa}$	$\bar{\theta}$	σ
1	1.53	0.10	7.32	1.18
2	2.31	0.09	7.36	0.60
3	3.25	0.09	7.51	0.40
4	4.23	0.09	8.49	0.33
5	5.23	0.08	9.48	0.28
6	6.21	0.08	10.57	0.25
7	7.22	0.08	11.67	0.24
8	8.20	0.08	12.32	0.20
9	9.21	0.07	13.65	0.20
10	10.22	0.07	14.94	0.20

Table 7.10: Clustering analysis estimated parameters of Cox-Ingersoll-Ross model.

APE	AAE	ARPE	RMSE
0.03	0.59	0.00	0.74

Table 7.11: Clustering analysis error metrics of Cox-Ingersoll-Ross approach. Cost function surface fitting.

- Admission health state
- Fixed costs
- Clinical complication

Furthermore the proposed approaches provide a general framework that can be applied to different departments.

2. BS model provides closed pricing formula and supplies a phenomenological interpretation of data. It describes the relation between the treatment of patient, the bed cost, and the number of patients hospitalized.

3. CIR model provide semi-closed pricing formula and supplies a phenomenological interpretation of data. It describes the evolution of the disease state during the hospitalization of the patients. In fact we assume that disease state will to the average state over time.

4. BS and CIR models provide a good fit of the real data in fact it performs the smallest discrepancy measures (AAE, APE, ARPE, and RMSE).

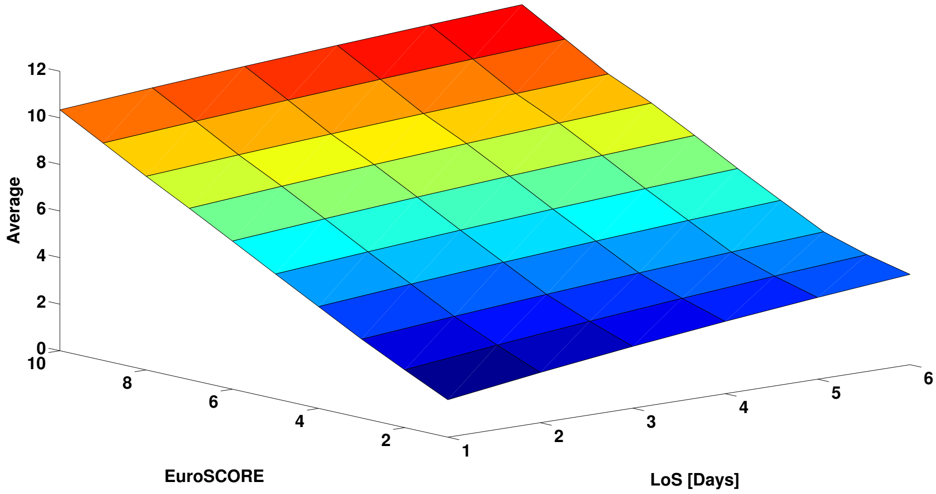


Figure 7.4: CIR average value estimation. - For each cluster the average value is estimated.

The provided framework used financial mathematical model in healthcare contest providing a new management method for healthcare optimization resources.

8

Nonlinear filtering methods

During the Ph.D. studies in addition to research on healthcare issues other topic were investigated. Particularly according to the Memorandum of Understanding between the Campus Bio-Medico University and the Financial Guard the Italian tax evasions phenomena and the Financial Guard law enforcement are analysed [172]. A stochastic version of Lotka-Volterra model is applied in order to describe the dynamic relation existing between the prevention activities and the illegal behaviour. The proposed model aim to provide guidelines to optimize the financial inspection resource consumption. Assuming that the prey represents the positive control and that the predators represent the control performed by Financial Guard, a forecast of the positive control number is provided.
In this Chapter describes a model to optimize the consumption of financial inspection resources for tax evasion by analysing the interaction between prevention/control activities and illegal behaviours.

8.1 Kalman filter

Nonlinear filtering is a challenging problem of system identification. It has a wide field of applications, beyond the planar tracking. To cite a few, nonlinear filtering is applied to electronics [173], robotics [174], navigation [175], geolocalization [176], aerospace [177, 178], biology [179], medical sciences [180], finance [181], meteorology

[182]. Consider the following system:

$$\begin{aligned} \mathbf{x}_{k+1} &= A_k \mathbf{x}_k + B_k \mathbf{u}_k + \mathbf{v}_k^{(1)} \\ \mathbf{y}_k &= C_k \mathbf{x}_k + D_k \mathbf{u}_k + \mathbf{v}_k^{(2)}, \end{aligned} \tag{1.1}$$

where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{u} \in \mathbb{R}^p$, $\mathbf{y} \in \mathbb{R}^q$ and $\mathbf{v}_k^{(1)} \in \mathbb{R}^n$ and $\mathbf{v}_k^{(2)} \in \mathbb{R}^q$ are the state noise and the measurement noise.

In order to estimate the state x using the Kalman filter the following hypothesis must hold:

(i) $\mathbf{v}_k^{(1)}$ and $\mathbf{v}_k^{(2)}$ are Gaussian with zero mean.

(ii) $\mathbf{v}_k^{(1)}$ and $\mathbf{v}_k^{(2)}$ are *white noise*: the samples are independent and have identical probability distribution. In particular:

$$\begin{aligned} \mathbb{E}[\mathbf{v}_{k-j}^{(1)} (\mathbf{v}_k^{(1)})^T] = \underline{0}, \quad \forall\ k,\ \forall\ j \neq 0, \\ \mathbb{E}[\mathbf{v}_{k-j}^{(2)} (\mathbf{v}_k^{(2)})^T] = \underline{0}, \quad \forall\ k,\ \forall\ j \neq 0. \end{aligned} \tag{1.2}$$

(iii) $\mathbf{v}_k^{(1)}$ and $\mathbf{v}_k^{(2)}$ are mutually uncorrelated. The error covariance matrix is

$$\mathbb{E}[\mathbf{v}_k^{(1)} (\mathbf{v}_k^{(2)})^T] = \underline{0}, \quad \forall\ k. \tag{1.3}$$

Without loss of generality we assume $\mathbf{v}_k^{(1)} = \tilde{F}_k N_k^{(1)}$ and $\mathbf{v}_k^{(2)} = \tilde{G}_k N_k^{(2)}$, where $N_k^{(1)}$ and $N_k^{(2)}$ are white standard sequences (succession of Gaussian random variables with zero mean uncorrelated for each instant of time). It is easy to verify that the assumptions *(i)*, *(ii)* and *(iii)* are satisfied. We can evaluate the state and measurement covariance matrix for each time k

$$\begin{aligned} \mathbb{E}[\mathbf{v}_k^{(1)} (\mathbf{v}_k^{(1)})^T] = \tilde{F}_k \tilde{F}_k^T = Q_k \\ \mathbb{E}[\mathbf{v}_k^{(2)} (\mathbf{v}_k^{(2)})^T] = \tilde{G}_k \tilde{G}_k^T = R_k. \end{aligned} \tag{1.4}$$

Then we can rewrite the system 1.1 in the following way

$$\begin{aligned} \mathbf{x}_{k+1} &= A_k \mathbf{x}_k + B_k \mathbf{u}_k + F_k N_k \\ \mathbf{y}_k &= C_k \mathbf{x}_k + D_k \mathbf{u}_k + G_k N_k, \end{aligned} \tag{1.5}$$

where

$$N_k = \begin{bmatrix} N_k^{(1)} \\ N_k^{(2)} \end{bmatrix}, \quad F_k = \left[\tilde{F}_k,\ \underline{0}\right], \quad G_k = \left[\underline{0},\ \tilde{G}_k\right], \tag{1.6}$$

so that $F_k N_k = \tilde{F}_k N_k^{(1)}$ and $G_k N_k = \tilde{G}_k N_k^{(2)}$.
Using the principle of superposition, we distinguish stochastic component $\mathbf{x}^s$ and the deterministic component $\mathbf{x}^d$ of 1.5 so that $\mathbf{x}_k = \mathbf{x}_k^d + \mathbf{x}_k^s$

$$\begin{aligned} \mathbf{x}_{k+1}^d &= A_k \mathbf{x}_k^d + B_k \mathbf{u}_k \\ \mathbf{x}_{k+1}^s &= A_k \mathbf{x}_k^s + F_k N_k. \end{aligned} \tag{1.7}$$

Setting $\mathbf{x}_0^d = \mathbb{E}[\bar{\mathbf{x}}_0]$, $\mathbb{E}[\mathbf{x}_0^s] = 0$ and $\mathbb{E}[\mathbf{x}_0^s (\mathbf{x}_0^s)^T] = \Psi_{\bar{\mathbf{x}}_0}$, where $\bar{\mathbf{x}}_0$ is the initial condition random variable, adding 1.7 we find the state Equation 1.5. In the same way the stochastic component of the measurement equation can be wrote

$$\mathbf{y}_k - C_k \mathbf{x}_k^d - D_k \mathbf{u}_k = \mathbf{y}_k^s = C_k \mathbf{x}_k^s + G_k N_k. \tag{1.8}$$

The solution of 1.7 can be calculated recursively. Focusing now on the state variable $\mathbf{x}^s$, we obtain the new equation system

$$\begin{aligned} \mathbf{x}_{k+1}^s &= A_k \mathbf{x}_k^s + F_k N_k \\ \mathbf{y}_k^s &= C_k \mathbf{x}_k^s + G_k N_k, \end{aligned} \tag{1.9}$$

for which we apply the minimum variance estimation.
In the following we call the stochastic component of the state and the measurement of 1.9 with $\mathbf{x}$ and $\mathbf{y}$.
The minimum variance estimation is the conditional expected value of $\mathbf{x}_k$ with respect to $\{\mathbf{y}_j\}_{j \le k}$. The optimal estimation $\hat{\mathbf{x}}$ at time k is given by

$$\hat{\mathbf{x}}_{k|k} = \mathbb{E}[\mathbf{x}_k | \mathcal{F}_k^{\mathbf{y}}], \tag{1.10}$$

where $\mathcal{F}_k^{\mathbf{y}}$ is the filtration $\sigma(\mathbf{y}_j, j \le k)$.
Calling with $\hat{\mathbf{x}}_{k|k-1} = \mathbb{E}[\mathbf{x}_k | \mathcal{F}_{k-1}^{\mathbf{y}}]$ the state prediction of $\mathbf{x}_k$ knowing the measures until the previous instant. Define the state innovation as

$$\begin{aligned} \nu_k^s &= \mathbb{E}[\mathbf{x}_k | \mathcal{F}_k^{\mathbf{y}}] - \mathbb{E}[\mathbf{x}_k | \mathcal{F}_{k-1}^{\mathbf{y}}] \\ &= \hat{\mathbf{x}}_{k|k} - \hat{\mathbf{x}}_{k|k-1}, \end{aligned} \tag{1.11}$$

and the measurement innovation as

$$\begin{aligned} \nu_k^o &= \mathbb{E}[\mathbf{y}_k | \mathcal{F}_k^{\mathbf{y}}] - \mathbb{E}[\mathbf{y}_k | \mathcal{F}_{k-1}^{\mathbf{y}}] \\ &= \mathbf{y}_k - \hat{\mathbf{y}}_{k|k-1}. \end{aligned} \tag{1.12}$$

From the conditional expected value properties, can be proved that the state and measurement innovations are white sequences with zero mean,

$$\mathbb{E}[\nu_k^s] = \underline{0}, \quad \mathbb{E}[\nu_k^o] = \underline{0}. \tag{1.13}$$

and

$$\mathbb{E}[\nu_k^s(\nu_j^s)^T] = 0, \quad \mathbb{E}[\nu_k^o(\nu_j^o)^T] = 0, \quad \forall\ j \neq k. \tag{1.14}$$

Moreover, for all $j \neq k$

$$\mathbb{E}[\nu_k^s(\nu_j^o)^T] = 0. \tag{1.15}$$

Last step of the Kalman filter is the Equivalence theorem: the filtration generated by the measurement sequence coincides with the filtration generated by the innovation measurement sequence. Therefore

$$\mathcal{F}_k^{\mathbf{y}} \equiv \mathcal{F}_k^{\nu^o} \tag{1.16}$$

then

$$\hat{\nu}^s = \mathbb{E}[\nu_k^s|\mathcal{F}_k^{\mathbf{y}}] = \nu^s = \mathbb{E}[\nu_k^s|\mathcal{F}_k^{\nu^o}]. \tag{1.17}$$

The optimal estimation of Gaussian vector is linear, then:

$$\nu^s = \mathbb{E}[\nu_k^s|\mathcal{F}_k^{\nu^o}] = \sum_{j=0}^{k} K(k,j)\nu_j^o, \tag{1.18}$$

where $K(k,j)$ is the Kalman gain matrix.
According to 1.15, $K(k,j)$ is different from 0 when $j = k$. Setting $K(k,k) = K_k$ we find that

$$nu_k^s = K_k\nu_k^o. \tag{1.19}$$

Therefore from 1.11, 1.12 and 1.9 we can find $\hat{\mathbf{x}}_k$

$$\hat{\mathbf{x}}_k = A_{k-1}\hat{\mathbf{x}}_{k-1} + K_k[\mathbf{y}_k - C_k A_{k-1}\hat{x}_{k|k-1}]. \tag{1.20}$$

In order to explicit the Kalman gain matrix K_k we define the error estimation covariance matrix and the error prediction covariance matrix

$$\begin{aligned} P_k &= \mathbb{E}[\hat{e}_k\hat{e}_k^T] \\ P_{k|k-1} &= \mathbb{E}[\hat{e}_{k|k-1}\hat{e}_{k|k-1}^T], \end{aligned} \tag{1.21}$$

where the estimation error is $\hat{e}_k = \mathbf{x}_k - \hat{\mathbf{x}}_k$ and the prediction estimation error is $\hat{e}_{k|k-1} = \mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1}$. It can be shown that 1.21 are linked with Kalman gain K_k with the following relation:

$$\begin{aligned} P_{k|k-1} &= A_{k-1}P_{k-1}A_{k-1}^T + Q_{k-1} \\ K_k &= P_{k|k-1}C_k^T[C_kP_{k|k-1}C_k^T + R_k]^{-1} \\ P_k &= P_{k|k-1}[I - K_kC_k]. \end{aligned} \tag{1.22}$$

In order to find the optimal estimation of $\mathbf{x}$, using 1.22, we can evaluate 1.20. Given that $\mathbf{x}_k = \mathbf{x}_k^d + \mathbf{x}_k^s$, we write 1.20 in the following way

$$\hat{\mathbf{x}}_k = A_{k-1}\hat{\mathbf{x}}_{k-1} + K_k[\mathbf{y}_k - C_kA_{k-1}\hat{x}_{k|k-1} - D_k\mathbf{u}_k]. \tag{1.23}$$

8.1.1 Non-linear filtering problem

In this section we extend the Kalman filter to non-linear system models to obtain an approximate filter the **Extended Kalman Filter** (**EKF**). In order to do that by finding an appropriate error system that is linear approximation of a non-linear system, it offers no guarantees of optimality in a mean squared error sense. However, for many system, the EKF has proven to be a useful method of obtaining good estimates of the system state. The non-linear filtering problem requires the recursive estimation of the state $\mathbf{x}_k \in \mathbb{R}^n$ of a nonlinear stochastic discrete-time system having the form:

$$\mathbf{x}_{k+1} = f(\mathbf{x}_k, \mathbf{u}_k, \mathbf{v}_k^{(1)}, k), \ \mathbf{x}_0 = x_0 \tag{1.24}$$

$$\mathbf{y}_k = h(\mathbf{x}_k, \mathbf{u}_k, \mathbf{v}_k^{(2)}, k), \tag{1.25}$$

where $\mathbf{u}_k \in \mathbb{R}^p$ is the deterministic control input, $\mathbf{y}_k \in \mathbb{R}^q$ is the available measurement vector, $\mathbf{v}_k^{(1)} \in \mathbb{R}^r$ is the state noise sequence, $\mathbf{v}_k^{(2)} \in \mathbb{R}^s$ is the measurement error sequence, and $f : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^r \to \mathbb{R}^n$ and $h : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^s \to \mathbb{R}^q$ are possibly non-linear functions. Moreover, the following properties are usually assumed:

(3.a) $\mathbf{v}_k^{(1)}$ and $\mathbf{v}_k^{(2)}$ are independent zero mean white sequences of random vectors with the covariance matrices$\Psi_{\mathbf{v}_k^{(1)}} \in \mathbb{R}^{n\times n}$ and$\Psi_{\mathbf{v}_k^{(2)}} \in \mathbb{R}^{q\times q}$, respectively;

(3.b) the initial state $x_0 \in \mathbb{R}^n$ is a random vector with the mean value $\bar{x}_0$ and covariance matrix$\Psi_{x_0} \in \mathbb{R}^{n\times n}$, independent of $\mathbf{v}_k^{(1)}$ and $\mathbf{v}_k^{(2)}$.

The estimation is required to be carried out starting from the initial estimate $\bar{x}_0$ and using the measurement sequence $\mathbf{Y}_k := \{\mathbf{y}_\tau, \tau = 0, 1, \ldots, k\}$.
From a probabilistic perspective, any statistical estimate of $x(k)$ will be some function of the conditional density:

$$\mathbb{P}\big(\mathbf{x}_k | \mathbf{Y}_k\big) \tag{1.26}$$

since it represents all the information which the measurement of the random vectors belonging to $\mathbf{Y}_k$ has conveyed about the random vector $\mathbf{x}_k$. This statistical estimate is denoted by $\tilde{\mathbf{x}}_k$. Suppose now that $\tilde{\mathbf{x}}_k$ is given as a fixed function of the random vectors in Y_k. Then $\tilde{\mathbf{x}}_k$ itself is a random vector and its actual value is known whenever the actual value of $\mathbf{Y}_k$ is known. In general, the actual value of $\tilde{\mathbf{x}}(k)$ is different from the unknown value of $\mathbf{x}_k$ [177]. To arrive to a rational way of determining $\tilde{\mathbf{x}}_k$, it is natural to define an optimality criterion. This can be done in different ways. Next, the two most used approaches are presented. It will be clear that both of them provide the same theoretical result about the optimal solution. However, it is worth noting that different approaches may yield vary different suboptimal solutions (e.g. EKF).

Minimum Variance Estimate Estimate optimality is usually defined by assigning a *penalty* or *loss* for incorrect estimates. Clearly, the loss should be a positive, nondecreasing function of the estimation error $\tilde{\mathbf{e}}_k := \mathbf{x}_k - \tilde{\mathbf{x}}_k$. The common choice for such a function is $L(\tilde{\mathbf{e}}_k) = \mathbb{E}\left[\| \tilde{\mathbf{e}}_k \|^2\right]$, $\| \cdot \|$ being the euclidean norm in $\mathbb{R}^n$. This choice leads up to the so called *minimum variance estimate* which is *the random vector* $\hat{\mathbf{x}}_k$ *that minimizes* $L(\tilde{\mathbf{e}}_k)$. It is well known that *the minimum variance estimate is the state conditional expectation*, i.e.

$$\hat{\mathbf{x}}_k = \mathbb{E}\left[\mathbf{x}_k | \mathbf{Y}_k\right]. \tag{1.27}$$

Moreover, $\mathbb{E}\left[\hat{\mathbf{x}}_k\right] = \mathbb{E}\left[\mathbf{x}_k\right]$, i.e. $\hat{\mathbf{x}}_k$ is *unbiased.*
It is useful to show that the minimum variance estimate can be also argued by using a geometrical approach, as done in [183]. Let $\big(\Omega, \mathcal{F}, \mathbb{P}\big)$ be a probability space. For any given sub σ-algebra $\mathcal{G}$ of $\mathcal{F}$, denote by $L^2(\mathcal{G}, n)$ the Hilbert space of the n-dimensional, $\mathcal{G}$-measurable, random vectors with finite second moment as

$$L^2(\mathcal{G}, n) := \left\{\mathbf{x} : \Omega \to \mathbb{R}^n, \mathcal{G}\text{-measurable}, \int_\Omega \| \mathbf{x}_\omega \|^2 \, dP_\omega < \infty\right\}, \tag{1.28}$$

where $\int_\Omega \cdot \, dP_\omega$ indicates the Lebesgue integral. This Hilbert space is endowed with the internal product

$$[\mathbf{x}, \mathbf{z}] := \int_\Omega \mathbf{x}_\omega^T \mathbf{z}_\omega dP_\omega = E\left[\mathbf{x}^T \mathbf{z}\right] \tag{1.29}$$

and the induced norm

$$\| \mathbf{x} \|_{L^2} := \sqrt{\int_\Omega \mathbf{x}_\omega^T \mathbf{x}_\omega dP_\omega} = \sqrt{\mathbb{E}\left[\| \mathbf{x} \|^2\right]}. \tag{1.30}$$

Moreover, when $\mathcal{G}$ is the σ-algebra generated by a random vector $\mathbf{y} : \Omega \rightarrow \mathbb{R}^m$, that is $\mathcal{G} = \sigma(\mathbf{y})$, the notation $L^2(\mathbf{y}, n)$ indicates $L^2(\sigma(\mathbf{y}), n)$. Finally, if $\mathcal{B}$ is a subspace of $L^2(\mathcal{F}, n)$, the symbol$\Pi(\, \mathbf{x}|\mathcal{B})$ indicates the orthogonal projection of $\mathbf{x} \in L^2(\mathcal{F}, n)$ onto $\mathcal{B}$.

From the Hilbert projection theorem [184] follows that the minimum variance estimate of a random vector $x \in L^2(\mathcal{F}, n)$ with respect to a random vector y corresponds to $\Pi(\mathbf{x}|L^2(\mathbf{y}, n))$. Therefore, by assuming that the state $\mathbf{x}_k$ has a finite second order moment, it results that

$$\hat{\mathbf{x}}(k) = \mathbb{E}\left[\mathbf{x}_k | \mathbf{Y}_k\right] = \Pi(\, \mathbf{x}_k | L^2\big(\mathbf{Y}_k, n\big). \tag{1.31}$$

8.1.2 Extended Kalman Filter

The EKF is the classical filter for non-linear systems. Despite it suffers from significant drawbacks, for many time, it has been considered the standard tool of choice for real tracking applications. Moreover, from a scientific point of view, it is considered as a first benchmark and its performances represent the *bare minimum* for any non-linear filtering method. The idea of EKF is to linearize the system equations (1.36)-(1.25) around the estimated trajectory. The linearized state equation has the form

$$\begin{aligned}
\mathbf{x}_{k+1} &\approx f(\hat{\mathbf{x}}_{k|k}, \mathbf{u}_k, 0, k) + A_k\big(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k}\big) + F_k N_k \\
&= A_k \mathbf{x}_k + \underbrace{f(\hat{\mathbf{x}}_{k|k}, \mathbf{u}_k, 0, k) - A_k \hat{\mathbf{x}}_{k|k}}_{\mathbf{u}_k^f} + F_k N_k \\
&= A_k \mathbf{x}_k + \mathbf{u}_k^f + F_k N_k
\end{aligned} \tag{1.32}$$

where the Jacobian matrices

$$A_k := \Big(\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}, \mathbf{u}_k, 0, k)\Big)\Big|_{\mathbf{x} = \hat{\mathbf{x}}_{k|k}}, \quad F_k = \Big(\frac{\partial}{\partial \mathbf{v}^{(1)}} f(\hat{\mathbf{x}}_{k|k}, \mathbf{u}_k, \mathbf{v}^{(1)}, k)\Big)\Big|_{\mathbf{v}^{(1)} = 0}, \tag{1.33}$$

and the artificial input u^{f_k} are known at the prediction time, since the state estimate $\hat{\mathbf{x}}_{k|k}$ is available.

The linearized measurement equation is

$$\begin{aligned}
\mathbf{y}_k &\approx h(\hat{\mathbf{x}}_{k|k-1}, \mathbf{u}_k, 0, k) + C_k\big)\big(x(k) - \hat{\mathbf{x}}_{k|k-1}\big) + G_k N_k \\
&= C_k \mathbf{x}_k + \underbrace{f(\hat{\mathbf{x}}_{k|k-1}, u_k, 0, k) - C_k \hat{\mathbf{x}}_{k|k-1}}_{\mathbf{u}_k^h} + G_k N_k \\
&= C_k \mathbf{x}_k + \mathbf{u}_k^h + G_k v_k, \qquad (1.34)
\end{aligned}$$

where the Jacobian matrices

$$C_k = \Big(\frac{\partial}{\partial \mathbf{x}} h(\mathbf{x}, \mathbf{u}_k, 0, k)\Big)\Big|_{\mathbf{x}=\hat{\mathbf{x}}_{k|k-1}}, \quad G_k = \Big(\frac{\partial}{\partial \mathbf{v}^{(2)}} h(\hat{\mathbf{x}}_{k|k-1}, \mathbf{u}_k, \mathbf{v}^{(2)}, k)\Big)\Big|_{\mathbf{v}^{(2)}=0}, \qquad (1.35)$$

and the artificial input u^{h_k} are known at the correction time, since the state prediction $\hat{\mathbf{x}}_{k|k-1}$ is available. Once system 1.36-1.25 is linearized relation 1.23 is applied. Because of linearization, the EKF is a suboptimal algorithm, but in practice it has been proved to work well for many applications. However, the use of EKF has two main well-known drawbacks: linearization can produce highly unstable filters if the assumptions of local linearity is violated; the derivations of the Jacobian matrices are non-trivial in most applications and often lead to significant implementation difficulties [185]. Research has attempted to adopt modifications to the basic algorithm in order to improve the performances and avoid critical behaviours. As a consequence, there is no such thing as the EKF, but rather there are hundreds of varieties of EKFs. In particular, 1) different coordinate systems [186, 187], 2) different factorizations of the covariance matrix, 3) second order (or higher order) Taylor series corrections to the state vector prediction and/or the measurement update [183, 188, 189], 4) iteration of the state vector update using measurements [175], 5) different orders with which to use sequential scalar valued measurements to update the state vector, 6) tuning process noise [190], 7) quasi-decoupling, and 8) combinations of all of the above [191]. Sometimes these practical tricks result in significant improvements, but often they result in no improvement or they make performance worse.

8.1.3 Kalman-Bucy Filter

Let the signal X_t and the observation Y_t are defined by Ito linear equations with respect to independent Wiener process $dW_t^{(1)}$ and $dW_t^{(2)}$

$$dX_t = A(t)X_t dt + B(t)dW_t^{(1)} \tag{1.36}$$

$$dY_t = C(t)X_t dt + D(t)dW_t^{(2)} \tag{1.37}$$

subject to Gaussian initial condition X_0 and Y_0 independent of $dW_t^{(1)}$ and $dW_t^{(2)}$. The deterministic functions $A(t)$, $B(t)$, $C(t)$, and $D(t)$ are assumed to be bounded and piece-wise continuous for $t \geq 0$. The main assumption here is

$$\inf_{t \geq 0} D^2(t) \geq 0 \quad \text{(for some } c > 0) \tag{1.38}$$

The filtering estimates $\hat{X}_t$ and the mean square filtering error $P(t)$ are defined by differential equations (Ito and Riccati)

$$d\hat{X}_t = A(t)\hat{X}_t dt + \frac{P(t)C(t)}{D^2(t)}(dY_t - A(t)\hat{X}_t dt) \tag{1.39}$$

$$dP(t) = 2A(t)P(t) + B(t) - \frac{P^2(t)C^2(t)}{D^2(t)} \tag{1.40}$$

subject to the initial condition:

$$\hat{X}_0 = \mathbb{E}[X_0] + \frac{cov(X_0, Y_0)}{cov(Y_0, Y_0)}(Y_0 - \mathbb{E}[Y_0]) \tag{1.41}$$

$$P(0) = cov(X_0, X_0) - \frac{cov^2(X_0, Y_0)}{cov(Y_0, Y_0)} \tag{1.42}$$

8.2 GDF case of study: Tax Evasion Dynamics by Fiscal Inspection Activity

Tax evasion in Italy is a serious issue: between a quarter and half of the Gross domestic product (GDP) seems to be hidden to the tax authorities Schneider [192]. This is of crucial importance at a macro economic level to ensure the reliability of official statistics and the efficiency of national productions. However they do not provide insights to policy makers that wish to investigate who are tax evaders and to start understanding the reasons why some taxpayers might consider under-declaring their income.

Measuring tax evasion is all but simple: Schneider [193] describes tax evasion measurement as a *scientific passion for knowing the unknown*. However, tax evasion analysis is relevant for public policy design and for estimating the bias that tax evasion introduces in some statistics, both at the macro and at the micro level.

It is possible to divide the study on phenomenon of tax evasion in two macro-interconnected areas. The first one is related to the methodologies of estimation of tax evasion and is based on the analysis and processing of datasets obtained from the field. The second concerns the study of the tax evasion models in relation with social and economic aspects. The phenomenon of deterrence, that is particularly important in this article, is treated separately in a special section.

Bernardi [194] offer an overview about the methodologies to estimate the tax evasion. Each methodology is strictly related to the availability of informations from the field and to the institutional context.

There are three different techniques to estimate the tax evasion:

1. Estimation of monetary and underground economy indicators. Starting from synthetic indicators of the actual *size* of the real economy we get the amount of evaded taxes with respect to the tax-detectable.

2. Method of national accounting. The total amount of taxable incomes reported is compared with the estimation of potential incomes. This is the most used method, it is particularly reliable and easily adaptable to different socio-economic contexts.

3. Sampled comparisons between declared and accrued income. It is used a micro-economic approach based on sampled income surveys, related to income estimation obtained from other sources (i.e. wealth and household consumption).

The 1) and 2), as macro-economic models, do not allow an accurate *resolution* of the results. 3) is strictly dependent on control methods and sampled population, and often provides inconsistent results.

However, any methodology must struggle with the need to calculate an entity that does not exist in the real world, and so inherently unknowable. An estimation also requires the comparison of complex, heterogeneous and often numerically inadequate data. It should also be noted that these methods are generally disconnected from the models of the decision to evade, which will be described later, and that a correlation between

the two fields have often provided inconsistent results.
Quantitative indicators on tax evasion have been produced since the early eighties in different countries. In United States the TCPM (Taxpayer Compliance Measurement program) is calculated by the IRS (Internal Revenue Service) and is estimated at about 17 percent of the total amount of income taxes Andreoni [195]. With similar methods, similar indicators were obtained in other advanced countries: i.e. in Holland Hessing [196], Switzerland Pommerehne [197], Canada KPMG [198] and Spain De Juan [199]. The level of tax evasion is higher in less developed countries. The Mediterranean countries also have the common characteristics of high administrative corruption and an unconsolidated civic education and this involves in a higher level of tax evasion Bovi [200]. It is estimated that the level of tax evasion seen in Italy is twice the Organization for Economic Cooperation and Developmen (OECD) countries average Tanzi [201].
Tax evasion can also be estimated using an indirect approach. Indirect methods estimate tax evasion considering it equal to the difference between aggregated macro indicators (e.g the discrepancy between income and expenditures or the difference between the actual demand for money and the demand for money estimated in absence of taxes). Direct methods aim at estimating tax evasion through the use of sample survey micro-data based on voluntary participation or the results of the auditing activity of tax authorities. In contrast to indirect methods, direct methods are more suitable to analyze tax evasion at the micro level and they can point out directions for policy.
Some of these methods have been applied to provide a measure of tax evasion in Italy. Among those who used indirect methods, Schneider [192] used the currency demand approach, Zizza [202] also the factorial analysis. Zizza estimates the share of the underground economy (excluding illegal and criminal activities) on GDP for the years 1984-2000 between a maximum of 17.6% (1991) and a minimum of 14.3% (2000). Schneider's estimates include also illegal and criminal activities. According to him the share of the underground economy on the italian GDP is very high and increasing (from 25.8% in 1994 to 27.8% in 1998), the highest rate among the OECD countries.
Calzaroni [203], Bernasconi [204], Marenzi [205], and Cannari [206] used direct methods. Calzaroni [203] estimates labor supply and labor demand functions by sectors using household and firm surveys, respectively, and compares results at the national and the regional level. The difference between the two is considered to be the number of the irregular workers. This amount, multiplied for the average sectorial productivity

estimated for regular workers gives a first measure of the underground economy. The overall incidence of the underground economy is calculated complementing this amount with coefficients correcting for the underestimation of the turnover and the balancing between aggregated input and output. This methodology relies heavily on the hypothesis that the Survey of (Household Income and Wealth) SHIW data set is representative of the population and of its subgroups. As Ministry of Finance (MF) data refers to the population, a measure of tax evasion based on this methodology requires the Bank of Italy (BI) data to be a good approximation of the population. This requirement must be verified carefully.

Cannari [206], Bernasconi [204] use total net income by group of taxpayers. However, the BI data set is quite reliable for the measurement of work income but it is much less so for other types of income such as capital, estate and building income Cannari [206]. This is due to two main reasons: first, these data are collected at the household level and they can only be imputed to the individual taxpayer; second, there is a tendency to misestimate the true value of these incomes, which is probably not voluntary and however common also to other similar surveys. For these reasons, we suggest here to focus only on work incomes.

The aim of this study is to analyse the mutual interaction between the tax evasion and tax assessment, therefore we introduce a dynamic model to explain the inspection activity related to instrumental controls carried out by Financial Guard. In order to describe populations that interact, thereby affecting each others growth rates, the application of a stochastic version of the famous Lotka-Volterra model is proposed. The model shows a good ability to determine the irregular control number within one month with respect to the time series analysed.

8.2.1 Overview on Modelling Tax Evasion

Allingham [207] (A/S Model) in 1972 produced the main tax evasion behavioural model. The A/S Model describes the tax evasion as a portfolio choice under uncertainty. This approach provides a correlation between the economic theory of crime Becker [208] and the theory of insurance choices against risk in an uncertainty context Arrow [209]. The A/S Model presents the choice to evade taxes as a decision of a rational economic agent who wants to maximize their expected utility given a certain probability of being subjected to scrutiny: if the agent evades gets a prize. The model is based on only two policy-control instruments: the frequency of tax audits and the retributions. The model also presents the following assumptions:

- Taxpayer is acting rationally
- Taxpayer shows a certain degree of risk aversion
- Taxpayer knows his actual income
- Tax rate is considered to be proportional to income
- Probability of incurring a tax audit is known to the taxpayer.

A large amount of empirical evidence accumulated over years suggests that, compared to the observed behavior of actual taxpayers, the A/S Model predicts too high levels of tax evasion. In the context of their model, the observed low evasion activity could be explained only with unreasonably high Arrow-Pratt-measures of risk aversion. This argument has been put forward by, among others, Alm [210] as well as by Graetz [211], Skinner [212] and, for Swiss data, by Pommerehne [213] and Frey [214]. Another problem of the A/S model is that it provides no correlation with the work done by the tax authorities. The model is also unable to describe misperceptions by the taxpayer (i.e. tax rates, tax audit probability) and administration operational constraints (i.e. frequency of tax audits, etc.). Many papers treat the empirical validation of the model. The studies about the tax-rates influence in the model offer mixed results: Clotfelder [215] states that there is a strong positive correlation between tax evasion and tax rates; Feinstein [216] argue just the opposite. Baldry [217] states that there is a close relation between the effect of the sanction and the frequency of audits and tends to be poorly received if that frequency is low. The study of audits probability will be explained in a later section. Moreover in the A/S Model the problem of the agents perfect rationality rules out any influence related to the socio-economic context, the degree of civic awareness and the political and administrative classes respectability.
Many works have been developed to extend the A/S model and to overcome its limitations. Bordignon [218] introduces a fairness constraint to a standard portfolio choice model of tax evasion. Erard [219] have integrated a variable accounting for psychological costs. Pommerehne [197] formulate a simulation model in which they assume that taxpayers follow a tit for tat strategy, so that *good* taxpayers, who initially not evade for

moral reasons, start to evade taxes when they recognize that politicians deviate from the citizen's preferences. Mittone [220] through experimental economy points out that the social environment affects significantly the attitudes of taxpayers. Flatters [221] focus on the degree of corruption of the administration. Engel [222] offer a detailed study that attempts to provide an estimation of taxpayer's attitude dynamics. The model considers cases in which individuals choose the fraction of income they report to the IRS while facing stochastic probabilities of being audited. Under these circumstances, a rational taxpayer's current evasion is a decreasing function of prior evasion, since, if audited and caught for evading this year, the taxpayer may incur penalties for past evasions. The aggregate behaviour of American taxpayers over the 1947-1993 period is consistent with the implications of this model.

8.2.2 Deterrence Phenomenon

The audits probability is the most important variable that can affect the deterrence. It was been shown that its increase reduces the expected return of evasion and increases the risk premium, leading to a tax evasion decrease. This conclusion is supported both by empirical that econometric studies Andreoni [195]. The effect of the audits probability is also influenced by the perception that the taxpayer receives as a result of changes in the frequency of the audits Keppler [223]. Alm [210] show that the deterrent effect is low in case of low frequency of audits. Alexander [224] argue that in order to reduce the risk of detection, a taxpayer can adopt strategies of *tax planning*, leading to a smaller deterrence effect.

Role of Financial Guard According to the Italian organizational structure, the Financial Guard (*Guardia di Finanza*, short GdF) is an Italian law enforcement agency under the authority of the Minister of Economy and Finance (MEF) and part of the Italian armed forces. The Guard is essentially responsible for dealing with financial crime and smuggling; it has also evolved into Italy's primary agency for suppressing the drugs trade, with functions in operational and unannounced control for preventing and combating all acts which have the effect of fraud and tax evasion. According to his duties, GDF realizes operational and unannounced controls for: verifying the compliance of trade; verifying the production, storage, movement and use of the property; detection and removal operations and illicit activities.

In recent years, the number of fiscal inspection actions that targeted with priority the control of taxpayers with high fiscal risk was growing. The growing internationalization of economic activities has created to multinational companies favourable conditions for fraudulent practices, price manipulation between subsidiaries of the same company beeing commonly used in order to reduce taxation in the country of production or destination. These procedures aim mainly to monitor the goods purchased from the Community in order to reduce the risk of tax evasion and stealing from the correct declaration of tax liabilities, respectively the risk of not collecting the obligations to the consolidated state budget. It was extended the collaboration with european tax administrations, including participation in the realization of multilateral controls to prevent and combat cross-border fraud, improve and perfect the techniques, methods and skills of control. From the results obtained by the tax inspection teams, the areas where fraud is more frequent or large scale are: construction and building materials, production and sale of food goods, production and marketing of energy products, transportation, wood exploitation and processing, production and sale of tobacco products, black and gray labor-use, production and recovery of alcohol and alcoholic-beverages. This project aims to identify and validate statistical indicators, based on the analytical information available, which are able to provide a strategy for a better and effective use of resources. Our study aims to introduce a dynamic model to explain the inspection activity related to instrumental controls carried out by GDF accross all Italian regions. More specifically, the analysis is based on the historical data of instrumental controls conducted by GDF during the period of time from 2002 to 2014. The inspections concern with infractions by public and private subjects in the field of tax receipts. They represent only a part of all activities of GDF however the data set provided represent a good framework of study for a proto-type model that can be extended in order to incorporate more sophisticated and demanding financial frauds. In rest of the chapter, with *positive control* we refer to an inspection where GDF officers have recorded an infraction.

8.2.3 Dynamic Model

In this section we introduce the the modelling framework needed to describe the methodological approach proposed Let $(\Omega, \mathbb{P}, \mathcal{F}, \{\mathcal{F}t\}_{t\geq 0})$ be a complete probability space with filtration $\{\mathcal{F}_t\}_{t\geq 0}$ satisfying the usual conditions, i.e., it is increasing and right continuous while $\mathcal{F}_0$ contains all $\mathbb{P}$-null sets. We propose a continuous time stochastic Lotka-Volterra type of model for the dynamics of the instrumental controls conducted by GDF. The model adopts a logistic form for the growth of the positive controls (*prey*), so as to take into account intra-specific competition. It can be described through the following equations:

$$dx_t = [rx_t(1-x_t) - qx_t y_t]\,dt + \epsilon x_t \left(\rho dW_t^{(1)} + \sqrt{(1-\rho^2)}dW_t^{(2)}\right), \tag{2.43}$$

$$dy_t = (cqx_t y_t - uy_t)\,dt + \eta y_t dW_t^{(1)}, \tag{2.44}$$

where

- x_t and y_t represent the number of positive controls and the total number of controls perfomed by GDF at time t, respectively;
- $r > 0$ is the specific growth rate of the positive control;
- c is the maximum production rate of the control activities;
- $u > 0$ is the specific decreasing rate of the control activities;
- $q > 0$ is a constant parameter representing the efficiency of the control process.

Moreover deterministic initial values $x_0,\ y_0 \geq 0$ are assumed. The random errors driven by $W^{(1)}$, $W^{(2)}$ represent the environmental stochasticity. They are assumed to be two one-dimensional and independently Weiner processes. The magnitude of these errors are proportional to the number of controls through the positive parameters ϵ and η, respectively. The environmental stochasticity is essentially related to the economic context and to the new regulations issued by the Italian government as well as to new instances introduced by GDF. Hence we allow for a dependence of the random errors affecting both the types of controls using a correlation coefficient $\rho \in [-1, 1]$. The lumped parameters r, c, u are controls-specific, whereas the behavioral parameter q in the functional response $qx_t y_t$ is the crucial and unknown parameter to be be estimated. This problem has been addressed in a biological framework by Boucher [225]. In the following we will use the notation $\theta = (r, c, u, q, \epsilon, \eta, \rho) \in \mathbb{R}_+^6 \times [-1, 1]$, to denote the set of model parameters.

We remark that the deterministic subclass of the Lotka-Volterra model are well-known and have been extensively investigated in the literature concerning ecological population modelling. One particularly interesting subclass describes the facultative mutualism of two species, where each one enhances the growth of the other, represented through the deterministic equations. The associated dynamics have been developed by, for example, Boucher [225], He [226] and Wolin [227]. In order to avoid having a solution that explodes at a finite time, additional conditions on the model parameters are needed Mao [228]. Nevertheless, this can be avoided, by introducing a stochastic environmental noise.

Note that, for a stochastic differential equation to have a unique global solution (i.e., no explosion in a finite time) for any given initial value, the coefficients of Equations 2.43-2.44 are generally required to satisfy both the linear growth condition and the local Lipschitz condition Revuz [229]. However, the coefficients of Equations 2.43-2.44 do not satisfy the linear growth condition, though they are locally Lipschitz continuous, so the solution may explode at a finite time. In Appendix 12 are reported some general results of system 2.43-2.44.

Estimation Methods We distinguish between two types of estimation problems. The first one consists in the estimation of parameter models given time observations of x_t, y_t. Following a general approach, we rewrite model equations 2.43-2.43 as a bivariate stochastic system:

$$dX_t = \mu(X_t;\theta)dt + \sigma(X_t;\theta)dW_t, \tag{2.45}$$

$$X_t = \begin{bmatrix} x_t \\ y_t \end{bmatrix}, \qquad W_t = \begin{bmatrix} W_t^{(1)} \\ W_t^{(2)} \end{bmatrix} \tag{2.46}$$

where the drift vector and the diffusion matrix take the following form:

$$\mu(X_t;\theta) = \begin{bmatrix} r x_t(1-x_t) - q x_t y_t \\ c q x_t y_t - u y_t \end{bmatrix}, \quad \sigma(X_t;\theta) = \begin{bmatrix} \varepsilon x_t \rho & \varepsilon \ {}_t x\sqrt{(1-\rho)^2} \\ \eta y_t & 0 \end{bmatrix}. \tag{2.47}$$

In order to estimate the parameter vector $\theta \in \mathbb{R}^6_+ \times [-1, 1]$ we use the likelihood function based on the logarithm of the Radon-Nikodym derivative Wang [230] of the transition distribution of X_t with respect to the Wiener measure, described as follows:

$$\begin{aligned} \log J(\theta) := \int_0^T & \mu^\top(X_t;\theta)\left[\sigma(X_t;\theta)\sigma^T(X_t;\theta)\right]^{-1} dX_t + \\ & -\frac{1}{2}\int_0^T \mu^\top(X_t;\theta)\left[\sigma(X_t;\theta)\sigma^T(X_t;\theta)\right]^{-1}\mu(X_t;\theta)dt. \end{aligned} \tag{2.48}$$

The maximum-likelihood estimator is

$$\widehat{\theta} = \text{argmax}_{\theta}\big\{\log J(\theta)\big\}, \tag{2.49}$$

and it provides the maximum likelihood parameters estimation of the model 2.45. In practice only discretely sampled data are available, namely $\widehat{X}_{t_0}$, $\widehat{X}_{t_1}, \dots,$ $\widehat{X}_{t_N}$, at times $t_k = t_0 + k\Delta t$, $k = 0, \dots, N$, $\Delta t > 0$. Hence, the minimum of the discretized minus log-likelihood $-\log J_N(\theta)$ is used. More precisely,

$$\begin{aligned} -\log J_N(\theta) := &- \sum_{k=0}^{N} \mu^\top(\widehat{X}_{t_k};\theta)\big[\sigma(\widehat{X}_{t_k};\theta)\sigma^T(\widehat{X}_{t_k};\theta)\big]^{-1}\Delta\widehat{X}_{t_k} + \\ &+ \frac{1}{2}\sum_{k=0}^{N} \mu^\top(\widehat{X}_{t_k};\theta)\big[\sigma(\widehat{X}_{t_k};\theta)\sigma^T(\widehat{X}_{t_k};\theta)\big]^{-1}\mu(\widehat{X}_{t_k};\theta)\Delta t. \end{aligned} \tag{2.50}$$

Of course the discretized version 2.50 of the log-likelihood is based on the application of the Euler-Maruyama discretization scheme of 2.45

$$X_{t_{k+1}} = X_{t_k} + \mu(X_{t_k}, \theta)\Delta t + \sigma(X_{t_k}, \theta)\big(W_{t_{k+1}} - W_{t_k}\big). \tag{2.51}$$

Therefore the likelihood estimator corresponds to

$$\widehat{\theta}_N = \text{argmax}_{\theta}\big\{-\log J_N(\theta)\big\}, \tag{2.52}$$

The accuracy of the estimator is improved as how much more the observations are numerous nearby, that is as the discretization step$\Delta\ t \to 0$. We refer the reader to Lipster [231] for a detailed and rigorous description of the estimation method described above.

From a different perspective, the variable x_t in model 2.43-2.44 can be used to describe both the positive inspections and the number of infractions undetected by GDF officers. Therefore, the model provides an indirect measure of tax evasion in the reference context. In this section we present the estimation procedure used in the case the variable x_t is unobservable. In this case we use a filtering technique based on a nonlinear version of the Kalman filter which linearizes about an estimate of the current mean and covariance Reif [232] and Kalman [177].

The estimation of the state variable x_t is obtained by applying an extended Kalman filter to a discrete approximation of system 2.43-2.44:

$$\hat{x}_t = \mathbb{E}\left[x_t \middle| y_s,\, 0 \le s \le t\right]. \tag{2.53}$$

Essentially, the Kalman filter is an analytical algorithm that recursively computes the first two moments of the distribution of the state variable conditional on observations up to that point. The algorithm consists of a prediction step when the current state is predicted, based on the last state, and an update step, when the prediction is latest information. For state space models that are linear in both state and measurement equations and Gaussian in both state transition and measurement densities, the predicted and updated densities are also Gaussian such that the algorithm is analytical and fast to execute.

We introduce a state-space notation for the system. Let us define $x_t^{(1)} = x_t$, $x_t^{(2)} = y_t$ and let Y_t be the (observable) state-measure process. Therefore we can rewrite system 2.43-2.44 as follows:

$$\begin{aligned} dx_t^{(1)} &= \big[rx_t^{(1)}(1-x_t^{(1)}) - qx_t^{(1)}x_t^{(2)}\big]dt + \varepsilon x_t^{(1)}\big[\rho dW_t^{(1)} + \sqrt{(1-\rho^2)}dW_t^{(2)}\big], \\ dx_t^{(2)} &= \big[cqx_t^{(1)}x_t^{(2)} - ux_t^{(2)}\big]dt + \eta x_t^{(2)}dW_t^{(1)}, \\ Y_t &= x_t^{(2)}. \end{aligned} \tag{2.54}$$

Unfortunately, system 2.54 is affected by nonlinearities with multiplicative noise that avoids the direct application of classical Kalman-based filters. However we can take advantage from the positivity of the solution established in Corollary 12.0.9 of Appendix 12, by considering the change of variable $\tilde{x}_t^{(i)} := \log x_t^{(i)}$, for $i = 1, 2$. Hence, by the application of the classical Ito's formula Revuz [229] we obtain the system

$$\begin{aligned} d\tilde{x}_t^{(1)} &= \Big[r\big(1-\varepsilon x^{\tilde{x}_t^{(1)}}\big) - q\varepsilon x^{\tilde{x}_t^{(1)}} - \frac{\varepsilon^2}{2}\Big]dt + \varepsilon\Big[\rho dW_t^{(1)} + \sqrt{(1-\rho^2)}\; dW_t^{(2)}\Big], \\ d\tilde{x}_t^{(2)} &= \Big[cq\varepsilon x^{\tilde{x}_t^{(1)}} - u - \frac{\eta^2}{2}\Big]dt + \eta dW_t^{(1)}, \\ \widetilde{Y}_t &= \tilde{x}_t^{(2)}. \end{aligned} \tag{2.55}$$

We apply the filtering technique to the discretized version of model 2.55

$$\begin{aligned} \widehat{X}_{k+1} &= f(\widehat{X}_k) + QN_k, \\ \widehat{Y}_t &= h(\widehat{X}_k), \end{aligned} \tag{2.56}$$

where $\widehat{X}_k = \Big[\hat{x}_k^{(1)}, \hat{x}_k^{(2)}\Big]^T$, e

$$f(\widehat{X}_k) = \begin{bmatrix} \hat{x}_k^{(1)} + \Big(r\big(1-\varepsilon x^{\hat{x}_k^{(1)}}\big) - q\varepsilon x^{\hat{x}_k^{(1)}} - \frac{\varepsilon^2}{2}\Big)\Delta t \\ \hat{x}_k^{(2)} + \Big(cq\varepsilon x^{\hat{x}_k^{(1)}} - u - \frac{\eta^2}{2}\Big)\Delta t \end{bmatrix}, \quad h(\widehat{X}_k) = \hat{x}_k^{(2)},$$

$$Q = \begin{bmatrix} \varepsilon\rho\sqrt{\Delta t} & \varepsilon\sqrt{(1-\rho^2)\Delta t} \\ \eta\sqrt{\Delta t} & 0 \end{bmatrix}, \quad N_k = \begin{bmatrix} N_k^{(1)} \\ N_k^{(2)} \end{bmatrix}.$$

In order to apply the EKF, we also need to the Jacobian matrices associated to $f(\widehat{X}_k)$ and $h(\widehat{X}_k)$, respectively given by:

$$J_f = \begin{bmatrix} 1 - r\varepsilon x^{\hat{x}_k^{(1)}}\Delta t & -q\varepsilon x^{\hat{x}_k^{(1)}}\Delta t \\ cq\varepsilon x^{\hat{x}_k^{(1)}}\Delta t & 1 \end{bmatrix}, \quad J_h = \begin{bmatrix} 0 & 1 \end{bmatrix}.$$

where $\hat{x}_k^{(i)}$ represents the approximation of $\tilde{x}_{t_k}^{(i)}$, at time $t_k = k\Delta t$, $k = 0, \dots, n-1$, $n > 1$, for $i = 1, 2$. The random terms $\{N_k^{(i)}\}_k$, for $i = 1, 2$, are independent standard normal variables. Even if the transformation increases the nonlinearity of the drift component, the noise takes an additive form, independent of the state variables. Given the previous representation, we can effectively apply an EKF to model 2.56 in order to deduce an estimate of the trajectory $\{\hat{x}_{t_k}^{(1)}\}_k$. Focusing also on the estimation of the interaction parameters in the model, q and r, we consider the extended state variable assigning additional components $\hat{x}_k^{(3)} = q$ and $\hat{x}_k^{(4)} = r$. In this case the structure of model 2.56 does not change. In fact, if the augmented state vector is $\widehat{X}_k := \left[\hat{x}_k^{(1)}, \hat{x}_k^{(2)}, \hat{x}_k^{(3)}, \hat{x}_k^{(4)}\right]^\top$, we can define

$$f(\widehat{X}_k) = \begin{bmatrix} \hat{x}_k^{(1)} + \left(\hat{x}_k^{(4)}\left(1 - \varepsilon x^{\hat{x}_k^{(1)}}\right) - \hat{x}_k^{(3)}\varepsilon x^{\hat{x}_k^{(1)}} - \frac{\varepsilon^2}{2}\right)\Delta t \\ \hat{x}_k^{(2)} + \left(c\hat{x}_k^{(3)}\varepsilon x^{\hat{x}_k^{(1)}} - u - \frac{\eta^2}{2}\right)\Delta t \\ \hat{x}_k^{(3)} \\ \hat{x}_k^{(4)} \end{bmatrix}, \quad h(\widehat{X}_k) = \hat{x}_k^{(2)},$$
$$Q = \begin{bmatrix} \varepsilon\rho\sqrt{\Delta t} & \varepsilon\sqrt{(1-\rho^2)\Delta t} \\ \eta\sqrt{\Delta t} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad N_k = \begin{bmatrix} N_k^{(1)} \\ N_k^{(2)} \end{bmatrix}. \tag{2.57}$$

In this case, the Jacobian matrices associated to $f(\widehat{X}_k)$ and $h(\widehat{X}_k)$ take the form

$$J_f = \begin{bmatrix} 1 - \hat{x}_k^{(4)}\varepsilon x^{\hat{x}_k^{(1)}}\Delta t & -\hat{x}_k^{(3)}\varepsilon x^{\hat{x}_k^{(1)}}\Delta t & -\varepsilon x^{\hat{x}_k^{(2)}}\Delta t & \left(1 - \varepsilon x^{\hat{x}_k^{(1)}}\right)\Delta t \\ c\hat{x}_k^{(3)}\varepsilon x^{\hat{x}_k^{(1)}}\Delta t & 1 & c\varepsilon x^{\hat{x}_k^{(1)}}\Delta t & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.58}$$
$$J_h = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} \tag{2.59}$$

Under the structure of model 2.56 it is possible to apply the EKF in order get an estimation of $\hat{x}_t^{(1)}$ and the parameters, conditional to a sample of discrete observations of $\widetilde{Y}_t$. For the sake of clarity, two simulated trajectories are presented in Figure 8.1, with parameters: $q = 4.092$, $r = 0.234$, $c = 0.43$, $u = 0.2$, $\varepsilon = 0.07$, $\eta = 0.05$, $\rho = 0.5$, with initial conditions$x_0^{(1)} = 0.1$, $x_0^{(2)} = 0.02$. Each trajectory consists of 700 points, with$\Delta\ t = 0.1$.

Moreover, in order to validate our method, we consider the percentage relative error, namely:

$$\epsilon_k^{(i)} = 100 \times \frac{\tilde{x}_k^{(i)} - \hat{x}_k^{(i)}}{\tilde{x}_k^{(i)}}, \tag{2.60}$$

for $i = 1, 2$ and initial condition $\hat{x}_0^{(1)} = \exp(\tilde{x}_0^{(1)}) = 0.018$, where $\hat{x}_k$ stands for the estimation of the state variable. The simulation results are presented in subsequent Figures 8.2-8.3-8.4-8.5-8.6. In the case of the augmented state system 2.57, we use the same set of parameters with initial condition $\hat{x}_0^{(1)} = \exp(\tilde{x}_0^{(1)}) = 0.018$, $\tilde{q} = \tilde{x}_0^{(3)} = 6$, $\tilde{r} = \tilde{x}_0^{(4)} = 3$. The simulation results show a substatial agreement with the simulated trajectories and the selected sample parameters. Since the model contains a large number of parameter models, we have concentrated our attention mainly on the estimation of q and r, given their crucial importance in the prey-predator model. Comparison

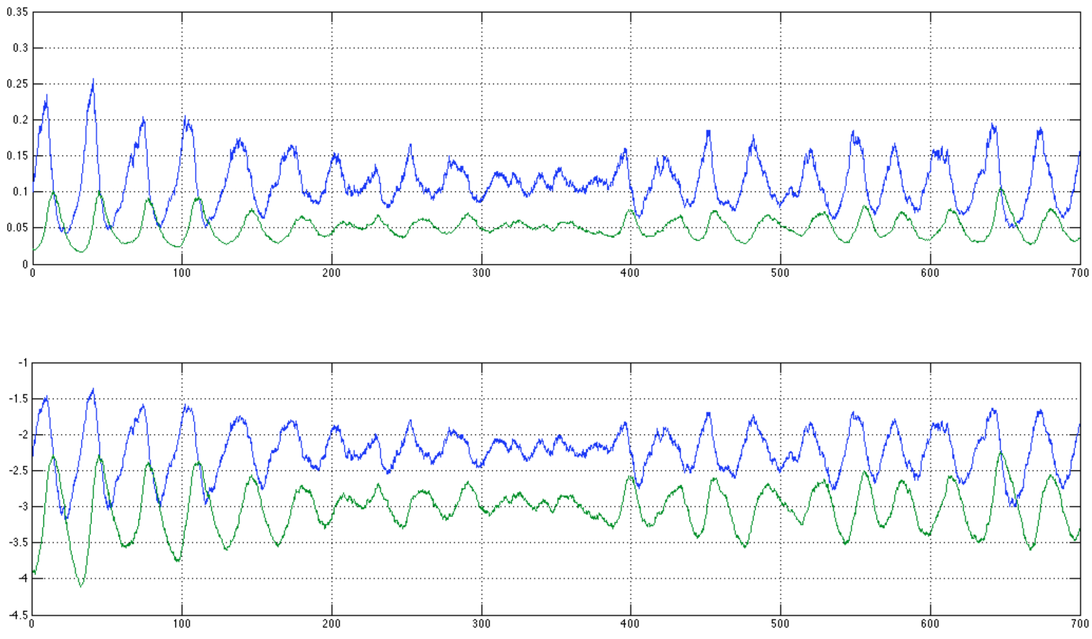


Figure 8.1: Estimated trajectories through the Lotka-Volterra model - The trajectories of processes $x_k^{(1)}$ (blue line), $x_k^{(2)}$ (green line) from model 2.54 (upper figure) and model 2.55 (*below figure*)

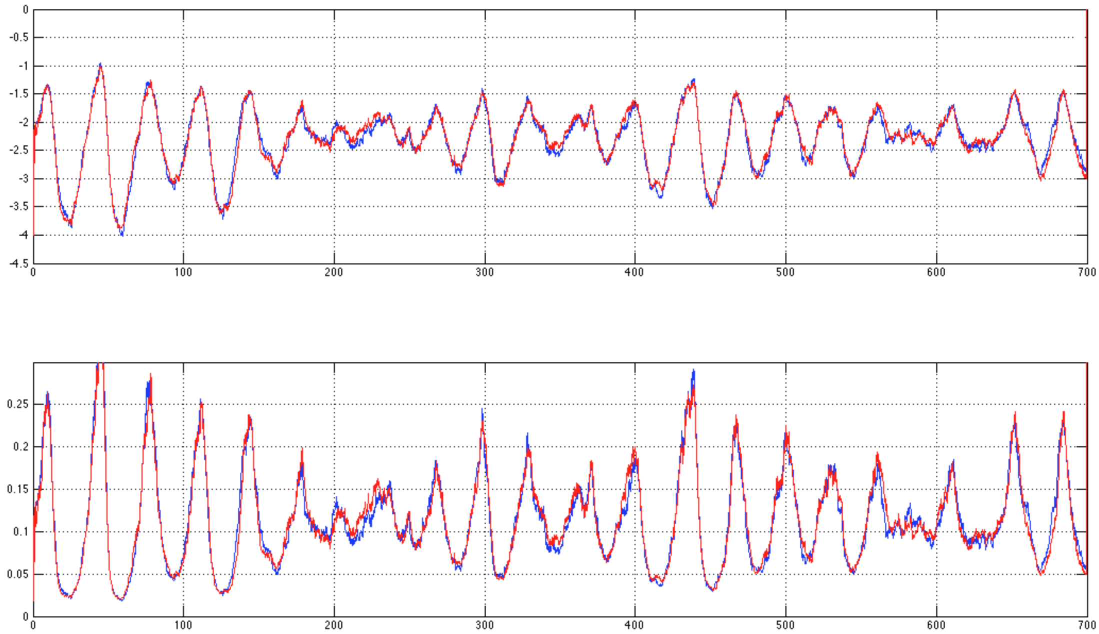


Figure 8.2: Trajectories comparison - Comparison between the simulated trajectories of $\tilde{x}_k^{(1)}$ (blue line) and its estimation $\hat{x}_k^{(1)}$ (red line) under the model 2.54 (*upper figure*) and under model 2.55 (*below figure*).

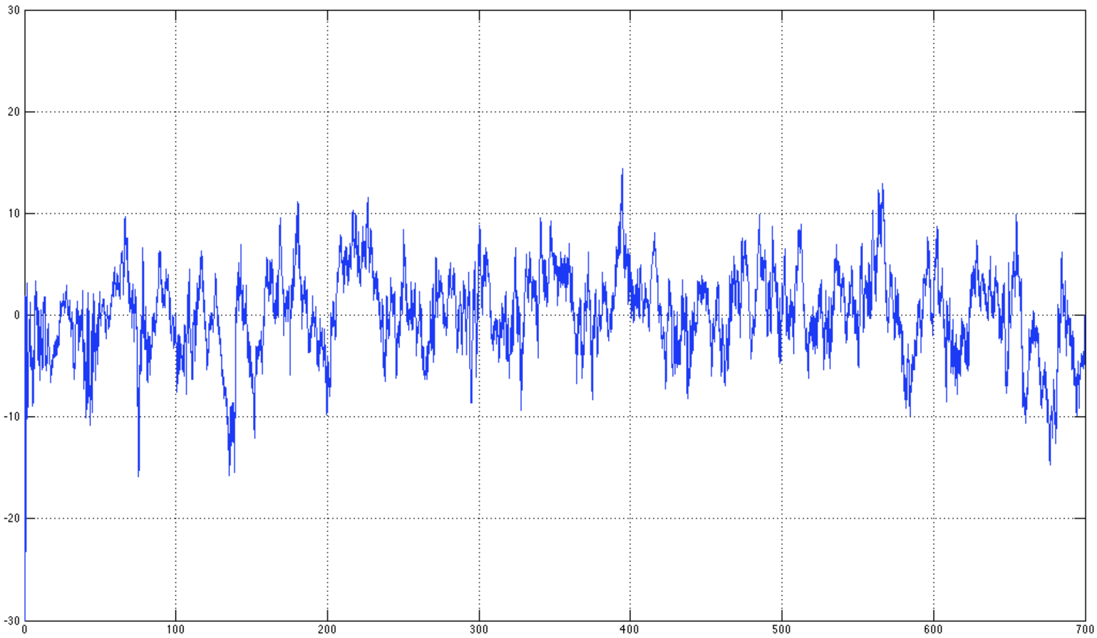


Figure 8.3: Error evaluation of state variable - Relative error estimation 2.60 for $x_k^{(1)}$ under model 2.55.

between the simulated trajectories of $\tilde{x}_k^{(1)}$ (blue line) and its estimation $\hat{x}_k^{(1)}$ (red line) under the augmented state model 2.54 (*upper figure*) and under the augmented state model 2.55 (*below figure*).

8.2.4 Estimation Results

In this study we consider anonymous data of GDF officers over all Italian regions and districts for the specific case of instrumental controls related to tax receipts infractions. We shall use historical data with daily frequency for a reference period from 2002 to 2014, provided by GDF information systems. Figure 8.7 shows the ratio between the positive control number and the number of hour used in prevention activities, data show that exist a big difference between macro-area.
In order to validate model 2.43 the common probability test are adopted to GDF instrumental control historical data. Table 8.1 reports the estimated parameters and Figures 8.8 and 8.9 show the *Q-Q plot* of noises $N_k^{(1)}$ and $N_k^{(2)}$. Furthermore in order to validate the Gaussian distribution noise hypothesis the results of *Kolmogorov-Smirnov test* and *Jarque-Bera test* are provided.

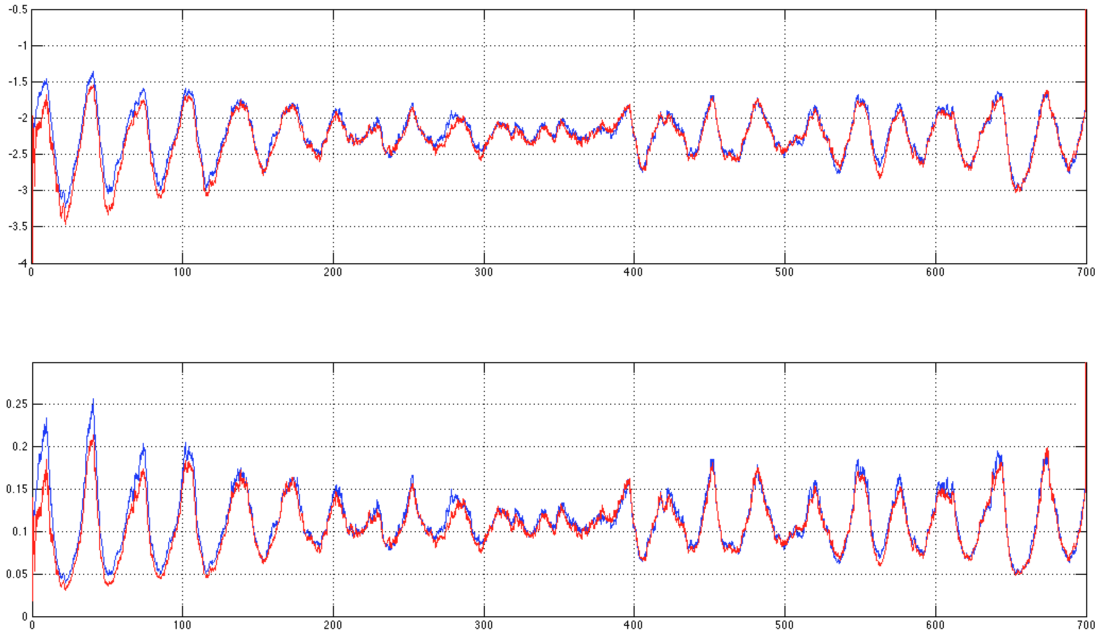


Figure 8.4: Trajectories comparison (augmented state model) - Comparison between the simulated trajectories of $\tilde{x}_k^{(1)}$ (blue line) and its estimation $\hat{x}_k^{(1)}$ (red line) under the augmented state model 2.54 (*upper figure*) and under the augmented state model 2.55 (*below figure*).

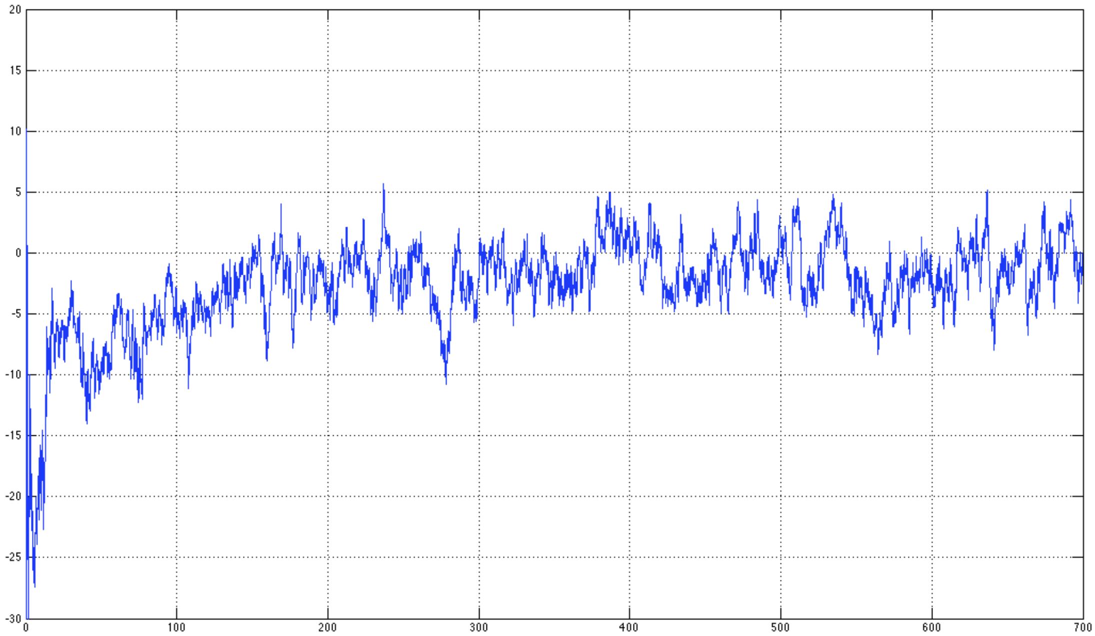


Figure 8.5: Estimation error for $x_k^{(1)}$ - Derived from the application of the filtering technique for the augmented state model.

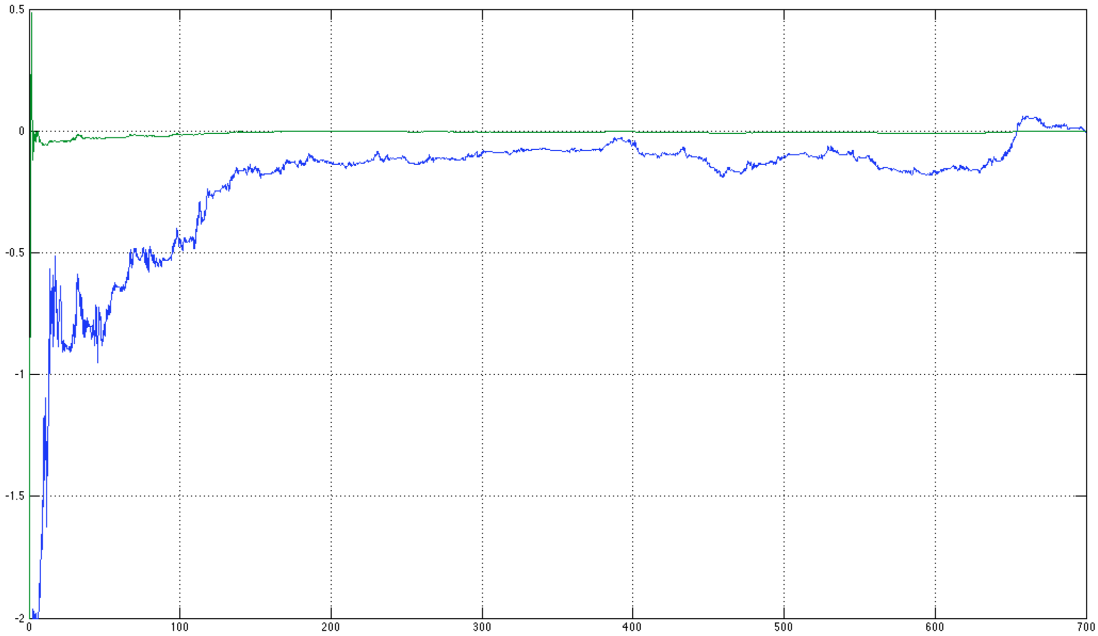


Figure 8.6: Estimation error for the unknown parameters q **and** r - Derived from the application of the filtering technique for the augmented state model.

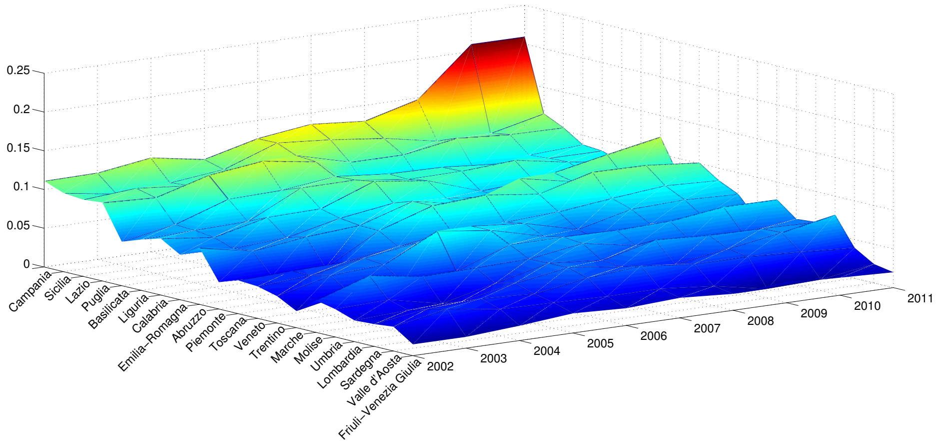


Figure 8.7: Regional analysis of the control activities efficiency - z-axes represent the ratio between the positive control number and the number of hour used in prevention activities.

	Estimated parameters		
q	0.0023	ε	0.3261
r	1.6631	η	0.2200
c	170.8702	ρ	0.0000
u	0.2233		

Table 8.1: Parameters estimation from the historical time series. Estimation is made on GDF data.

A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set[1]. If we assume that the noises are Normally distributed, the points in the Q-Q plot will approximately lie on the diagonal of the plot. Although a Q-Q plot is based on quantiles, in a standard Q-Q plot it is not possible to determine which point in the Q-Q plot determines a given quantile but allows to compare the quantile of two distribution. Figures 8.8 and 8.9 show that noise $N_k^{(2)}$ is Normally distributed otherwise $N_k^{(1)}$ has evident deviations from the Normal distribution.

[1]Quantile is the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. If we consider a continuous distributions with cumulative density function F, $\alpha-$quantile is $F(q_\alpha) = \alpha$.

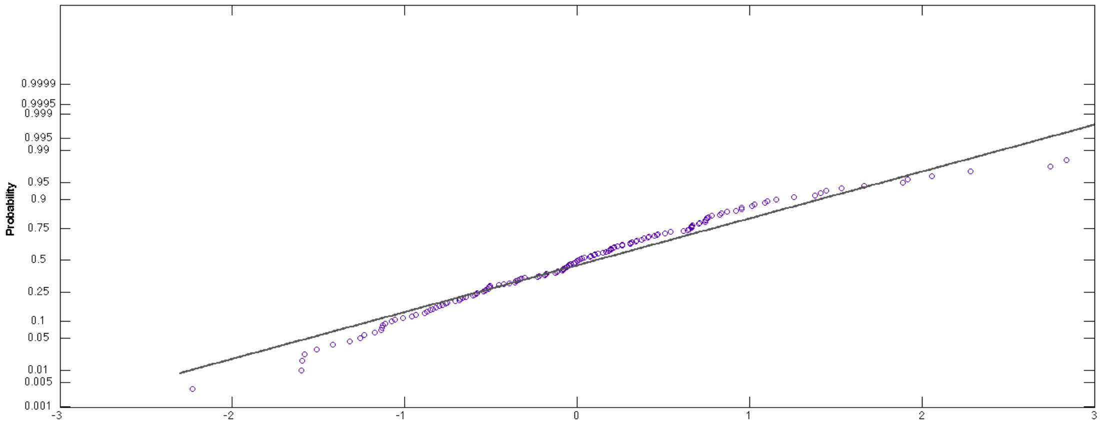


Figure 8.8: Noise $N_k^{(1)}$ probability plot - *Q-Q plot*

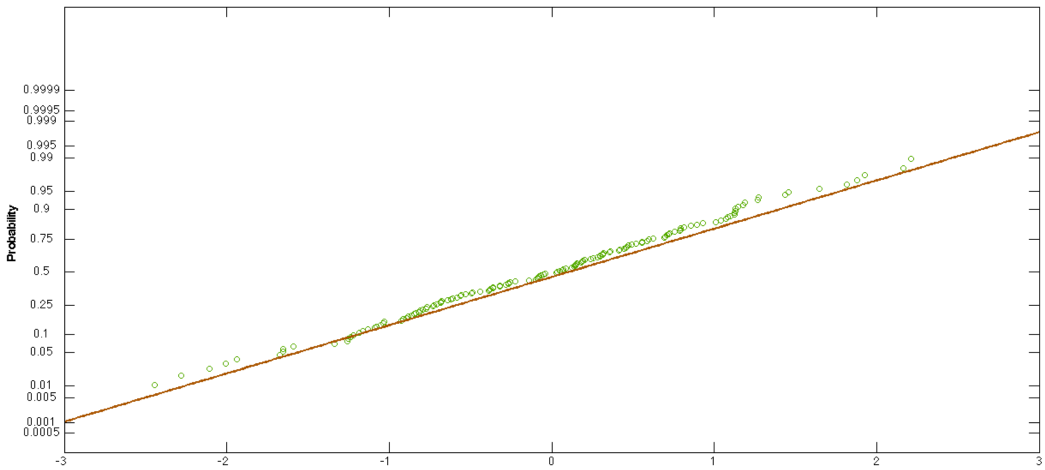


Figure 8.9: Noise $N_k^{(2)}$ probability plot - *Q-Q plot*

	$N_k^{(1)}$	$N_k^{(2)}$
KS	0.0602	0.0426
JB	139.7487	2.7505

Table 8.2: Kolmogorov-Smirnov (KS) and Jarque-Bera (JB) test results on GDF data.

Goodness-of-fit tests and inspection activity dataset In order to evaluate the deviation from the Normal distribution fitting test are proposed. The Kolmogorov-Smirnov (KS) test is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution, or to compare two samples. The Kolmogorov-Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples.
Noise vectors $N_k^{(1)}$ and $N_k^{(2)}$ are compared with a Normal distribution vector. If the test result is 1 we reject the hypothesis that data are Normally distributed, otherwise we can not reject the Normally distributed hypothesis.
Jarque-Bera (JB) test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution.The test statistic JB is defined as

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right),$$

where n is the observation number, S is the *skewness*, and K is the kurtosis. Table 8.2 reports the test results. Despite the few number of observed data, results show that the model proposed, in particular for $N_k^{(1)}$, describes only partially the dynamic analysed. We propose the forecast of May 2014 data, in order to obtain the forecast the model is calibrated using data from 2002 to April 2014. For each regions we apply the *prey-predator* model and our observations $\{\hat{y}_{t_k}\}_k$, $\{\hat{x}_{t_k}\}_k$, are represented by the time series of total (positive and negative) inspections conducted at time $t_k = k\Delta t$, and the fraction between the number of positive inspections over the total number of inspections at time t_k, respectively. We perform an ex-post analysis in order to verify the goodness-of-fit of our model, through an investigation of its ability to predict the future value of positive inspections, given past observations:

$$\hat{x}_{t_{k+1}} = \mathbb{E}[x_{t_{k+1}}|x_{t_k}, y_{t_k}]. \tag{2.61}$$

Areas	Empirical	Empirical positive	Estimated positive	Positive error estimation
MI	200	66	77	+5.50%
FI	58	15	12	−5.17%
NA	69	46	57	+15.94%
RM	327	209	210	+0.31%
PA	44	12	21	+20.45%
TO	233	168	107	−26.18%
CS	39	10	13	+7.69%
VE	27	5	8	+11.11%
MC	13	1	2	+7.69%
GE	58	34	30	−6.90%

Table 8.3: Model forecast validation. The estimated number of positive controls is compared with the number of irregularities found by GDF in each area.

To this end, we have selected few representative districts inside the macro-regions, that are differentiated by their demographic dimension, gross domestic product and the density of economic and business activities. Since we have daily data, we chosen to set the year as the unit time, hence the discretization step$\Delta\ t$ is fixed as 1/30. For each month m, available in the historical time series of inspections, and for each day d in the month m, we estimated the model for every district using past observations up to time t_m^d, which corresponds to end day of month m; then we have computed the simple average of the conditional expectations 2.61 for the days between t_m^{month} and t_{m+1}^{month} and we used it as the prediction. Table 8.3 shows the main results obtained where **Empirical** represents the overall control number, **Empirical positive** is the number of positive number found, **Estimated positive** is the number of positive control estimated using the model, and **Positive error estimation** is the error metric define in 2.60. In some areas the estimation provide is very close to empirical data. In Figure 8.10 are reported positive control number time series of the Rome district, according to this series the model estimates 210 positive controls during May 2014. The empirical number of positive control recorded in that period is 209.

8.3 Summary

In order to provide guidelines to optimize the financial inspection resource consumption is analysed the mutual interaction between instrumental controls carried out by

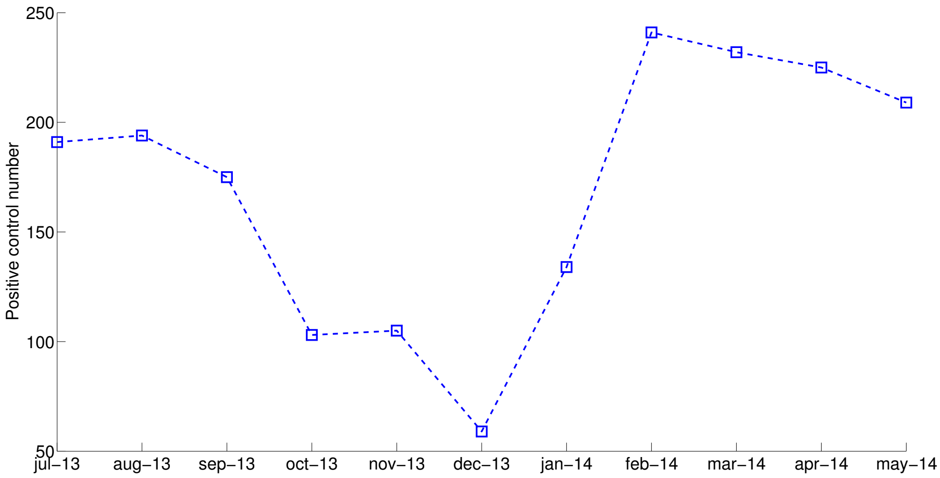


Figure 8.10: Empirical number of positive control time trend - Rome district data recorded from July 2013 to May 2014

Financial Guard and tax evasion. The positive control number oscillations is described using a stochastic version of the Lotka-Volterra model where the predators represent the control performed by Financial Guard. The stochastic dynamic model proposed shows a good ability to determine the irregular control number within one month with respect to the time series analysed. Reported results show that the proposed approach produces a good fit of empirical distribution and provides also a phenomenological interpretation of the analysed problem.

9

Volatility model

This Chapter proposes a part of a research project in cooperation with Enel a new stochastic volatility model for the calibration of option prices is developed [233]. In order to avoid the estimation of the initial volatility, a weighted average formulation for the Heston stochastic volatility option price is presented. This approach has been developed in the literature for the estimation of the distribution of stock price changes (returns), showing an excellent agreement with real market data. This method is extended to the calibration of option prices considering a large class of probability distributions assumed for the initial volatility parameter. The estimation error is shown to be less than the case of the simple pricing formula. Our results are also validated with a numerical comparison on observed call prices, between the proposed calibration method and the classical approach.

Over the last decade stochastic volatility models (SV) have become an industry standard for option pricing. The most popular stochastic volatility models are, among others, the models by [234], [235] and [236]. Nowadays several financial institutions have incorporated them into their front office systems. Also the literature on SV models is rapidly increasing and demonstrates the ongoing interest in such models ([237]; [238], [239], [240], [241]). We also refer to [242] or [243] for general surveys. More generally, SV models including Lévy jumps have also been proposed with the aim of increasing the smile effect in the short end. The drawback of SV models is that the more realistic dynamics comes at the cost of an additional theoretical complexity and a greater difficulty in the numerical solution of the pricing problem and model calibration. For instance, a typical calibration of a SV model requires efficient numerical schemes for the solution of a two-dimensional partial differential equation. Moreover, while returns are readily known from financial time series data, the volatility is not a tradable asset, so it acts as a hidden stochastic variable. Therefore, pricing under SV models involves an additional volatility risk term. The purpose of this work is to present a new approach in the calibration of SV models in order to provide an efficient approximation of observed plain vanilla options. We consider the simple stochastic volatility model proposed by [236]. The choice of such a model is motivated by the fact that it has a closed-form expression for the characteristic function of its transitional probability density function from which options can be efficiently priced; a feature of the Heston model that has received considerable attention in the literature. The Heston model is the most popular one because of its three main features: it does not allow negative volatilities, it allows the correlation between asset returns and volatility, and it has a closed-form pricing formula. We follow closely the approach presented by [244]. Using the Fourier and Laplace transforms, the authors solve the Fokker-Planck equation for the Heston model exactly and find the joint density function of log-returns and variance as a function of time, conditional on the initial volatility (v_0). Thus, they integrate the joint density function over the initial volatility and obtain the marginal density function of log-returns, unconditional with respect to the initial variance. The approximated probability density function, found in [244], provides an excellent agreement with observed historical financial data. The intuition consists in supposing that the initial volatility is a random variable distributed according to the stationary distribution of the volatility process. This stationary distribution is a Gamma distribution with coefficients depending on the Heston model parameters. Our estimation method extends, with rigorous arguments, this technique to the option pricing problem. We also derive a representation formula for the price of a call option similar to that of the celebrated Heston closed-form solution, without increasing the computational complexity required for the evaluation. Furthermore, we prove that, for a large class of probability distributions assumed for the initial volatility parameter, the estimation error in the calibration procedure of option prices is less than the case of the simple pricing formula. Our results are validated with a numerical comparison, on observed call prices, between the proposed calibration method and the classical approach. It should be stressed that the Heston model is used only as a specific example to allow our methodology to be fully developed. Our technique itself is not limited to any particular model and the extension to other models, eventually involving jumps, is a matter of detail alone and requires no further significant conceptual development. The theoretical framework needed to show that our pricing formula fullfills a no-arbitrage principle, is possible even though further research is required. We aim to come back to this and other related topics in a forthcoming paper.

9.1 Model

Options are usually priced under the risk-neutral measure and incorporate a volatility premium. Let $\{W_t^1\}_{t\geq 0}$, $\{W_t^2\}_{t\geq 0}$ be two independent Brownian motions on the probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{Q})$, where $\mathbb{Q}$ is the risk-neutral probability measure. The Heston model (1993) assumes risk-neutral dynamics of the form:

$$\begin{aligned} dS_t &= rS_t dt + S_t\sqrt{v_t}dW_t^1, \\ dv_t &= \kappa(\theta - v_t)dt + \sigma\sqrt{v_t}\left(\rho dW_t^1 + \sqrt{1-\rho^2}dW_t^2\right), \end{aligned} \tag{1.1}$$

with $S_0,\ v_0 > 0$ and constant parameters $r \in \mathbb{R}$, $\kappa, \theta, \sigma > 0$, $\rho \in [-1, 1]$. S is the stock price, v is the state variable driving volatility, and $r > 0$ is the risk-free interest rate. From the theory of Bessel processes [229], we impose the condition $2\kappa\theta \geq \sigma^2$ to ensure that the volatility is strictly positive in finite time. One of the reasons for the popularity of the Heston model is that it provides a closed-form solution for pricing vanilla options. This is of great benefit in particular when calibrating against market prices. The call t-time price of the European call with strike K and maturity T is the expected discounted value under the risk-neutral measure $\mathbb{Q}$, namely:

$$\begin{aligned} C_t &= e^{-r\tau}\mathbb{E}_t^{\mathbb{Q}}[(S_T - K)^+] \\ &= e^{-r\tau}\mathbb{E}_t^{\mathbb{Q}}[S_T \mathbb{1}_{S_T>K}] - e^{-r\tau}K\mathbb{E}_t^{\mathbb{Q}}[\mathbb{1}_{S_T>K}], \end{aligned} \tag{1.2}$$

where $\mathbb{E}_t^{\mathbb{Q}}[\cdot]$ denotes the conditional $\mathbb{Q}$-expected value, given $\mathcal{F}_t$. By analogy with the Black-Scholes formula, the guessed solution of this European option is of the form $C_t = C^H(t, S_t, V_t)$, where the deterministic function C^H takes the form

$$C^H(t, S, v) = SP_1(T - t, \log(S), v) - e^{-r(T-t)}KP_2(T - t, \log(S), v), \tag{1.3}$$

for any $S,\ v > 0$ and $0 \leq t \leq T$. The function $P_j(\tau, x, v)$, defined for $\tau > 0$, $x \in \mathbb{R}$, $v > 0$, represents the probability (under suitable probability measures on $(\Omega, \mathcal{F}_T)$) of the call expiring in-the-money, conditional on the value $x_t = \log(S_t)$ of the stock and on the value v_t of the volatility at time t. In the Heston model, the expression for these probabilities are given by

$$P_j(\tau, x, v) = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \mathrm{Re}\left[\frac{e^{-\imath\phi\log(K)} f_j(\phi; x, v)}{\imath\phi}\right] d\phi, \tag{1.4}$$

with

$$f_j(\phi; x, v) = \exp(C_j(\tau,\phi) + D_j(\tau,\phi)v + \imath\phi x), \tag{1.5}$$

$$D_j(\tau,\phi) = \frac{b_j - \rho\sigma\imath\phi + d_j}{\sigma^2}\left(\frac{1 - e^{d_j\tau}}{1 - g_j e^{d_j\tau}}\right), \tag{1.6}$$

$$C_j(\tau,\phi) = r\imath\phi\tau + \frac{a}{\sigma^2}\left[(b_j - \rho\sigma\imath\phi + d_j)\tau - 2\log\left(\frac{1 - g_j e^{d_j\tau}}{1 - g_j}\right)\right], \tag{1.7}$$

$$d_j = \sqrt{(b_j - \rho\sigma\imath\phi)^2 - \sigma^2(2u_j\imath\phi - \phi^2)},$$

$$g_j = \frac{b_j - \rho\sigma\imath\phi + d_j}{b_j - \rho\sigma\imath\phi - d_j},$$

for $j = 1, 2$. where $u_1 = \frac{1}{2}$, $u_2 = -\frac{1}{2}$, $a = \kappa\theta$, $b_1 = \kappa - \rho\sigma$, $b_2 = \kappa$.

The results obtained in [245] for a general class of pricing problems based on reflecting diffusion processes with jumps (including also the Heston model) allow to state that $C^H(t, S, v)$ is the unique viscosity solution, in the sense of [246] of the Dirichlet problem

$$\begin{cases} \mathcal{L}\, C^H(t, S, V) = 0, & (t, s, v) \in (0, T) \times (0, \infty)^2, \\ C^H(T, S, v) = (S - K)^+ & (s, v) \in (0, \infty)^2, \end{cases} \tag{1.8}$$

where $\mathcal{L}$ is the differential operator

$$\mathcal{L} = -r + \frac{\partial}{\partial t} + rS\frac{\partial}{\partial S} + \kappa(\theta - v)\frac{\partial}{\partial v} + \frac{1}{2}S^2 v\frac{\partial^2}{\partial S^2} + \frac{1}{2}\sigma^2 v\frac{\partial^2}{\partial v^2} + \rho\sigma S v\frac{\partial}{\partial S\partial v}. \tag{1.9}$$

The results of [247] in (2010) and, more recently, the contribution of [248] (see Theorem 9.2.1) ensure the following properties:

1) $C^H \in C(([0, T) \times [0, \infty)^2) \cap C^{1,0,1}([0, T) \times (0, \infty) \times [0, \infty))$.

2) $C^H \in C^{1,2,2}([0, T) \times (0, \infty)^2)$.

3) For every $t \in [0, T)$, $v > 0$, the function $S \mapsto C^H(t, S, v)$ is increasing and strictly convex on $(0, \infty)$;

4) For every $t \in [0, T)$, $S > 0$, the function $v \mapsto C^H(t, S, v)$ is strictly increasing.

The Fourier pricing setup introduced by [249] generalizes previous works on Fourier transform methods (see e.g. [250]). In the case of a call option, the price reduces to the following formula:

$$C_0 \quad = \quad S_0 - \frac{\sqrt{S_0 K e^{-rT}}}{\pi} \int_0^{+\infty} \mathrm{Re}\Big[e^{\imath uk} \varphi_T\left(u - \frac{\imath}{2}\right)\Big] \frac{du}{u^2 + \frac{1}{4}}. \tag{1.10}$$

In the Heston model, the expression for the characteristic function under the measure $\mathbb{Q}$ is

$$\varphi_T(u; x, v) = e^{-\imath u(x+rT)} f_2(u; x, v) = \exp(C_2(T, u) + D_2(T, u)v - \imath urT), \tag{1.11}$$

for any $u \in (0, \infty)$, $x \in \mathbb{R}$, $v > 0$. Here the functions $C_2(\cdot, \cdot)$ and $D_2(\cdot, \cdot)$ are given by 1.6-1.7. Plugging this expression in 1.10, an alternative formula for the price of the European call option with maturity T and strike K at time $t = 0$ can be easily derived. We will use this reduced formula in next sections.

9.2 The calibration problem

The calibration of SV models to synthetic and market option data forms one of the major theme in the literature. Calibrating methods to market data (either option prices or implied volatilities) allows to infer the (risk-neutral) market parameters for the different models and thus to use these models for pricing and hedging purposes.
The cost of using such models, however, is that the calibration and pricing techniques that must be employed are usually quite onerous. The choice of a calibrating routine requires a trade-off between its computational complexity and its accuracy. This leads to a complication that plagues SV models in general. A common solution is to find those parameters which produce the correct market prices of vanilla options. This is called an inverse problem, as we solve for the parameters indirectly through some implied structure. A well documented and popular method of fitting pricing models to observed data is to find a set of model parameter values that minimizes the square of the differences between the empirical values and the corresponding model values. More

specifically, the squared differences between vanilla option market prices and model theoretical prices are minimized over the parameter space:

$$\inf_{\Theta} \sum_{i=1}^{N} w_i \Big(C^{Model}(S_0, K_i, T_i; \Theta) - C_i^{Market}(K_i, T_i) \Big)^2. \tag{2.12}$$

where Θ is the vector of parameter values, $C^{Model}(S_0, T_i, K_i; \Theta)$ and $C_i^{Market}(K_i, T_i)$ denote the i^{th} option price from the model and market dataset, respectively, with strike K_i and maturity T_i, whereas N is the number of options used for calibration. The coefficients w_i, $i = 1, \dots, N$, denote suitable weights; their choice will be discussed later. Note that we could also do this for model and market implied volatilities, however, this adds to the complexity of the calibration routine ([251] and [252]). The minimization above is therefore an inverse, ill-posed problem, making the choice of an optimization algorithm tricky. In general, the objective function (F) is neither convex nor does it have any particular structure. This poses some complications:

- Finding the minimum of F is not as simple as finding those parameter values that make the gradient of F zero. The function might also have many local minima, making purely gradient based schemes ineffective and necessitating a careful choice of initial calibration parameters.
- Finding a global minimum is difficult (and very dependent on the optimization method used). As a result, we have to choose carefully between using a local or a global optimization routine. Global optimization schemes tend to be less sensitive to initial parameter estimates than local ones and should handle complicated objective functions better. They usually take longer to converge to a solution however.
- Unique solutions to 2.12 need not necessarily exist, in which case only local minima can be found. This has some implications regarding the stationarity of parameter values which are important in these types of models.

There are several alternative calibration methods that have been experimented. For instance, the *regularization method* involves adding a penalty function, $p = p(\Theta)$, to 2.12 such that the objective function

$$\sum_{i=1}^{N} w_i \Big(C^{Model}(S_0, T_i, K_i; \Theta) - C_i^{Market}(K_i, T_i) \Big)^2 + \alpha p(\Theta) \tag{2.13}$$

is convex. The parameter α is a regularization parameter. The underlying idea is to find an approximation which is as close to the true solution as possible. To achieve this, the problem is replaced with one which is close to the original, but does not possess the ill conditioning which the makes the original intractable. For a detailed discussion, we refer to [253]. When applied to a given set of market prices, these methods yield a single set of model parameters calibrated to the market but also require the extra step of determining the regularization parameter α, see for instance [254].

Another important consideration is the choice of weights w_i, for $i = 1, \ldots, N$. One possible choice is to set $w_i = 1/N$, for all $i = 1, \ldots, N$, making equation 2.12 a measure of mean squared errors [252]. Alternatively, we could let $w_i = |bid_i - ask_i|^{-1}$, where bid_i and ask_i stand for the bid and ask prices of the i^{th} option in the dataset. This would allow us to place more weight on options which are more liquid in the market. A third option that has also been suggested is to use the implied volatilities of the sampled options as weights, a method explored by [254].

9.2.1 Averaging over volatility

In the Heston model there are essentially four (risk-neutral) parameters that need estimation: $\kappa > 0$, $\theta > 0$, $\sigma > 0$ and $\rho \in [-1, 1]$. Nevertheless, in SV models the log-returns $\{x_t = \log(S_t)\}_{t \geq 0}$ are directly known from financial time series, whereas the volatility is a hidden variable that has to be estimated. Inevitably, such an estimation is done with some degree of uncertainty, which precludes a clear-cut direct comparison between model prices and financial data. In fact, several research contributions have shown that the implied parameters that produce the correct vanilla option prices and their time-series estimate counterparts are different [255]. The common approach adopted to overcome this estimation problem, is considering the initial volatility $v_0 > 0$ as an additional parameter in the calibration procedure. An alternative approach can be perfomed with at-the-money (ATM) implied variance, based on the following result from [256].

Theorem 9.2.1. (Term structure of the Black-Scholes implied volatility in the Heston Model)

$$\sigma^2_{ATM} \approx \frac{1}{T}\int_0^T [(v_0 - \theta')e^{-\kappa' t} + \theta]dt = (v_0 - \theta')\frac{1 - e^{-\kappa' T}}{\kappa' T} + \theta', \tag{2.14}$$

where $\kappa' = \kappa - \frac{1}{2}\rho\sigma$ and $\theta' = \kappa\theta/\kappa'$. The ATM Black-Scholes implied variance in the Heston model converges (in probability) to v_0, as $T \to 0$.

The practical significance of the previous theorem is that, if we assume that the stock process follows the Heston dynamics, then v_0 should be consistent with the short dated at-the-money volatility: there is a linear relationship between the initial variance, v_0, and the Black-Scholes implied variance returned by the Heston model. This estimation method has been considered as a satisfactory estimate for the initial variance v_0, in the Heston model. An interesting approach has been proposed by [244] in the estimation of the Heston model on time series of log-returns. Using the Fourier and Laplace transforms, the authors solve the Fokker-Planck equation exactly and they find the joint probability density function of log-returns and volatility as a function of time, conditional on the initial value v_0 of the volatility. Thus, they integrate the joint density function over the volatility parameter v_0 and obtain a proxy for the unconditional marginal density function of log-returns. The latter density function is then directly compared with financial data, with an excellent agreement between the results and market data (over a 20 year period of time). The Fokker-Planck equation for the transition density of the volatility admits a stationary solution given by the density of the gamma distribution, see [80]:

$$\Pi_\star(v_0) = \frac{\alpha^\alpha}{\Gamma(\alpha)} \frac{v_0^{\alpha-1}}{\theta^\alpha} e^{-\alpha v_0/\theta}, \tag{2.15}$$

where $\alpha = 2\kappa\theta/\sigma^2$ is the ratio of the average volatility θ to the characteristic fluctuation of variance $\sigma^2/2\kappa$ during the relaxation time $1/\kappa$. In [244], the unconditional density function of log-returns is approximated by averaging the conditional density function over v_0 with the weight$\Pi_\star(v_0)$. However there is neither a reasonable mathematical explanation nor a financial motivation for the use of such a distribution. We propose a method of calibration of the Heston model which resume and extend the argument followed in [244] and, moreover, we give a mathematical and numerical justification of our approach.

In the following, we will consider the set $\mathcal{P}$ of all non-negative Lebesgue-integrable functions $f : \mathbb{R} \longrightarrow \mathbb{R}$ such that $f(v) = 0$ for any $v \leq 0$, almost everywhere (*a.e.*), and

$$\int_0^\infty f(v)\, dv = 1. \tag{2.16}$$

Clearly $\mathcal{P}$ is a subset of all probability density functions on $\mathbb{R}$. For every $f \in \mathcal{P}$ and any bounded measurable function $\varphi : (0, \infty) \to \mathbb{R}$, let define

$$\mathbb{E}_f[\varphi] := \int_0^\infty \varphi(v) f(v)\, dv. \tag{2.17}$$

Moreover, we shall denote with M_f the (extended) moment generating function of f, namely

$$M_f(z) = \int_0^\infty e^{zv} f(v)\, dv. \tag{2.18}$$

Clearly $M_f(z)$ is well defined for all $z \in \mathbb{C}$, with $\mathrm{Re}(z) \leq 0$, and it coincides with the *generalized* Fourier transform of f at $-\imath\, z$.

Let the actual price of a call option with maturity T and strike K - in the framework of the Heston model - be denoted as $C^H(S_0, v_0, T, K; \Theta)$, where

$$\Theta \in \mathcal{H} := \left\{(\kappa, \theta, \sigma, \rho) \in \mathbb{R}^4 \ : \ \kappa, \theta, \sigma > \ 0, \quad \rho \in [-1, 1]\right\}. \tag{2.19}$$

Note that $\mathcal{H}$ is a convex subspace of $\mathbb{R}^4$. In fact

$$\begin{aligned} & 2(\lambda\kappa_1 + (1-\lambda)\kappa_2)(\lambda\theta_1 + (1-\lambda)\theta_2) \geq \\ & \quad \geq \lambda^2(2\kappa_1\theta_1) + (1-\lambda)^2(2\kappa_2\theta_2) + 2\lambda(1-\lambda)(\kappa_1\theta_2 + \kappa_2\theta_1) \\ & \quad \geq \lambda^2\sigma_1^2 + (1-\lambda)^2\sigma_2^2 + 2\lambda(1-\lambda)\sqrt{(2\kappa_1\theta_1)(2\kappa_2\theta_2)} \\ & \quad \geq \lambda^2\sigma_1^2 + (1-\lambda)^2\sigma_2^2 + 2\lambda(1-\lambda)\sigma_1\sigma_2 = (\lambda\sigma_1 + (1-\lambda)\sigma_2)^2 \end{aligned} \tag{2.20}$$

and, obviously, $\lambda\rho_1 + (1-\lambda)\rho_2 \in [-1, 1]$, for any $(k_i, \theta_i, \sigma_i, \rho_i) \in \mathcal{H}$, $i = 1, 2$, $0 \leq \lambda \leq 1$.

By *averaging over volatility*, we mean the option price functional given by

$$C_f^H(S_0, T, K; \Theta) := \mathbb{E}_f\left[C^H(0, S_0, \cdot, T, K; \Theta)\right], \tag{2.21}$$

for every $S_0 > 0$, $T > 0$, $K > 0, \Theta \ \in \mathcal{H}$, $f \in \mathcal{P}$.

We observe that if f is replaced with the Dirac delta function $\delta(v - v_0)$ centered at $v_0 > 0$, then 2.21 reduces to $C^H(S_0, v_0, T, K; \Theta)$. Thus, the case of a constant initial volatility can be loosely seen as a special case of considering v_0 as a random variable distributed according with the density $f \in \mathcal{P}$. We wonder if the use of 2.21 as an approximating call price for the Heston model, under a suitable probability density function f, can lead to an improvement of the calibration results of the model against the observed option prices. Let $C_i^M = C_i^{Market}(K_i, T_i)$, $i = 1, \ldots, N$ be a basket of call prices, all written on the same underlying asset with price S_0, at time $t = 0$. Let $\{w_i \geq 0, \text{ for } i = 1, \dots, N\}$ be a set of given weights. For the calibration purpose, we consider two objective functionals $J : \mathcal{H} \times (0, \infty) \to [0, \infty)$ and $J' : \mathcal{H} \times \mathcal{P} \to [0, \infty)$, defined as follows:

$$J(\Theta, v_0) \quad := \quad \sum_{i=1}^{N} w_i \, |C_i^M - C^H(0, S_0, v_0, T_i, K_i; \Theta\,)|^2, \tag{2.22}$$

$$J'(\Theta, f) \quad := \quad \sum_{i=1}^{N} w_i \, |C_i^M - C_f^H(S_0, T_i, K_i; \Theta\,)|^2. \tag{2.23}$$

Let $\mathcal{P}'$ be a non-empty subset of $\mathcal{P}$, then define I, I' be the infimum of J over $\mathcal{H} \times (0, \infty)$ and the infimum of J' over $\mathcal{H} \times \mathcal{P}'$, respectively. The following result states that if $\mathcal{P}'$ includes a sequence of densities weakly converging to the Dirac delta centered at an arbitrary $v_0 > 0$, then the calibration obtained through the averaged call price 2.21 improves the calibration result.

Theorem 9.2.2. *For every $f \in \mathcal{P}$, the integral in 2.21 is bounded. If $\mathcal{P}' \subseteq \mathcal{P}$ is such that for every $v_0 > 0$, there exists a sequence $\{f_n\}_n \subseteq \mathcal{P}'$ satisfying*

$$\lim_{n \to \infty} \int_{\mathbb{R}} f_n(v) g(v) \, dv = g(v_0), \tag{2.24}$$

for all bounded, continuous functions $g : \mathbb{R} \to \mathbb{R}$, then $I' \leq I$. Furthermore

$$\inf_{f \in \mathcal{P}'} C_f^H(S_0, T, K; \Theta\,) < C^H(0, S_0, v_0, T, K; \Theta\,) < \sup_{f \in \mathcal{P}'} C_f^H(S_0, T, K; \Theta\,) \tag{2.25}$$

for any $\Theta \in \mathcal{H}$, S_0, $v_0 > 0$, T, $K > 0$.

Remark 9.2.3. We remark that 2.24 is equivalent to the weak convergence of $\{f_n\}_n$ to $\delta(\cdot - v_0)$. The chain of inequalities 2.25 shows that the averaged price 2.21 yields a wider range of prices than the standard Heston model, when v_0 varies within a bounded interval, and this holds in practice. In fact, if the observed market prices belong to the no-arbitrage range, that is

$$(S_0 - Ke^{-rT})^+ < C_i^M < S_0, \qquad \forall\, i = 1, \dots, N, \tag{2.26}$$

then, for every$\Theta \in \mathcal{H}$, there exist two numbers $0 < \underline{v}(\Theta) \leq \overline{v}(\Theta)$ such that

$$C^H(0, S_0, \underline{v}(\Theta), T_i, K_i; \Theta) = \underline{C}_i^M := \min_{1 \leq j \leq N} C_j^M, \tag{2.27}$$

$$C^H(0, S_0, \overline{v}(\Theta), T_i, K_i; \Theta) = \overline{C}_i^M := \max_{1 \leq i \leq N} C_i^M. \tag{2.28}$$

Therefore, by the strict monotonicity of C^H as a function of v, it is easy to see that

$$I = \inf_{(\Theta,v_0)\in\times(0,\infty)} J(\Theta, v_0) = \inf_{(\Theta,v_0)\in\mathcal{K}} J(\Theta, v_0), \tag{2.29}$$

where

$$\mathcal{K} = \{(\Theta, v_0) \in \mathcal{H} \times (0,\infty) : \;\; v_0 \in [\underline{v}(\Theta), \overline{v}(\Theta)]\}. \tag{2.30}$$

Therefore, for every fixed Θ, the optimal initial volatility belongs to a bounded interval.

Before we give the proof of Theorem 9.2.2, we first formulate an example for the set $\mathcal{P}'$ which satisfies the assumption of Theorem 9.2.2. In light of the conjecture of [244] that consider the stationary solution 2.15 as a prior distribution for v_0, we put our attention on the subset $\mathcal{G} \subset \mathcal{P}$ of the probability density functions associated with the Gamma distribution:

$$\mathcal{G} = \left\{ g_{\alpha,\beta} \in \mathcal{P} \, : \; g_{\alpha,\beta}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} 1_{x>0}, \quad \alpha, \beta \quad > 0 \right\}. \tag{2.31}$$

We have the following result.

Remark 9.2.4. For every $\alpha, \beta >\; 0$ and $z \in \mathbb{C}$, with $\mathrm{Re}(z) < \beta$, the moment generating function of $g_{\alpha,\beta} \in \mathcal{G}$ is given by

$$M_{\alpha,\beta}(z) = \int_0^\infty e^{zv} \frac{\beta^\alpha}{\Gamma(\alpha)} v^{\alpha-1} e^{-\beta v}\, dv = \left(1 - \frac{1}{\beta} z\right)^{-\alpha}. \tag{2.32}$$

The set $\mathcal{G}$ satisfies condition 2.24 in Theorem 9.2.2.
Let $v_0 > 0$ and $g_n \in \mathcal{G}$ such that $g_n = g_{\alpha_n,\beta_n}$, with $\alpha_n = n$ $\beta_n = n/v_0$. The characteristic function of g_n is

$$\phi_n(t) = \left(1 - \frac{\imath t}{\beta_n}\right)^{-\alpha_n} = \left(1 - \frac{\imath t \bar{v}}{n}\right)^{-n}, \tag{2.33}$$

for all $t \in \mathbb{R}$. g_n converges weakly to $\delta(\cdot - v_0)$, since $\phi_n(t) \to e^{\imath t v_0}$, for any $t \in \mathbb{R}$, where

$$\phi(t) = \int_{\mathbb{R}} e^{\imath t x} \delta(x - v_0) dx = e^{\imath t v_0}, \tag{2.34}$$

which is the characteristic function associated to the delta function, centered at v_0.

Remark 9.2.5. Although our analysis is focused on the Gamma distribution, we observe that different types of probability distributions satisfy condition 2.24. In particular we mention the Inverse Gaussian distribution (IG) with density function

$$IG_{\alpha,\beta}(x) = \left[\frac{\alpha}{2\pi x^3}\right]^{1/2} \cdot \exp\left(-\frac{\alpha(x-\beta)^2}{2\beta^2 x}\right), \tag{2.35}$$

for $x > 0$, where $\beta > 0$ is the mean and $\alpha > 0$ is the shape parameter. Then it is easy to see that IG_{α_n,v_0} converges weakly to $\delta(\cdot - v_0)$, for any sequence $\alpha_n \to \infty$. In particular, the Inverse Gaussian and Gamma distributions are special cases of the generalized Inverse Gaussian distribution (GIG) having density function

$$GIG_{a,b,p}(x) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})} x^{p-1} \cdot \exp\left[-\frac{1}{2}\left(ax + \frac{b}{x}\right)\right], \tag{2.36}$$

for $x > 0$, with parameters $a,\ b > 0$, $p \in \mathbb{R}$. Here K_p denotes the modified Bessel function of the second kind. This is the Gamma distribution if $a = 2\beta$ and $b \to 0$, $p = \alpha$; it is the Inverse Gaussian if $a = \alpha/\beta^2$, $b = \alpha$ and $p = -1/2$. Another suitable distribution is a scaled version of noncentral-χ^2 distribution. In fact, as documented in several papers, the distribution of the Heston volatility v_T, conditional on v_t, for $t < T$, is distributed according to such a distribution with parameters derived from the Heston model, see for example [257].

Proof. - Theorem 9.2.2. The function $u(t, S, v) = S$ is a super-solution of 1.8, since $\mathcal{L}u = 0$ and $u(T, S, v) = S > (S - K)^+$. Thus by the comparison principle for 1.8, proved in [245], it holds $C^H(0, S_0, v_0, T, K; \Theta) \leq S_0$, for every $S_0 > 0$, $v_0 > 0$, $K,\ T > 0$, $\Theta \in \mathcal{H}$. This yields

$$0 \leq C_f^H(S_0, T, K; \Theta) = \int_0^\infty C^H(0, S_0, v, T, K; \Theta) f(v)\, dv \leq S_0 < \infty, \tag{2.37}$$

for any $f \in \mathcal{P}$.
The Jensen's inequality implies

$$\begin{aligned} |C_i^M - C_f^H(S_0, T, K; \Theta)|^2 &= |\mathbb{E}_f[C_i^M - C^H(0, S_0, \cdot, T_i, K_i; \Theta)]|^2 \\ &\leq \mathbb{E}_f\left[|C_i^M - C^H(0, S_0, \cdot, T_i, K_i; \Theta)|^2\right], \end{aligned} \tag{2.38}$$

for every $f \in \mathcal{P}$. Summing over $i = 1, \ldots, N$, we get

$$J'(\Theta, f) \leq \mathbb{E}_f[J(\Theta, \cdot)] \qquad \forall\, f \in \mathcal{P}'. \tag{2.39}$$

Let $v_0 > 0$ and $\{f_n\}_n \subset \mathcal{P}'$, satisfying 2.24. Writing inequality 2.39 for f_n, leads to

$$I' \leq J'(\Theta, f_n) \leq \mathbb{E}_{f_n}[J(\Theta, \cdot)] = \int_{\mathbb{R}} J(\Theta, v) f_n(v) dv. \tag{2.40}$$

By 2.37, the function $v \mapsto J(\Theta, v)$ is bounded, for any$\Theta\ \in \mathcal{H}$. Hence, we can take the limit on the right-hand side as $n \to \infty$ to obtain

$$I' \leq J(\Theta, v_0), \tag{2.41}$$

for any arbitrary Θ and $v_0 > 0$, and taking the infimum over $(\Theta, v_0) \in \mathcal{H} \times (0, \infty)$. This proves the inequality $I' \leq I$.
We prove the inequality for the supremum in 2.25, the other relations can be obtained with similar arguments. Let $\bar{v} > v_0 > 0$, and $\{f_n\}_n \subset \mathcal{P}'$ be a sequence of density functions associated to $\bar{v}$. Let $g \in C^\infty(\mathbb{R})$ be such that $0 \leq g(v) \leq 1$, everywhere, $g(v) = 0$, for $|v| \geq 1$, $g(v) = 1$ for $|v| \leq 1/2$. Define $g_\varepsilon(v) = g\left(\frac{v-\bar{v}}{\varepsilon}\right)$, for every $\varepsilon > 0$. Since $v \mapsto C^H(0, S_0, v, T, K; \Theta)$ is strictly increasing, for every $0 < \varepsilon < \bar{v} - v_0$, we can write

$$\begin{aligned} C^H_{f_n}(S_0, T, K; \Theta) &\geq \int_{\bar{v}-\varepsilon}^{\bar{v}+\varepsilon} C^H(0, S_0, v, T, K; \Theta) f_n(v)\, dv \\ &\geq C^H(0, S_0, \bar{v}-\varepsilon, T, K; \Theta) \int_0^\infty g_\varepsilon(v) f_n(v)\, dv. \end{aligned} \tag{2.42}$$

Taking the limit as $n \to \infty$, we get

$$\sup_{f \in \mathcal{P}'} C^H_f(S_0, T, K; \Theta) \geq C^H(0, S_0, \bar{v}-\varepsilon, T, K; \Theta) > C^H(0, S_0, v_0, T, K; \Theta), \tag{2.43}$$

for any $S_0 > 0$, $v_0 > 0$, T, $K > 0$,$\Theta \in \mathcal{H}$. □

Following Theorem 9.2.2 the calibration procedure using the average call price 2.21, under the Gamma distribution set $\mathcal{G}$, consists in minimizing the functional

$$I_{\mathcal{G}}(\alpha, \beta, \Theta) := \sum_{i=1}^N \omega_i |C_i^M - \mathbb{E}_{g_{\alpha,\beta}}[C^H(0, S_0, \cdot, T_i, K_i; \Theta)]|^2 \tag{2.44}$$

over $(\alpha, \beta, \Theta) \in (0, \infty)^2 \times \mathcal{H}$. Thus, the calibration is achieved by adding to the set of parameters $\mathcal{H}$ two real parameters that describe the distribution of the initial volatility v_0. We also observe that the averaged call price 2.21 is strictly increasing with respect to the scale parameter $\gamma = 1/\beta$. In fact, by the regularity properties of the Heston call price it holds:

$$\begin{aligned} &\frac{\partial}{\partial \gamma} \mathbb{E}_{g_{\alpha,\beta}}[C^H(0, S_0, \cdot, T, K; \Theta)] \\ &= \frac{\partial}{\partial \gamma} \int_0^\infty C^H(S_0, \gamma w, T, K; \Theta) \frac{1}{\Gamma(\alpha)} w^{\alpha-1} e^{-w}\, dw \\ &= \int_0^\infty \frac{\partial C^H}{\partial v}(S_0, \gamma w, T, K; \Theta) \frac{1}{\Gamma(\alpha)} w^{\alpha-1} e^{-w}\, dw > 0. \end{aligned} \tag{2.45}$$

Hence we can conjecture that the scale parameter represents an estimate of the "true" volatility in the Heston model.

9.3 Average call price formula

In this section we derive a closed-form formula for the averaged call price 2.21, given a probability density function$\Pi \in \mathcal{P}$.

$$C_{\Pi}(S_0,T,K;\Theta) = \int_0^{+\infty} C^H(S_0,v,T,K;\Theta)\Pi(v)\,dv. \tag{3.46}$$

We will state a result that yields a simplified form, reducing the expression of the price above to a single integration. This will be of great convenience for numerical computation purposes.

Theorem 9.3.1. (**Average Call Price**) *If $\Pi \in \mathcal{P}$ satisfies $\mathbb{E}_{\Pi}[v] < \infty$, then the following relation holds true:*

$$C_{\Pi}(S_0,T,K;\Theta) = S_0 Q_1(S_0,T,K;\Theta) - e^{-rT} K Q_2(S_0,T,K;\Theta), \tag{3.47}$$

where

$$Q_j = \frac{1}{2} + \frac{1}{\pi}\int_0^{\infty} \mathrm{Re}\Big[\frac{e^{C_j(T,\phi)+\imath\phi\log(\frac{S_0}{K})} M_{\Pi}\big(D_j(T,\phi)\big)}{\imath\phi}\Big]\,d\phi, \tag{3.48}$$

for $j = 1,\ 2,\ S_0,\ T,\ K > 0,\ r > 0,\ \Theta \in \mathcal{H}$, with $\rho \in (-1,1)$, M_{Π} being the moment generating function 2.18 related to Π.

Remark 9.3.2. If Π is the pdf associated to the Gamma distribution with parameters (α,β), the integrand in 3.48 reduces to

$$Q_j = \frac{1}{2} + \frac{1}{\pi}\int_0^{\infty} \mathrm{Re}\left[\frac{e^{C_j(T,\phi)+\imath\phi\log(\frac{S_0}{K})}}{\imath\phi\big(1-\beta D_j(T,\phi)\big)^{\alpha}}\right]\,d\phi. \tag{3.49}$$

In the cases of the Inverse Gaussian distribution (IG) and the Generalized Inverse Guassian distribution (GIG), we can also find an explicit expression for Q_j which are based on the moment generating function of these distributions, respectively given by:

$$M_{IG}(z) = \exp\left[\frac{\alpha}{\beta}\left(1-\sqrt{1-\frac{2\beta^2 z}{\alpha}}\right)\right], \tag{3.50}$$

$$M_{GIG}(z) = \left(\frac{a}{a-2z}\right)^{p/2} \frac{K_p(\sqrt{b(a-2\imath z)})}{K_p(\sqrt{ab})}, \tag{3.51}$$

where K_p is a modified Bessel function of the second kind.

In order to prove the previous result, we need some technical results, which state some crucial properties for the coefficients in the Heston formula 1.6-1.7. For our knowledge, the inequality in Lemma 9.3.3 is not proved in the literature, hence we will give a detailed proof of such a result. Proposition 9.3.4 provides useful information about the asymptotic behavior of the coefficients $C_j(\phi, T)$ an $D_j(\phi, T)$, as $\phi \to \infty$ and as $\phi \to 0^+$. The proof can be found in [258] - Propositions 3.1, 3.2, 3.3. See also [259].

Lemma 9.3.3. *For any* $\phi, \kappa, \theta, \sigma > 0$, $\rho \in (-1, 1)$, $\tau > 0$ *we have that*

$$\operatorname{Re}(D_j(\tau, \phi)) < 0, \tag{3.52}$$

for $j = 1, 2$.

Let us introduce the notation for the integrands in the Heston model 1.4-1.5:

$$p_j(\phi, \log(S_0), v) = \operatorname{Re}\left[\frac{e^{-\imath\phi\log(K)} f_j(\phi; \log(S_0), v)}{\imath\phi}\right] \qquad j = 1, 2, \tag{3.53}$$

then we can state the following asymptotics.

Proposition 9.3.4. *Assuming that* κ, θ, σ , S_0, K, $T > 0$ *and* $\rho \in (-1, 1)$, *then the following asymptotics hold:*

$$\lim_{\phi \to \infty} \frac{C_j(T, \phi)}{\phi} = -\frac{\kappa\theta}{\sigma} T \left(\sqrt{1 - \rho^2} + \imath\rho\right) + \imath r T, \tag{3.54}$$

$$\lim_{\phi \to \infty} \frac{D_j(T, \phi)}{\phi} = -\frac{\sqrt{1 - \rho^2} + \imath\rho}{\sigma}, \tag{3.55}$$

$$\lim_{\phi \to 0^+} p_j(\phi, \log(S_0), v) = \log(S_0/K) + \operatorname{Im}(\partial_\phi C_j(T, 0)) + \operatorname{Im}(\partial_\phi D_j(T, 0))v, \tag{3.56}$$

where

$$\operatorname{Im}(\partial_\phi C_j(T, 0)) = rT + a\frac{e^{-b_j T} + b_j T - 1}{2b_j^2} \tag{3.57}$$

$$\operatorname{Im}(\partial_\phi D_j(T, 0)) = (-1)^j \cdot \frac{e^{-b_j T} - 1}{2b_j}, \tag{3.58}$$

for $j = 1, 2$, *provided that* $b_1 = \kappa - \rho\sigma \neq 0$, *otherwise, we obtain* $\operatorname{Im}(\partial_\phi C_j(0, T)) = aT^2/4$ *and* $\operatorname{Im}(\partial_\phi D_j(0, T)) = \frac{T}{2}$ *for* $j = 1, 2$.

Proof. - Theorem 9.3.1. According to Heston model 1.2,

$$P_j = \frac{1}{2} + \frac{1}{\pi}\int_0^{\infty} p_j(\phi, \log(S_0), v)\, d\phi, \qquad j = 1,\ 2. \tag{3.59}$$

Therefore it suffices to prove the equation $Q_j = \frac{1}{2} + \frac{1}{\pi}I$, where

$$I = \int_0^{\infty}\int_0^{\infty} p_j(\phi, \log(S_0), v)\Pi(v)\, d\phi\, dv. \tag{3.60}$$

Let's first verify the convergence of I. To this end, Proposition 9.3.4-3.54 yields

$$\mathrm{Re}(C_j(T,\phi)) < -c\phi, \tag{3.61}$$

as $\phi > \eta$, with $\eta > 0$ chosen large enough, where $c = \kappa\theta T\sqrt{1-\rho^2}/(2\sigma) > 0$. Moreover, it is easy to see that $\mathrm{Im}(D_j(T,\phi)) \to 0$, as $\phi \to 0^+$. Therefore, by Proposition 9.3.4-3.56, setting $v = 0$, there exists $\varepsilon > 0$ such that

$$|p_j(\phi, \log(S_0), 0)| \;\leq\; 2|\log(S_0/K)| + 2|\mathrm{Im}(\partial_\phi C_j(0,T))|, \tag{3.62}$$

$$|\mathrm{Im}(D_j(T,\phi))| \;\leq\; d_j\phi, \tag{3.63}$$

for every $0 < \phi < \varepsilon$, for a constant coefficient d_j depending only on the coefficient in Proposition 9.3.4-3.57.
Now it suffices to show that the integrals I_∞ and I_0, defined below, are bounded. They represent respectively the integral on $(\phi, v) \in (\eta, \infty) \times (0, \infty)$ and the integral over $(\phi, v) \in (0, \varepsilon) \times (0, \infty)$. Let us consider them separately:

$$\begin{aligned} I_\infty &= \\ &\int_0^{\infty}\left[\int_\eta^{\infty} \Pi(v)\frac{1}{\phi}e^{\mathrm{Re}(C_j)+\mathrm{Re}(D_j)v}\left|\sin\left(\mathrm{Im}(C_j) + \mathrm{Im}(D_j)v + \phi\log\frac{S_0}{K}\right)\right| d\phi\right] dv \\ &\leq \int_0^{\infty}\left[\Pi(v)\int_0^{\infty}\frac{e^{\mathrm{Re}(C_j)}}{\phi}e^{\mathrm{Re}(D_j)v}\, d\phi\right] dv \leq \int_\eta^{\infty}\frac{e^{\mathrm{Re}(C_j(T,\phi))}}{\phi}\, d\phi < \infty. \end{aligned} \tag{3.64}$$

Here we have omitted the dependence of C_j and D_j on (T,ϕ), and we have used Lemma 9.3.3 in the last inequality. Still using Lemma 9.3.3 and inequality $|\sin(x)| \leq$

$|\sin(y)| + |x - y|$, for any pair of real numbers $x, \in y$, we get

$$
\begin{aligned}
& I_0 = \\
& \int_0^\infty \Pi(v) \left[\int_0^\varepsilon \frac{1}{\phi} e^{\mathrm{Re}(C_j)+\mathrm{Re}(D_j)v} \left| \sin\left(\mathrm{Im}(C_j) + \mathrm{Im}(D_j)v + \phi \log \frac{S_0}{K} \right) \right| d\phi \right] dv \\
& \leq \int_0^\infty \Pi(v) \left[\int_0^\varepsilon \frac{1}{\phi} e^{\mathrm{Re}(C_j)} \left[\left| \sin\left(\mathrm{Im}(C_j) + \phi \log \frac{S_0}{K} \right) \right| + |\mathrm{Im}(D_j)| v \right] d\phi \right] dv \\
& \leq \int_0^\varepsilon |p_j(\phi, \log(S_0), 0)| \, d\phi + d_j \varepsilon \, \mathbb{E}_\Pi \left[v \right]. \qquad (3.65)
\end{aligned}
$$

Hence, I_0 is bounded under the assumption $\mathbb{E}_\Pi[v] < \infty$ and by 3.62. So we are allowed to change the order of integration in 3.60:

$$
\begin{aligned}
I &= \int_0^\infty \left[\int_0^\infty \mathrm{Re}\Big[\frac{e^{-\imath\phi \log(K)} f_j(\phi; \log(S_0), v)}{\imath\phi} \Big] \Pi(v) dv \right] d\phi \\
&= \int_0^\infty \left[\mathrm{Re}\Big[\int_0^\infty \frac{e^{-\imath\phi \log(K)} f_j(\phi; \log(S_0), v)}{\imath\phi} \Pi(v) \Big] dv \right] d\phi \\
&= \int_0^\infty \mathrm{Re}\Big[\frac{e^{\imath\phi \log \frac{S_0}{K} + C_j(T,\phi)}}{\imath\phi} \int_0^\infty e^{D_j(T,\phi)v} \, \Pi(v) dv \Big] d\phi \\
&= \int_0^\infty \mathrm{Re} \left[\frac{e^{\imath\phi \log \frac{S_0}{K} + C_j(T,\phi)}}{\imath\phi} M_\Pi(D_j(T,\phi)) \right] d\phi. \qquad (3.66)
\end{aligned}
$$

We remark that $\mathrm{Re}(D_j(T,\phi)) < 0$ implies that $M_\Pi\big(D_j(T,\phi)\big)$ is well defined for all $T > 0$, $\phi > 0$. Note also that the switch between the integral and the real is allowed since

$$
\begin{aligned}
\left| \frac{e^{-\imath\phi \log(K)} f_j(\phi; \log(S_0), v)}{\imath} \Pi(v) \right| &= e^{\mathrm{Re}(C_j(T,\phi)) + \mathrm{Re}(D_j(T,\phi))v} \Pi(v) \\
&< e^{\mathrm{Re}(C_j(T,\phi))} \, \Pi(v), \qquad (3.67)
\end{aligned}
$$

and, given that $\mathrm{Re}(C_j(T,\phi))$ does not depend on v, this function is integrable with respect to $v \in (0, \infty)$, for every ϕ. □

Proof. - Lemma 9.3.3. As is well known (see [236]), $\tau \mapsto D_j(\tau,\phi)$ solves, for every $\phi > 0$, a Riccati type equation:

$$
\frac{\partial D_j}{\partial \tau}(\tau,\phi) = A_j(\phi) - B_j(\phi) D_j(\tau,\phi) + R D_j^2(\tau,\phi), \qquad (3.68)
$$

and $D_j(0,\phi)=0$, where

$$A_j(\phi) = \imath u_j\phi - \frac{1}{2}\phi^2 \tag{3.69}$$

$$B_j(\phi) = b_j - \rho\sigma\phi\imath \tag{3.70}$$

$$R = \frac{1}{2}\sigma^2. \tag{3.71}$$

Thus, the function $w(\tau,\phi)=\exp\left(-R\int_0^\tau D_j(t,\phi)dt\right)$ solves the second order differential equation

$$\partial_\tau^2 w(\tau,\phi) + B_j(\phi)\partial_\tau w(\tau,\phi) + RA_j(\phi)w(\tau,\phi) = 0. \tag{3.72}$$

In the sequel we shall use the notation w' to denote the partial derivative $\partial_\tau w(\tau,\phi)$. and $\bar{w}(\tau,\phi)$ for the conjugate of $w(\tau,\phi)$. Computing the real part of D_j leads to

$$\mathrm{Re}(D_j) = -\frac{1}{R}\mathrm{Re}\left(\frac{w'\bar{w}}{|w|^2}\right) = -\frac{1}{R}\frac{w'_R w_R + w'_I w_I}{|w|^2}, \tag{3.73}$$

where $w = w_R + \imath w_I$ and $w' = w'_R + \imath w'_I$. Let $\xi(\tau,\phi)=|w(\tau,\phi)|^2$, then of course we can write $\xi' = w'\bar{w} + w\bar{w}'$ and $\mathrm{Re}(D_j) = -\xi'/(2R\xi)$. Moreover, by using equation 3.72, we find

$$\begin{aligned}
\xi'' &= -B_j w'\bar{w} - A_j Rw\bar{w} + 2w'\bar{w}' - w(\bar{B}_j\bar{w}' + \bar{A}_j R\bar{w})\\
&= -\mathrm{Re}(B_j)(w'\bar{w}+w\bar{w}') - \imath\mathrm{Im}(B_j)(w'\bar{w}-w\bar{w}') - 2R\,\mathrm{Re}(A_j)\xi\\
&\quad + 2w'\bar{w}' = -\mathrm{Re}(B_j)\xi' + 2\mathrm{Re}(\imath w\bar{w}')\mathrm{Im}(B_j) - 2R\mathrm{Re}(A_j)\xi +\\
&\quad +2w'\bar{w}' = -b_j\xi' - 2\mathrm{Re}(\imath w\bar{w}')\sigma\rho\phi + \frac{1}{2}\sigma^2\phi^2\xi + 2w'\bar{w}',
\end{aligned} \tag{3.74}$$

and, by the definition of w, we argue that

$$\mathrm{Re}(\imath w\bar{w}') = -R\,\mathrm{Im}(D_j)|w|^2 = -R\,\mathrm{Im}(D_j)\,\xi, \tag{3.75}$$

$$|w'|^2 = R^2\xi|D_j|^2. \tag{3.76}$$

By these relations, we deduce that ξ solves the Cauchy problem:

$$\begin{cases}
\xi''(\tau) + b_j\xi'(\tau,\phi) - \gamma(\tau,\phi)\xi(\tau,\phi) = 0,\\
\xi'(0,\phi) = 0,\\
\xi(0,\phi) = 1,
\end{cases} \tag{3.77}$$

where $\gamma(\tau,\phi) = (1/2)\sigma^2\phi^2 + 2[R^2|D_j(\tau,\phi)|^2 + R\sigma\rho\phi\,\mathrm{Im}(D_j(\tau,\phi))]$. In order to prove that $\mathrm{Re}(D_j(\tau,\phi))<0$, it suffices to show that $\tau\in(0,\infty)\mapsto\xi(\tau,\phi)$ is strictly increasing. Assume that this is not true, by way of contradiction. Suppose that there exists $\tau'>0$ such that $\xi'(\tau')<0$. Since $D_j(0,\phi)=0$, 3.77 yields

$$\xi''(0,\phi) = \frac{1}{2}\sigma^2\phi^2 + 2\gamma(0) = \frac{1}{2}\sigma^2\phi^2 > 0, \tag{3.78}$$

for all $\phi > 0$; hence, by the continuity of ξ' and ξ'' as functions of τ, the supremum

$$\tau_0 = \sup\{\tau \in (0, \tau'] : \xi'(\tau,\phi) \geq 0\}, \tag{3.79}$$

is well defined, since the related set is non-empty, and $\tau_0 < \tau'$, $\xi'(\tau_0, \phi) = 0$. We show that $\xi''(\tau_0, \phi) > 0$. From the differential equation 3.77 and $\mathrm{Re}(D_j(\tau_0, \phi)) = 0$, this is equivalent to state the inequality

$$\phi^2 + 2\rho\phi A + A^2 > 0, \tag{3.80}$$

where $A = \sigma \mathrm{Im}(D_j(\tau_0, \phi))$. If $A = 0$ the last inequality reduces to $\phi^2 > 0$, that is obviously true. Otherwise, if $A \neq 0$, since $|\rho| < 1$, we get

$$\phi^2 + 2\rho\phi A + A^2 > \phi^2 - 2|A|\phi + A^2 = (\phi - |A|)^2 \geq 0. \tag{3.81}$$

Thus $\xi'(\tau,\phi)$ is positive in a right neighborhood of τ_0. This is in contradiction with the definition of τ_0 and we have proved that $\xi'(\tau,\phi) > 0$ for every τ, $\phi > 0$, implying inequality 3.52. □

The assertion of Lemma 9.3.3 is still valid if $D_2(\tau,\phi)$ is replaced by $D_2(\tau,\phi - \imath/2)$. In fact, the coefficient $\kappa - \rho\sigma/2$ substitutes $b_2 = \kappa$, without compromising the proof. The asymptotic behavior established in Proposition 9.3.4 holds true for $C_2(\tau, u - \imath/2)$, as $u \to \infty$ ([259]). This remark is relevant in the context of the Lewis approach 1.10, where we can obtain a reduced form expression for the averaged call price which involves only single integration.

Theorem 9.3.5. (**Lewis Average Call Price**) *If $\Pi \in \mathcal{P}$ satisfies $\mathbb{E}_\Pi[v] < \infty$, then the following relation holds true:*

$$C_\Pi(S_0, T, K; \Theta) = S_0 - \frac{\sqrt{S_0 K} e^{-rT}}{\pi} \int_0^{+\infty} \frac{F(u)\, du}{u^2 + \frac{1}{4}}, \tag{3.82}$$

where

$$F(u) = e^{C_2(T,u-\imath/2)+\imath u \log(S_0/K) - \frac{rT}{2}} M_\Pi \left(D_2(T, u - \imath/2)\right). \tag{3.83}$$

for every S_0, T, $K > 0$, $r > 0$, $\Theta \in \mathcal{H}$, with $\rho \in (-1, 1)$.

Proof. By Lewis call price 1.10, we can write

$$C_\Pi(S_0, T, K, \Theta) \;=\; S_0 - \frac{\sqrt{S_0 K e^{-rT}}}{\pi} I, \tag{3.84}$$

where

$$I = \int_0^{+\infty} \left[\int_0^{+\infty} \frac{1}{u^2 + \frac{1}{4}} \mathrm{Re}\Big[e^{\imath uk} \phi_T(u - \imath/2) \Big] \Pi(v) du \right] dv. \tag{3.85}$$

The convergence of the integral I can be obtained by the same argument used in the proof of Theorem 9.3.1. Thus, we can change the integration order:

$$\begin{aligned} I = & \int_0^{+\infty} \frac{1}{u^2 + \frac{1}{4}} \left[\int_0^{+\infty} \mathrm{Re}\Big[e^{C_2(T,u-\frac{\imath}{2}) + D_2(T,u-\frac{\imath}{2})v + \imath u(k-rT) - \frac{rT}{2}} \Big] \Pi(v) dv \right] du \\ = & \int_0^{+\infty} \frac{1}{u^2 + \frac{1}{4}} \left[\mathrm{Re}\Big[e^{C_2(T,u-\frac{\imath}{2}) + \imath u(k-rT) - \frac{rT}{2}} \int_0^{+\infty} e^{D_2(T,u-\frac{\imath}{2})v} \Pi(v) \Big] dv \right] du \\ = & \int_0^{+\infty} \frac{1}{u^2 + \frac{1}{4}} \mathrm{Re}\Big[e^{C_2(T,u-\frac{\imath}{2}) + \imath u \log(S_0/K) - \frac{rT}{2}} M_\Pi\big(D_2(T, u - \imath/2)\big) \Big] du. \end{aligned} \tag{3.86}$$

We remark that Lemma 9.3.3 allows the change of order between the integration with respect to v and the real part, and it implies that $M_\Pi\big(D_2(T, u - \frac{\imath}{2})\big) < \infty$ in the expression above. Clearly 3.86 proves the equation 3.82. □

9.3.1 Calibration to option prices

The general approach to the calibration of parametric models, such as the Heston model, is to apply a least-square type procedure either in price or implied volatility. Unfortunately, this kind of approach will in general be very sensitive to the choice of the initial point, which will often in practice drive the selection of the local minima the algorithm will converge to. The various explicit formulas come into play to receive a pertaining initial point. Estimates for the volatility parameter v_0, with the structural parameters $= \{\kappa,\ \theta, \rho\ , \sigma\ \}$ will be needed. The calibration procedure consists in the minimization of the functional in 2.12, where C^{Model} is the Heston call price $C^H(S_0, v_0, T, K; \Theta)$ in the standard case or, otherwise the weighted average call price $C_\Pi(S_0, T, K; \Theta)$ 9.3.1-9.3.5, for a given probability distribution density$\Pi\ \in \mathcal{P}$. In this second case, the density is chosen according to a parameterized family of density functions related to a probability distribution, implying that the set of parameters includes also the parameters of such a distribution. By the results of Section 9.2.1, we have compared the standard method which considers v_0 as an additional parameter and our approach under three distributions: the Gamma (GAM), the Inverse Gaussian (IG) and the Generalized Inverse Gaussian (GIG), for which the integrands appearing in 9.3.1 are explicitly known thanks to the relations 3.49, 3.50 and 3.51. From a numerical point of view, the calculation of the option price is made somewhat complicated by the fact that the integrands have oscillatory nature. However, the integration can be done in a reasonably simple fashion by the aid of Gauss-Lobatto quadrature. This integration method is capable of handling a wide range of functional forms. Since the Gauss-Lobatto algorithm is designed to operate on a closed bounded interval, we have used a transformation of the original integral boundaries $(0, \infty)$ to the finite interval $[0, 1]$, as presented in Kahl and Jackel (2005). In order to evaluate the performance of those methods, we have used three discrepancy measures documented in several works in the literature: the average prediction error (APE), the root mean-square error (RMSE) and the average relative prediction error (ARPE). They are defined as follows:

$$APE = \sum_{n=1}^{N} \frac{|C_i^{Model} - C_i^{Market}|}{\sum_{n=1}^{N} C_i^{Market}} \tag{3.87}$$

$$ARPE = \frac{1}{N} \sum_{i=1}^{N} \frac{|C_i^{Model} - C_i^{Market}|}{C_i^{Market}} \tag{3.88}$$

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{|C_i^{Model} - C_i^{Market}|^2}{N}}, \tag{3.89}$$

where $C_i^{Model} = C^H(S_0, v_0, T_i, K_i; \Theta)$ under the simple Heston model (H) and $C_i^{Model} = C_\Pi(S_0, T_i, K_i; \Theta)$ under the average price method denoted WAPH in the following. T_i and K_i denote respectively the maturity and the strike price of the ith option, all written on a stock with current price S_0. The admissible parameter set for the Heston model has been specified in 2.19. The coefficients w_i are chosen as described in Section 9.2. In fact, the use of the bid-ask spread (ask price minus the bid price) of the market price is suitable since it assigns a greater weight to options where the spread is small, and less weight to options with a larger spread.

9.4 Estimation results

Our empirical analysis is conducted on a dataset of option prices on the Standard and Poor's 500 Index, which represents the main capitalization-weighted index of 500 stocks in the US market. The index is designed to measure performance of the broad domestic economy through changes in the aggregate market value of 500 stocks representing all major industries. We have considered a first set of market composed by prices from September 1, 2010 to September 30, 2010 and a second set of market composed by prices from September 1, 2015 to September 30, 2015. It is considered only call options that verify standard no-arbitrage bounds. Moreover we test the model on call options, with the constraint on the moneyness $0.9 < M < 1.1$, where M is the moneyness defined by $\frac{K}{S_0}$. Overall, we have considered $8,315$ call prices divided into 21 trading dates and 9 expiry dates for the 2010's set and $1,091$ call prices divided into 21 trading dates and 7 expiry dates for the 2015's set. For each of the considered models we have calibrated everyday the corresponding parameters (that are 5 in the H case, 6 in both cases of WAPH with GAM and with IG, 7 in the case of WAPH with GIG). The results of the estimation are summarized in Table 9.2 and in Table 9.3. Precisely, the averages of daily error measures and the parameters for all methods are reported. For what concerns all the error measures, the averaged call price under the GIG distribution seems to perform better in the considered period, while for what concerns the case with the Gamma distribution (GAM) and the Inverse Gaussian distribution (IG) we observe a substantial equality of the performance of these approaches. In fact, as observed in Remark 9.2.5, the GIG case includes the GAM and the IG as special cases. The Heston model does not perform badly, but it is systematically beaten by the weighted average price model, especially for what concerns the RMSE criterion.

	APE	**RMSE**	**ARPE**
Heston	2.1905	0.5182	3.5480
WAPH-GAM	1.2878	0.2145	2.3781
WAPH-IG	1.1230	0.4167	2.5181
WAPH-GIG	1.3400	0.1104	1.0824

Table 9.1: Averages of the daily error measures APE, RMSE and ARPE for the different pricing methods on 2010 S&P option price database.

In order to better analysed the models performance we estimate the implied Black and Scholes volatility and for each trading days. Every trading day is associated with the standard deviation of the implied volatility (σ_σ) and the results of the parameters estimation are clustered considering different level standard deviation. In Table 9.5 are reported cluster average RMSE, results show that the GIG model error does not

	APE	**RMSE**	**ARPE**
Heston	4.5448	3.4457	0.0844
WAPH-GAM	3.9909	3.4239	0.1908
WAPH-IG	4.0201	3.1184	0.0923
WAPH-GIG	3.7336	2.9810	0.0700

Table 9.2: Averages of the daily error measures APE, RMSE and ARPE for the different pricing methods on 2015 S&P option price database.

	k	θ	σ	ρ	v_0		
Heston	0.0252	4.4944	0.4599	-0.6062	0.0024		
	k	θ	σ	ρ	α (shape)	β (rate)	
GAM	0.1178	0.9404	0.4567	-0.6057	0.0016	2.7953	
	k	θ	σ	ρ	α (shape)	β (mean)	
IG	0.1053	0.8843	0.5103	-0.6342	0.0023	0.0082	
	k	θ	σ	ρ	a	b (mean)	p
GIG	0.0145	1.2084	0.4432	-0.6401	5.4502	0.0001	0.0027

Table 9.3: Averages of the daily estimated parameters under the considered pricing methods for the 2010 S&P option price database.

depend on the standard deviation of the implied volatility and the conclusion that can be drawn is that clusters with more variability are better described by GIG model.
A natural extension of the Heston model is to include jumps in the stock price process and in the volatility process. Intuitively, it makes sense that jump in the stock price process should trigger a correlated jump in the volatility process in that sudden, large movements in the stock price would cause increased market anxiety around that stock. In our formulation of the jump stochastic volatility model (SVJJ) we have the following risk-neutral dynamics:

$$\begin{aligned} dS_t &= (r - \lambda\mu_J)S_t dt + S_t\sqrt{v_t}dW_t^1 + JS_t dN_t, \\ dv_t &= \kappa(\theta - v_t)dt + \sigma\sqrt{v_t}\left(\rho dW_t^1 + \sqrt{1-\rho^2}dW_t^2\right) + ZdN_t. \end{aligned} \tag{4.90}$$

N_t represents a Poisson process under the risk neutral measure, with jump intensity λ. The jump terms in the model are defined as follows:

$$\begin{aligned} Z &\sim \text{Exponential}(\mu_V) \\ (1+J)Z &\sim \text{Log-normal}(\mu_S + \rho_J Z, \sigma_S^2) \end{aligned} \tag{4.91}$$

where

$$\mu_J = \frac{e^{\mu_S + \frac{\sigma_S^2}{2}}}{1 - \rho_J\mu_V} - 1 \tag{4.92}$$

	k	θ	σ	ρ	v_0		
Heston	6.5140	0.0919	0.4732	-0.9999	0.0471		
	k	θ	σ	ρ	α (shape)	β (rate)	
GAM	4.3786	0.0282	0.4770	-0.9999	0.0100	4.3888	
	k	θ	σ	ρ	α (shape)	β (mean)	
IG	0.1505	0.2313	0.3483	0.8742	0.0067	0.0124	
	k	θ	σ	ρ	a	b (mean)	p
GIG	4.4033	0.0233	0.2754	-0.8743	2.7055	0.0675	101.0082

Table 9.4: Averages of the daily estimated parameters under the considered pricing methods for the 2015 S&P option price database.

	Heston	**WAPH-GIG**
$1.5\% < \sigma_\sigma \leq 2.0\%$	3.9526	2.5608
$2.0\% < \sigma_\sigma \leq 2.5\%$	4.6239	2.1010
$2.5\% < \sigma_\sigma \leq 3.0\%$	4.1979	1.8574
$3.0\% < \sigma_\sigma \leq 3.5\%$	5.5774	3.1035
$3.5\% < \sigma_\sigma \leq 4.0\%$	6.4538	2.7909
$4.0\% < \sigma_\sigma \leq 4.5\%$	7.8989	3.0723

Table 9.5: Average RMSE for different standard deviation cluster.

In Table 9.6 are reported the estimated parameters of the WAPH-GIG model and the SVJJ model. Table 9.7 shows that the WAPH-GIG and SVJJ provide same results.

This empirical analysis shows that our approach is quite promising and represents

	k	θ	σ	ρ	a	b (mean)	p
WAPH-GIG	2.9894	0.0242	0.3796	-0.8499	1.5416	0.0348	15.8841
	k	θ	σ	ρ	v_0		
SVJJ	4.9689	0.0352	0.5913	-0.9999	0.0580		
	λ	μ_S	σ_S	ρ_J	μ_V		
	0.0254	-0.8516	1.4978	-0.0440	39.9852		

Table 9.6: Pricing models estimated parameters on September 1, 2015 database.

	APE	**RMSE**	**ARPE**
WAPH-GIG	2.5048	2.9857	0.0491
SVJJ	2.3049	3.0306	0.0452

Table 9.7: Error measures APE, RMSE and ARPE for the different pricing models on September 1, 2015.

an improvement over the Heston model, while retaining the same degree of analytical tractability.

9.5 Summary

The model of [236] is a mathematical tool still widely used as a basis for the valuation of financial derivatives. In this work, we have described a new method for the calibration of the Heston model in order to improve the effectiveness of such model. Our method overcomes the problem of the non-observability of the initial volatility and it is inspired by a previous work of [244] for the estimation of the historical probability density function. We have formulated a generalization of this insight in a rigorous way in order to reduce the estimation error of the parameters in the calibration of option prices. Thus, we have established some theoretical results that allow to derive a new pricing relation. For future research directions, we aim to investigate the theoretical foundation of our approach with the purpose to prove that the weighted average option price can be interpreted as a no-arbitrage price.

10

Conclusion

Part I

The increase of the life expectancy and the increment of technology usage in the healthcare sector required an increase in resource consumption. Financing is a critical element determining the quantity, distribution, and quality of health service. Funding for recurrent operating and log-term development costs for health services may come from different primary sources:

1. Public source of financing
2. Private sources of financing
3. External financing.

Therefore it is possible to categorize the various types of healthcare systems along the main characteristics of their financing dimension because one type of financing will generally dominate a national system. Given the general finite resources allocated in the health system and the endless demand, there is a need to efficiently and effectively plan and manage all resources with particular emphasis on hospitals. Hospitals are the key element of many health systems, they provide a wide kind of service to a large number of patients. Hospitals are characterized by complex and variable dynamics, therefore in order to optimize the resource allocation process a framework is applied for healthcare planing and control that breaks down all functions of the hospital system. The main reliable and valid proxies for measuring the consumption of hospital resources are analysed:

1. A new PH distribution is introduced to describe hospital **length of stay** distribution

2. The single factor model is applied in order to describe the **discharge rate** and the **admission rate**.

Furthermore in order to better evaluate the service provided, the **quality approach** is implemented. The intents are mainly focused towards the identification of need through measurements and indices suitable to assess the performance. The reduction of interventions, the increase of mechanical support, and the decrement of the average LoS are some of evaluated indices. Also the **control charts** are used for on-line process surveillance. This tool puts in evidence a time trend that could be explained by the evolution of the department.
The analyses carried out draw attention to the optimization of resources in order to improve the service provided. The aim of the hospital bed management is to allocate beds to patients while taking into account capacity constraints. A **stochastic dynamic programming** approach allows the evaluation of the optimal bed allocation policy. In this way the hospital manager could balance the cost of empty beds against the cost of turning patients away, thus facilitating a good choice of bed provision in order to have a low cost and a high access to service. The models allow to analyze some aspects of biomedical resources consumption. Developing the approach based on dynamic programming is possible to establish a methodology that supports the decision-making process. Hence feasible extensions are:

- Consider a more generic model for the hospital discharge probability

- Use a cost function that take into account when a patient is rejected

- Modelling the emergency patients flow

In order to describe the variable cost associated to patient hospitalization a new approach is provided. Starting from the **Real Option Approach** the cost are described and a phenomenological interpretation is given. In according to Nelson-Siegel model a cost function is estimated and the result are compared with value obtained from option approach.

Part II

In addition to research on healthcare issues, other topics were investigated. Particularly according to the Memorandum of Understanding between the Campus Bio-Medico University and the Financial Guard, the Italian tax evasions phenomena and the Financial Guard law enforcement are analysed. A **stochastic version of Lotka-Volterra model** is applied in order to describe the dynamic relation existing between the prevention activities and the illegal behaviour. The proposed approach produces a good fit of empirical distribution and provides also a phenomenological interpretation of the analysed problem.

As part of a research project in cooperation with Enel a new **stochastic volatility model** is proposed. In order to avoid the estimation of the initial volatility, a weighted average formulation for the Heston stochastic volatility option price is presented. This approach has been developed in the literature for the estimation of the distribution of stock price changes (returns), showing an excellent agreement with real market data. This method is extended to the calibration of option prices considering a large class of probability distributions assumed for the initial volatility parameter. The estimation error is shown to be less than the case of the simple pricing formula. Results are also validated with a numerical comparison on observed call prices, between the proposed calibration method and the classical approach.

11

Appendix A

11.1 Probability Space, Random Variables

A **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where

- Ω is a *sample space*;
- $\mathcal{F}$ is a σ-algebra of Ω;
- $\mathbb{P}$ is the *physical* probability measure.

The simple space is the set of all possible outcome of an experiment. The σ-algebra $\mathcal{F}$ on Ω is a family $\mathcal{F}$ of subsets of Ω with the following properties:

1. $\emptyset \in \mathcal{F}$
2. $E \in \mathcal{F} \Rightarrow E^c \in \mathcal{F}$
3. $A_1, A2, ... \in \mathcal{F} \Rightarrow A := \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

The pair $(\Omega; \mathcal{F})$ is called a measurable space. A probability measure $\mathbb{P}$ on a measurable space $(\Omega; \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \to [0; 1]$ such that

1. $0 \leq \mathbb{P}(A) \leq 1$
2. $\mathbb{P}(\Omega) = 1$

3. For any sequence of events $A_1, A_2 ...$ that are mutually exclusive, that is, events for which $A_i \bigcap A_j = \emptyset$ when $i \neq j$ (where $\emptyset$ is the null set),

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \tag{1.1}$$

If $\mathcal{F}$ contains all subsets of Ω the probability space is *complete*. The subsets F ofΩ which belong to $\mathcal{F}$ are called $\mathcal{F} - measurablesets$. In a probability context these sets are called *events* and we use the interpretation

$$\mathbb{P}(F) = \textit{the probability that the event } F \textit{ occurs} \tag{1.2}$$

In particular, if $\mathbb{P}(F) = 1$ we say that F *occurs with probability 1*, or *almost surely*. Given any family $\mathcal{U}$ of subsets of Ω there is a smallest σ-algebra $\mathcal{H}_{\mathcal{U}}$ containing $\mathcal{U}$, namely

$$\mathcal{H}_{\mathcal{U}} = \bigcap\{\mathcal{H}; \mathcal{H}\ \sigma\text{-algebra of}\Omega\ ,\ \mathcal{U} \subset \mathcal{H}\} \tag{1.3}$$

We call $\mathcal{H}_{\mathcal{U}}$ the *σ-algebra generated* by $\mathcal{U}$.

If U is the collection of all open subsets of a topological space Ω (e.g. $\Omega = \ \mathbb{R}^n$), then $\mathcal{B} = \mathcal{H}_{\mathcal{U}}$ is called the *Borel σ-algebra* on Ω and the elements $B \in \mathcal{B}$ are called *Borel sets*. $\mathcal{B}$ contains all open sets, all closed sets, all countable unions of closed sets, all countable intersections of such countable unions etc.

If $(\Omega, \mathcal{F}, \mathbb{P})$ is a given probability space, then a function $Y : \Omega \rightarrow \mathbb{R}^n$ is called $\mathcal{F}$-*measurable* if

$$Y^{-1}(U) := \{\omega \in \Omega;\ Y(\omega) \in U\} \in \mathcal{F} \tag{1.4}$$

for all open sets $U \in \mathbb{R}^n$, (or, equivalently, for all Borel sets $U \subset \mathbb{R}^n$). If $X : \Omega \rightarrow \mathbb{R}^n$ is any function, then the σ-algebra $\mathcal{H}_{\mathcal{X}}$ generated by X is the smallest σ-algebra onΩ containing all the sets

$$X^{-1}(U); \quad U \subset \mathbb{R}^n\text{open} \tag{1.5}$$

It is not hard to show that

$$\mathcal{H}_{\mathcal{X}} = \{X^{-1}(B);\ B \in \mathcal{B}\} \tag{1.6}$$

where B is the Borel σ-algebra on $\mathcal{R}^n$. Clearly, X will then be $\mathcal{H}_{\mathcal{X}}$-measurable and $\mathcal{H}_{\mathcal{X}}$ is the smallest σ-algebra with this property. The following result is useful. It is a special case of a result sometimes called the *Doob-Dynkin* lemma.

Lemma 11.1.1. *If X, $Y : \Omega \to \mathbb{R}^n$ are two given functions, then Y is $\mathcal{H}_X$-measurable if and only if there exists a Borel measurable function $g : \mathbb{R}^n \to \mathbb{R}^n$ such that*

$$Y = g(X) \tag{1.7}$$

In the following we let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a given complete probability space. A **random variable** X is an $\mathcal{F}$-measurable function $X : \Omega \to \mathbb{R}^n$. Every random variable induces a probability measure μ_X on $(R)^n$, defined by

$$\mu_X(B) = \mathbb{P}(X^{-1}(B)) \tag{1.8}$$

μ_X is called the *distribution of* X. If $\int_\Omega |X(\omega)| d\mathbb{P}(\omega) < \infty$ then the number

$$\mathbb{E}[X] := \int_\Omega X(\omega) d\mathbb{P}(X) = \int_{\mathbb{R}^n} x d\mu_X(x) \tag{1.9}$$

is called *the expectation* of X. More generally, if $f : \mathbb{R}^n \to \mathbb{R}$ is Borel measurable and $\int_\Omega |X(\omega)| d\mathbb{P}(\omega) < \infty$ then we have

$$\mathbb{E}[f(X)] := \int_\Omega f(X(\omega)) d\mathbb{P}(X) = \int_{\mathbb{R}^n} f(x) d\mu_X(x) \tag{1.10}$$

Definition 11.1.2. *A sequence $\{X_1, X_2, ...\}$ of random variables is said to* ***converge in distribution****, or converge weakly, or converge in law to a random variable X if*

$$\lim_{n\to\infty} F_n(x) = F(x) \tag{1.11}$$

for every number $x \in \mathbb{R}$ at which F is continuous. Here F_n and F are the cumulative distribution functions of random variables X_n and X correspondingly. Convergence in distribution may be denoted as $X_n \to^d X$

Definition 11.1.3. *Let $\{X_n\}$ be a sequence of random variables, and let X be a random variables. Then $\{X_n\}$ is said to* ***converge in probability*** *to X if for every $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0. \tag{1.12}$$

We write $X_n \longrightarrow^{Pr}$ to indicate convergence in probability. Thus, the Weak Law says that $\frac{1}{n}\sum_{i=1}^n X_i$ converges in probability to μ, provided $\{X_i\}$ is a sequence of i.i.d. random variables with expectation μ.

Let $X_1, ..., X_n$ be a random sample of size n that is, a sequence of independent and identically distributed random variables drawn from distributions of expected values given by μ and finite variances given by σ^2. Suppose we are interested in the sample average

$$S_n := \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1.13}$$

of these random variables. By the law of large numbers, the sample averages converge in probability and almost surely to the expected value μ as n tends to infinity. The classical central limit theorem describes the size and the distributional form of the stochastic fluctuations around the deterministic number μ during this convergence. More precisely, it states that as n gets larger, the distribution of the difference between the sample average S_n and its limit μ, when blown up by the factor $\sqrt{n}$ (that is $\sqrt{n}(S_n - \mu)$), approximates the normal distribution with mean 0 and variance σ^2. For large enough n, the distribution of S_n is close to the normal distribution with mean μ and variance $\frac{1}{n^2}$. The usefulness of the theorem is that the distribution of $\sqrt{n}(S_n - \mu)$ approaches normality regardless of the shape of the distribution of the individual X_i's. Formally, the theorem can be stated as follows

Theorem 11.1.4. *Suppose $\{X_i\}_i \in \mathbb{B}$ is a sequence of i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}[X_i] = \sigma^2 < \infty$. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$:*

$$\sqrt{n}\Big(\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) - \mu\Big) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \tag{1.14}$$

The mathematical model for independence is the following:

Definition 11.1.5. *Two subsets A, $B \in \mathcal{F}$ are called independent if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B) \tag{1.15}$$

A collection $\mathcal{A} = \{\mathcal{H}_i; i \in I\}$ of families $\mathcal{H}_i$ of measurable sets is independent if

$$\mathbb{P}(H_{i_1} \cap \cdots \cap H_{i_k}) = \mathbb{P}(H_{i_1}) \cdots \mathbb{P}(H_{i_k}) \tag{1.16}$$

for all choices of $H_{i_1} \in \mathcal{H}_{i_1} \cdots H_{i_k} \in \mathcal{H}_{i_k}$, with different indices $i_1, ..., i_k$. A collection of random variables $\{X_i;\ i \in I\}$ is independent if the collection of generated σ-algebras $\mathcal{H}_{X_i}$ is independent.

If two random variables $X, Y : \Omega \to \mathbb{R}$ are independent then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[X] \tag{1.17}$$

provided that $\mathbb{E}[|X|] < \infty$ and $\mathbb{E}[|Y|] < \infty$.

Definition 11.1.6. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *a probability space, then* $L^2(\Omega, \mathcal{F}, \mathbb{P})$ *is the space of random variable* $X : (\Omega, \mathcal{F}) \to \mathbb{R}$ *such that*

$$\mathbb{E}[X^2] < +\infty \tag{1.18}$$

Definition 11.1.7. *A stochastic process is a parametrized collection of random variables*

$$\{X_t\}_{t \in T} \tag{1.19}$$

defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ *and assuming values in* $\mathbb{R}^n$.

The parameter space T is usually the haline $[0, \infty)$, but it may also be an interval $[a, b]$, the non-negative integers and even subsets of $\mathbb{R}^n$ for $n \geq 1$. Note that for each $t \in T$ xed we have a random variable

$$\omega \to X_t(\omega); \quad \omega \in \Omega. \tag{1.20}$$

On the other hand, fixing $\omega \in \Omega$ we can consider the function

$$t \to X_t(\omega); \quad t \in T. \tag{1.21}$$

which is called a *path* of X_t.
It may be useful for the intuition to think of t as time and each ω as an individual particle or experiment. With this picture $X_t(\omega)$ would represent the position (or result) at time t of the particle (experiment) ω. Sometimes it is convenient to write $X(t,\omega)$ instead of $X_t(\omega)$. Thus we may also regard the process as a function of two variables

$$(t,\omega) \to X(t,\omega) \tag{1.22}$$

from $T \times \Omega$ into $\mathbb{R}^n$. This is often a natural point of view in stochastic analysis, because (as we shall see) there it is crucial to have $X(t,\omega)$ jointly measurable in (t,ω).
Finally we note that we may identify each ω with the function $t \to X_t(\omega)$ from T into $\mathbb{R}^n$. Thus we may regard Ω as a subset of the space $\tilde{\Omega} = (\mathbb{R}^n)^T$ of all functions from T into $\mathbb{R}^n$. Then the σ-algebra $\mathcal{F}$ will contain the σ-algebra $\mathcal{B}$ generated by sets of the form

$$\{\omega;\ \omega(t_1) \in F_1, \cdots, \omega(t_k) \in F_k\}, \quad F_i \subset \mathbb{R}^n \text{Borel sets} \tag{1.23}$$

($\mathcal{B}$ is the same as the Borel σ-algebra on $\tilde{\Omega}$ if $T = [0, \infty)$ and $\tilde{\Omega}$ is given the product topology). Therefore one may also adopt the point of view that a stochastic process is *a probability measure P on the measurable space* $((\mathbb{R}^n)^T, \mathcal{B})$. The *(finite dimensional) distributions* of the process $X = \{X_t\}_{t\in T}$ are the measures $\mu_{t_1,\dots t_k}$ defined on $\mathbb{R}^{nk}$, $k = 1, 2, \dots$ by

$$\mu_{t_1,\dots t_k}(F_1 \times \cdots \times F_k) = \mathbb{P}[X_{t_1} \in F_1, \cdots, X_{t_k} \in F_k] \quad t_i \in T. \tag{1.24}$$

Here $F_1, \cdots, F_k$ denote Borel sets in $\mathbb{R}^n$.
The family of all finite-dimensional distributions determine many (but not all) important properties of the process X.
Conversely, given a family $\nu_{t_1,\dots,t_k}$; $k \in \mathbb{N}$, $t_i \in T$ of probability measures on $\mathbb{R}^{nk}$ it is important to be able to construct a stochastic process $Y = \{Y_t\}_{t\in T}$ having $\nu_{t_1,\dots,t_k}$ as its finite-dimensional distributions. One of Kolmogorovs famous theorems states that this can be done provided $\nu_{t_1,\dots,t_k}$ satisfies two natural consistency conditions:

Theorem 11.1.8 (Kolmogorovs extension theorem)**.** *For all $t_1, \dots, t_k \in T$, $k \in \mathbb{N}$ let $\nu_{t_1,\dots,t_k}$ be the probability measures on $\mathbb{R}^{nk}$ such that*

$$\nu_{t_{\sigma(1)},\dots,t_{\sigma(k)}}(F_1 \times \cdots \times F_k) = \nu_{t_1,\dots,t_k}(F_{\sigma^{-1}(1)} \times \cdots \times F_{\sigma^{-1}(k)}) \tag{1.25}$$

for all permutations σ on $\{1, 2, \dots, k\}$ and

$$\nu_{t_1,\dots,t_k}(F_1 \times \cdots \times F_k) = \nu_{t_1,\dots,t_k,t_{k+1},\dots,t_{k+m}}(F_1 \times \cdots \times F_k \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n) \tag{1.26}$$

for all $m \in \mathbb{N}$ where the set on the right hand side has a total of $k + m$ factors.
Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$and a stochastic process $\{X_t\}$ on Ω, $X_t : \Omega \to \mathbb{R}^n$, such that

$$\nu_{t_1,\dots,t_k}(F_1 \times \cdots \times F_k) = \mathbb{P}[X_{t_1} \in F_1, \cdots, X_{t_k} \in F_k], \tag{1.27}$$

for all $t_i \in T$, $n \in \mathbb{N}$ and all Borel sets F_i.

11.1.1 Brownian Motion

In 1828 the Scottish botanist Robert Brown observed that pollen grains suspended in liquid performed an irregular motion. The motion was later explained by the random collisions with the molecules of the liquid. To describe the motion mathematically it is natural to use the concept of a stochastic process $B_t(\omega)$, interpreted as the position at time t of the pollen grain ω. We will generalize slightly and consider an n-dimensional analog.

To construct $\{B_t\}_{t0}$ it suffices, by the Kolmogorov extension theorem, to specify a family $\{\nu_{t_1,\ldots,t_k}\}$ of probability measures satisfying 1.25 and 1.26. These measures will be chosen so that they agree with our observations of the pollen grain behaviour:
Fix $x \in \mathbb{R}^n$ and define

$$p(t,x,y) = (2\pi t)^{\frac{n}{2}} \cdot e^{\frac{|x-y|^2}{2t}} \quad \text{for } y \in \mathbb{R}^n \; t > 0. \tag{1.28}$$

If $0 \leq t_1 \leq t_2 \leq \cdots t_k$ define a measure $\nu_{t_1,\ldots,t_k}$ on $\mathbb{R}^{nk}$ by

$$\nu_{t_1,\ldots,t_k}(F_1 \times \cdots \times F_k) = \int_{F_1\times\cdots\times F_k} p(t_1,x,x_1)p(t_2-t_1,x_1,x_2)\cdots p(t_k-t_{k-1},x_{k-1},x_k)dx_1\cdots dx_k \tag{1.29}$$

where we use the notation $dy = dy_1 \cdots dy_k$ for Lebesgue measure and the convention that $p(0,x,y)dy = \delta_x(y)$, the unit point mass at x. Extend this definition to all finite sequences of t_i's by using 1.25. Since $\int_{\mathbb{R}^n} p(t,x,y)dy = 1$ for all $t \geq 0$, 1.26 holds, so by Kolmogorovs theorem there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a stochastic process $\{B_t\}_{t0}$ on Ω such that the nite-dimensional distributions of B_t are given by 1.29, i.e.

$$\mathbb{P}^x(B_{t_1} \in F_1, \cdots, B_{t_k} \in F_k) = \int_{F_1\times\cdots\times F_k} p(t_1,x,x_1)p(t_2-t_1,x_1,x_2)\cdots p(t_k-t_{k-1},x_{k-1},x_k)dx_1\cdots dx_k \tag{1.30}$$

Definition 11.1.9. *Such a process is called (a version of) Brownian motion starting at x (observe that $\mathbb{P}^x(B_0 = x) = 1$).*

The Brownian motion thus defined is not unique, i.e. there exist several quadruples $(B_t, \Omega, \mathcal{F}, \mathbb{P})$ such that 1.30 holds. However, for our purposes this is not important, we may simply choose any version to work with. As we shall soon see, the paths of a Brownian motion are (or, more correctly, can be chosen to be) continuous, a.s. Therefore we may identify (a.a.) $\omega \in \Omega$ with a continuous function $t \to B_t(\omega)$ from $[0,\infty)$ into $\mathbb{R}^n$. Thus we may adopt the point of view that Brownian motion is just the space $C([0,\infty),\mathbb{R}^n)$ equipped with certain probability measures $\mathbb{P}^x$ (given by 1.29 and 1.30 above).This version is called the *canonical* Brownian motion. Besides having the advantage of being intuitive, this point of view is useful for the further analysis of measures on $C([0,\infty),\mathbb{R}^n)$, since this space is Polish (i.e. a complete separable metric space). We state some basic properties of Brownian motion:

1. B_t is a *Gaussian process*, i.e. for all $0 \leq t_1 \leq \cdots \leq t_k$ the random variable $Z = (B_{t_1}, \cdots, B_{t_k}) \in \mathbb{R}^{nk}$ has a *(multi)normal distribution*. This means that there exists a vector $M \in \mathbb{R}^{nk}$ and a non-negative definite matrix $C = [c_{jm}] \in \mathbb{R}^{nk \times nk}$ (the set of all $nk \times nk$-matrices with real entries) such that

$$\mathbb{E}^x \left[\exp\left(i \sum_{j=1}^{nk} u_j Z_j \right) \right] = \exp\left(-\frac{1}{2} \sum_{j,m} u_j c_{jm} u_m + i \sum_j u_j M_j \right) \tag{1.31}$$

for all $u = (u_1, ..., u_{nk}) \in \mathbb{R}^{nk}$, where $i = \sqrt{-1}$ is the imaginary unit and $\mathbb{E}^x$ denotes expectation with respect to $mathbbP^x$. Moreover, if 1.31 holds then

$$M = \mathbb{E}^x[Z] \quad \text{is the mean value of } Z \tag{1.32}$$

and

$$c_{jm} = \mathbb{E}^x[(Z_j - M_j)(Z_m - M_m)] \quad \text{is the covariance matrix of } Z. \tag{1.33}$$

To see that 1.31 holds for $Z = (B_{t_1}, ..., Bt_k)$ we calculate its left hand side explicitly by using 1.30 and obtain 1.31 with

$$M = \mathbb{E}^x[Z] = (x, x, ..., x) \in \mathbb{R}^{nk} \tag{1.34}$$

and

$$\boldsymbol{A} = \begin{bmatrix} t_1 I_n & t_1 I_n & \cdots & t_1 I_n \\ t_1 I_n & t_2 I_n & \cdots & t_2 I_n \\ \vdots & \vdots & & \vdots \\ t_1 I_n & t_2 I_n & \cdots & t_k I_n \end{bmatrix}. \tag{1.35}$$

Hence

$$\mathbb{E}^x[B_t] = x \quad \text{for all } t \geq 0 \tag{1.36}$$

and

$$\begin{aligned} \mathbb{E}^x[(B_t - x)^2] &= nt \\ \mathbb{E}^x[(B_t - x)(B_s - x)] &= n\min(s,t). \end{aligned} \tag{1.37}$$

Moreover

$$\mathbb{E}^x[(B_t - B_s)^2] = n(t - s) \quad \text{if } t \geq s, \tag{1.38}$$

since

$$\begin{aligned}\mathbb{E}^x[(B_t - B_s)^2] &= \mathbb{E}^x[(B_t - x)^2 - 2(B_t - x)(B_s - x) + (B_s - x)^2] \\ &= n(t - 2s + s) = n(t - s), \quad \text{when } t \geq s\end{aligned} \tag{1.39}$$

2. B_t has *independent increments*, i.e.

$$B_{t_1}, B_{t_2} - B_{t_1}, ..., B_{t_k} - B_{t_{k-1}} \tag{1.40}$$

are independent for all $0 \leq t_1 \leq \cdots \leq t_k$. To prove this we use the fact that normal random variables are independent if they are uncorrelated. So it is enough to prove that

$$\mathbb{E}^x[(B_{t_i} - B_{t_{i-1}})(B_{t_j} - B_{t_{j-1}})] = 0 \quad \text{when } t_i < t_j \tag{1.41}$$

which follows from the form of C:

$$\begin{aligned}\mathbb{E}^x[B_{t_i}B_{t_j} - B_{t_i}B_{t_{j-1}} - B_{t_{i-1}}B_{t_j} + B_{t_{i-1}}B_{t_{j-1}}] = \\ n(t_i - t_{i-1} - t_i + t_{i-1}) = 0.\end{aligned} \tag{1.42}$$

From this we deduce that B_sB_t is independent of $\mathcal{F}_t$ if $s > t$.

3. Brownian motion has a *continuous version*. To prove that we need the following important concept:

Definition 11.1.10. *Suppose that $\{X_t\}$ and $\{Y_t\}$ are stochastic processes on $(\Omega, \mathcal{F}, \mathbb{P})$. Then we say that $\{X_t\}$ is a version of (or a modification of) $\{Y_t\}$ if*

$$\mathbb{P}(\{\omega;\ X_t(\omega) = Y_t(\omega)\}) = 1 \quad \text{for all } t. \tag{1.43}$$

Note that if $\{X_t\}$ is a version of $\{Y_t\}$, then $\{X_t\}$ and$\{Y_t\}$ have the same finite-dimensional distributions. Thus from the point of view that a stochastic process is a probability law on $(\mathbb{R}^n)^{[0,\infty)}$ two such processes are the same, but nevertheless their path properties may be different.

The continuity question of Brownian motion can be answered by using another famous theorem of Kolmogorov:

Theorem 11.1.11 (Kolmogorovs continuity theorem)**.** *Suppose that the process $X = \{X_t\}_{t\geq 0}$ satisfies the following condition: For all $T > 0$ there exist positive constants α, β, D such that*

$$\mathbb{E}[|X_t - X_s|^{\alpha}] \leq D \cdot |t-s|^{1+\beta}; \quad 0 \leq s\ t \leq T. \tag{1.44}$$

Then there exists a continuous version of X.

For Brownian motion B_t it is not hard to prove that

$$\mathbb{E}[|B_t - B_s|^4] = n(n+2)|t-s|^2. \tag{1.45}$$

So Brownian motion satises Kolmogorovs condition 1.44 with $\alpha = 4, D = n(n+2)$ and $\beta = 1$, and therefore it has a continuous version. From now on we will assume that B_t is such a continuous version.

Finally we note that if $B_t = (B_t^{(1)}, ..., B_t^{(n)})$ is n-dimensional Brownian motion, then the 1-dimensional processes $\{B_t^{(j)}\}_{t\geq 0}$, $1 \leq j \leq n$ are independent.

Definition 11.1.12. *Denote $\Lambda_B^p([\alpha,\beta])$ the equivalence class space of real process $X = (\Omega, \mathcal{F}, (\mathcal{F}_t)_{\alpha\leq t\leq\beta}, (X_t)_{\alpha\leq t\leq\beta}, \mathbb{P})$ progressively measurable such that*

$$\mathbb{P}\Big(\int_{\alpha}^{\beta} |X_s|^p ds < +\infty\Big) = 1. \tag{1.46}$$

Definition 11.1.13. *$M_B^p([\alpha,\beta])$ is the equivalence class space of real process progressively measurable such that*

$$\mathbb{E}\Big[\int_{\alpha}^{\beta} |X_s| ds\Big] < +\infty \tag{1.47}$$

11.1.2 Exchangeability

In probability theory, the random variables $X_1, ..., X_N$ are said to be **exchangeable** (or permutable or symmetric) if their joint distribution $F(x_1, ..., x_N)$ is symmetric; that is, if F is invariant under permutation of its arguments, so that

$$F(z_1, ..., z_N) = F(x_1, ...x_N) \tag{1.48}$$

whenever $z_1, ..., z_N$ is a permutation of $x_1, ..., x_N$ [260] [261].
Exchangeable random variables are identically distributed, and iid variables are exchangeable. Now suppose that $X_1, ..., X_N$ are iid given an unknown parameter θ that

indexes their joint distribution. Such variables will not be unconditionally independent when θ is a random variable, but will be exchangeable. Consider, for example, the case in which $X_1, ..., X_N$ have a joint density. The unconditional density of $X_1, ..., X_N$ will be

$$f(x_1, ..., x_N) = \int_\theta f(x_1, ..., x_N|\theta)dF(\theta) = \int_\theta \prod_i f(x_i|\theta)dF(\theta) \tag{1.49}$$

Exchangeability of $X_1, ..., X_N$ follows from the identity of the marginal densities in the product. However, given that these densities depend on θ, the integral and product cannot be interchanged, so that $f(x_1, ..., x_N) \neq \prod_i f(x_i)$ We thus have that a mixture of iid sequences is an exchangeable sequence, but not iid except in trivial cases.

11.2 Ito Integrals

11.2.1 Construction of the Ito Integral

We now turn to the question of finding a reasonable mathematical interpretation of the *noise* term in the equation 2.50

$$\frac{dX}{dt} = b(t, X_t) + \sigma(t, X_t) \cdot noise \tag{2.50}$$

where b and σ are some given functions. Let us first concentrate on the case when the noise is 1-dimensional. It is reasonable to look for some stochastic process W_t to represent the noise term, so that

$$\frac{dX}{dt} = b(t, X_t) + \sigma(t, X_t) \cdot W_t \tag{2.51}$$

Based on many situations, for example in engineering, one is led to assume that W_t has, at least approximately, these properties:

1. $t_1 \neq t_2 \Rightarrow W_{t_1}$ and W_{t_2} are independent.
2. $\{W_t\}$ is stationary.
3. $\mathbb{E}[W_t] = 0$ for all t.

However, it turns out there does not exist any *reasonable* stochastic process satisfying (1.) and (2.): Such a W_t cannot have continuous paths. If we require $\mathbb{E}[W_{t_2}] = 1$ then the function $(t,\omega) \to W_t(\omega)$ cannot even be measurable, with respect to the σ-algebra $\mathcal{B} \times \mathcal{F}$, where $\mathcal{B}$ is the Borel σ-algebra on $[0, \infty)$. Nevertheless it is possible to represent W_t as a generalized stochastic process called the *white noise process.* That the process is *generalized* means that it can be constructed as a probability measure on the space $\mathcal{S}'$ of tempered distributions on $[0,\infty)$, and not as a probability measure on the much smaller space $\mathbb{R}^{[0,\infty)}$, like an ordinary process can. We will avoid this kind of construction and rather try to rewrite equation 2.51 in a form that suggests a replacement of W_t by a proper stochastic process. Let $0 = t_0 < t_1 < \cdots < t_m = t$ and consider a discrete version of 2.51:

$$X_{k+1} - X_k = b(t_k, X_k)\Delta t_k + \sigma(t_k, X_k)W_k\Delta t_k \tag{2.52}$$

where

$$X_j = X(t_j), \quad W_k = W_{t_k}, \quad \Delta t_k = t_{k+1} - t_k. \tag{2.53}$$

We abandon the W_k-notation and replace $W_k\Delta t_k$ byΔ $V_k = V_{t_{k+1}} - V_{t_k}$, where $\{V_t\}_{t\geq 0}$ is some suitable stochastic process. The assumptions (1.), (2.) and (3.) on W_t suggest that V_t should have *stationary independent increments with mean* 0. It turns out that the only such process with continuous paths is the Brownian motion B_t. Thus we put $V_t = B_t$ and obtain from 2.52

$$X_k = X_0 + \sum_{j=0}^{k-1} b(t_j, X_j)\Delta t_j + \sum_{j=0}^{k-1} \sigma(t_j, X_j)\Delta B_j. \tag{2.54}$$

It possible to prove that the limit of the right hand side of 2.54 exists whenΔ $t_j \to 0$. Then by applying the usual integration notation we should obtain

$$X_t = X_0 + \int_0^t b(t_s, X_s)ds + \int_0^t \sigma(t_s, X_s)dB_s \tag{2.55}$$

and we would adopt as a convention that 2.51 really means that $X_t = X_t(\omega)$ is a stochastic process satisfying 2.55. Suppose $0 \leq S < T$ and $f(t,\omega)$ is given so we want to define

$$\int_S^T f(t,\omega)dB_t(\omega) \tag{2.56}$$

Definition 11.2.1. *Let $\mathcal{V} = \mathcal{V}(S,T)$ be the class of functions*

$$f(t,\omega) : [0,\infty) \times \Omega \to \mathbb{R} \tag{2.57}$$

such that

(i) *$(t,\omega) \to f(t,\omega)$ is $\mathcal{B}\times\mathcal{F}$-measurable, where $\mathcal{B}$ denotes the Borel σ-algebra on $[0,\infty)$.*

(ii) $f(t,\omega)$ *is* $\mathcal{F}_t$*-adapted.*

(iii) $\mathbb{E}[\int_S^T f(t,\omega)dt] < \infty$.

Definition 11.2.2 (The Ito integral). *Let* $f \in \mathcal{V}(S,T)$. *Then the Ito integral of f (from S to T) is defined by*

$$\int_S^T f(t,\omega)dB_t(\omega) = \lim_{n\to\infty} \int_S^T \phi_n(t,\omega)dB_t(\omega) \tag{2.58}$$

limit in $L^2(P)$, *where* $\{\phi_n\}$ *is a sequence of elementary functions such that*

$$\mathbb{E}\Big[\int_S^T f(t,\omega) - \phi_n(t,\omega)dt\Big] \to 0 \quad \text{as } n \to \infty. \tag{2.59}$$

It can be proven that such a sequence $\{\phi_n\}$ satisfying 2.59 exists. Moreover, the limit in 2.58 exists and does not depend on the actual choice of $\{\phi_n\}$, as long as 2.59 holds [50].

11.2.2 Some properties of the Ito integral

First we observe the following:

Theorem 11.2.3. *Let* $f, g \in \mathcal{V}(0,T)$ *and let* $0 \le S < U < T$. *Then*

(i) $\int_S^T f dB_t = \int_S^U f dB_t + \int_U^T f dB_t$ *for a.a.* ω

(ii) $\int_S^T (cf + g)dB_t = c \cdot \int_S^T f dB_t + \int_S^T g dB_t$ *c constant for a.a.* ω

(iii) $\mathbb{E}[\int_S^T f dB_t] = 0$

(iv) $\int_S^T f dB_t$ *is* $\mathcal{F}_t$*-measurable.*

An important property of the Ito integral is that it is a *martingale.*

Definition 11.2.4 (Martingale). *A filtration (on* $(\Omega, \mathcal{F})$*) is a family* $\mathcal{M} = \{\mathcal{M}_t\}_{t\ge 0}$ *of* σ*-algebras* $\mathcal{M}_t \subset \mathcal{F}$ *such that*

$$0 \le s < t \Rightarrow \mathcal{M}_s \subset \mathcal{M}_t \tag{2.60}$$

An n-dimensional stochastic process $\{M_t\}_{t\ge 0}$ *on* $(\Omega, \mathcal{F}, \mathbb{P})$ *is called a martingale with respect to a filtration* $\{\mathcal{M}_t\}_{t\ge 0}$ *and with respect to* $\mathbb{P}$ *if*

(i) $\{M_t\}$ *is* $\{\mathcal{M}_t\}$*-measurable for all t*

(ii) $\mathbb{E}[|M_t|] < \infty$ *for all* t

(iii) $\mathbb{E}[M_s|\mathcal{M}_t] = M_t$ *for all* $s \geq t$

Here the expectation in (ii) and the conditional expectation in (iii) is taken with respect to $\mathbb{P} = \mathbb{P}^0$.

Let $f(t,\omega) \in \mathcal{V}(0,T)$ for all T. Then

$$M_t(\omega) = \int_0^t f(s,\omega)dB_s \tag{2.61}$$

is a martingale with respect to a filtration $\mathcal{F}_t$ and

$$\mathbb{P}\Big[\sup_{0\leq t\leq T} |M_t| \geq \lambda\Big] \leq \frac{1}{\lambda^2} \cdot \mathbb{E}\Big[\int_0^T f(s,\omega)^2 ds\Big]; \quad \lambda, T > 0 \tag{2.62}$$

for demonstration see [262].

11.3 The Ito Formula

Definition 11.3.1 (Ito Process). *Let* B_t *be* 1*-dimensional Brownian motion on* $(\Omega, \mathcal{F}, \mathbb{P})$. *A (*1*-dimensional) Ito process (or stochastic integral) is a stochastic process* X_t *on* $(\Omega, \mathcal{F}, \mathbb{P})$ *of the form*

$$X_t = X_0 + \int_0^t u(s,\omega)ds + \int_0^t v(s,\omega)dB_s \tag{3.63}$$

where $v \in \mathcal{W}_{\mathcal{H}}$ *so that*

$$\mathbb{P}\Big[\int_0^t v(s,\omega)^2 ds < \infty\Big] = 1 \quad \text{for all } t \geq 0 \tag{3.64}$$

we also assume that u *is* $\mathcal{H}_t$*-adapted (where* $\mathcal{H}_t$ *is increasing family of* σ*-algebras such that* B_t *is a martingale with respect to* $\mathcal{H}_t$*) and*

$$\mathbb{P}\Big[\int_0^t |u(s,\omega)|ds < \infty\Big] = 1 \quad \text{for all } t \geq 0 \tag{3.65}$$

If X_t is an Ito process of the form 3.63 the equation 3.63 is sometimes written in the shorter differential form

$$dX_t = udt + vdB_t \tag{3.66}$$

Theorem 11.3.2 (The 1-dimensional Ito formula). *Let X_t be an Ito process given by*

$$dX_t = udt + vdB_t. \tag{3.67}$$

Let $g(t,x) \in C^2([0,\infty) \times \mathbb{R})$ (i.e. g is twice continuously differentiable on $[0,\infty) \times \mathbb{R}$). Then

$$Y_t = g(t, X_t) \tag{3.68}$$

is again an Ito process, and

$$dY_t = \frac{\partial g}{\partial t}(t, X_t)dt + \frac{\partial g}{\partial x}(t, X_t)dX_t + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}(t, X_t)(dX_t)^2 \tag{3.69}$$

where $(dX_t)^2 = (dX_t)(dX_t)$ is computed according to the rules

$$dt \cdot dt = dt \cdot dB_t = dB_t \cdot dt = 0, \quad dB_t \cdot dB_t = dt \tag{3.70}$$

11.3.1 The Multi-dimensional Ito Formula

We now turn to the situation in higher dimensions. Let $B(t,\omega) = (B_1(t,\omega), ..., B_m(t,\omega))$ denote m-dimensional Brownian motion. If each of the processes $u_i(t,\omega)$ and $v_{ij}(t,\omega)$ is an Ito process ($1 \le i \le n, 1 \le j \le m$) then we can form the following n Ito processes

$$\begin{cases} dX_1 = u_1 dt + v_{11} dB_1 + \cdots + v_{1m} dB_m \\ \quad \vdots \\ dX_n = u_n dt + v_{n1} dB_1 + \cdots + v_{nm} dB_m \end{cases} \tag{3.71}$$

Or, in matrix notation simply

$$dX_t = udt + vdB_t, \tag{3.72}$$

where

$$X_t = \begin{pmatrix} X_1(t) \\ \vdots \\ X_n(t) \end{pmatrix}, \quad u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}, \quad v = \begin{pmatrix} u_{11} & \cdots & u_{1m} \\ \vdots & & \vdots \\ u_{n1} & \cdots & u_{nm} \end{pmatrix}, \quad dB(t) = \begin{pmatrix} dB_1(t) \\ \vdots \\ dB_n(t) \end{pmatrix} \tag{3.73}$$

Such a process $X(t)$ is called an n-**dimensional Ito process** (or just an Ito process).

Theorem 11.3.3 (The general Ito formula). *Let*

$$dX(t) = udt + vdB(t) \tag{3.74}$$

be an n-dimensional Ito process as above. Let $g(t,x) = (g_1(t,x),...,g_p(t,x))$ be a C^2 map from $[0,\infty) \times \mathbb{R}^n$ into $\mathbb{R}^p$. Then the process

$$Y(t,\omega) = g(t, X(t)) \tag{3.75}$$

is again an Ito process, whose component number k, Y_k , is given by

$$dY_k = \frac{\partial g_k}{\partial t}(t,X)dt + \sum_i \frac{\partial g_k}{\partial x_i}(t,X)dX_i + \frac{1}{2}\sum_{i,j}\frac{\partial^2 g_k}{\partial x_i \partial x_j}(t,X)dX_i dX_j \tag{3.76}$$

12

Appendix B

In the following section are provided some general results of system (2.43)-(2.44) of Chapter 8.

Let $\mu(\mathbf{x}) = \big(\mu_i(\mathbf{x})\big)_{1\leq i\leq n}$ and $\sigma(\mathbf{x}) = \big(\sigma_{ij}(\mathbf{x})\big)_{1\leq i\leq n, 1\leq j\leq m}$ be measurable functions defined on $D \subseteq \mathbb{R}^n$ with values on $\mathbb{R}^n$ and $\mathbb{R}^{n\ timesm}$. Let also W_t be a $m-$dimensional Winer process defined on $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$. We consider the following system of differential stochastic equations

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t. \tag{0.1}$$

The following Kolmogorov operator is associated with the system

$$\begin{aligned} Lg(\mathbf{x}) =& \sum_{i=1}^{n} \partial_{x_i} g(\mathbf{x})\mu_i(\mathbf{x}) + \frac{1}{2}\sum_{i,j=1}^{n} q_{ij}(\mathbf{x})\partial_{x_i,x_j} g(x) \\ =& \langle \nabla g(\mathbf{x}), \mu(\mathbf{x})\rangle + \frac{1}{2} Tr\big[q(\mathbf{x})\nabla^2 g(\mathbf{x})\big], \quad \mathbf{x} \in \mathbb{R}^n \end{aligned} \tag{0.2}$$

where $g : D \subseteq \mathbb{R}^n \to \mathbb{R}$, with $g \in C^2(D)$ e $q(\mathbf{x}) = \sigma(\mathbf{x})\sigma^T(\mathbf{x})$ and ∂_{x_i} is the partial derivative with respect to x_i.

Definition 12.0.4. *Let the* **Lyapunov generalized function** *of (0.1) be the function* $V : D \subseteq \mathbb{R}^n \to \mathbb{R}$ *that satisfies the following condition:*

(i) $V(\mathbf{x}) \geq 0, \quad \forall\ \mathbf{x} \in D,$ *e* $\{\mathbf{x} \in D : V(\mathbf{x}) = 0\}$ *is limited.*

(ii) $\exists\ \zeta > 0 : LV(\mathbf{x}) \leq \zeta\big(1 + V(\mathbf{x})\big), \quad \forall\ \mathbf{x} \in D.$

(iii) $\lim_{\mathbf{x}\to\widehat{\mathbf{x}}} V(\mathbf{x}) = +\infty, \quad \forall\ \widehat{\mathbf{x}} \in \partial D,$ *e* $\lim_{\|\mathbf{x}\|\to+\infty} V(\mathbf{x}) = +\infty, \quad \forall\ \mathbf{x} \in D$

Where ∂D is the frontier of D.

Theorem 12.0.5. *Consider the differential stochastic equations system (2.43). Let $D = [0, +\infty] \times [0, +\infty]$ and $(\tilde{x}, \tilde{y}) = (\frac{u}{cq}, \frac{r}{q} - \frac{ru}{cq^2})$ the equilibrium point of the deterministic model associated to (2.43) (where $\varepsilon = \eta = 0$).*
The function $V(x, y) : D \subset \mathbb{R}^2 \to \mathbb{R}$

$$V(x, y) = V_1(x) + kV_2(y) \tag{0.3}$$

where $k \in (0, \frac{1}{c}]$ and

$$V_1(x) = x - \tilde{x} - \tilde{x}\log\frac{x}{\tilde{x}}$$
$$V_2(y) = y - \tilde{y} - \tilde{y}\log\frac{y}{\tilde{y}}$$

is a Lyapunov generalized function of the system (2.43).

Proof. In order to prove that (0.3) is a Lyapunov generalized function of the system (2.43), the properties of Definition 12.0.4 must hold. Given that $(\tilde{x}, \tilde{y})$ is the absolute minimum in which $V(\tilde{x}, \tilde{y}) = 0$, therefore *(i)* e la *(iii)* are verified.
Property *(ii)* has to be verified. Given that

$$\nabla V(x, y) = \begin{pmatrix} 1 - \frac{\tilde{x}}{x} \\ k\left(1 - \frac{\tilde{y}}{y}\right) \end{pmatrix}, \quad \nabla^2 V(x, y) = \begin{pmatrix} \frac{\tilde{x}}{x^2} & 0 \\ 0 & k\frac{\tilde{y}}{y^2} \end{pmatrix},$$

$$\mu(x, y) = \begin{pmatrix} rx(1 - x) - qxy \\ cqxy - uy \end{pmatrix},$$

$$\sigma(x, y) = \begin{pmatrix} \varepsilon x\rho & \varepsilon x\sqrt{1 - \rho^2} \\ \eta y & 0 \end{pmatrix},$$

$$q(x, y) = \sigma(x, y)\sigma^T(x, y) = \begin{pmatrix} \varepsilon^2 x^2 & \varepsilon\eta\rho xy \\ \varepsilon\eta\rho xy\eta & {}^2y^2 \end{pmatrix},$$

using the Kolmogorov operator (0.2) on V

$$\begin{aligned} LV(x, y) \quad &= rx - r\tilde{x} + \tfrac{\varepsilon^2\tilde{x}}{2} - rx^2 + \tfrac{k\eta^2\tilde{y}}{2} - kuy + \\ &+ku\tilde{y} - qxy + q\tilde{x}y + rx\tilde{x} + ckqxy - ckqx\tilde{y}. \end{aligned} \tag{0.4}$$

Increasing equation 0.4)

$$\begin{aligned} LV(x, y) &\leq rx + \frac{\varepsilon^2\tilde{x}}{2} + \frac{k\eta^2\tilde{y}}{2} + ku\tilde{y} + q\tilde{x}y + rx\tilde{x} + qxy(ck - 1) \\ &\leq xr(1 + \tilde{x}) + yq\tilde{x} + \frac{\varepsilon^2\tilde{x}}{2} + \frac{k\eta^2\tilde{y}}{2} + ku\tilde{y} \\ &= Ax + By + C, \end{aligned} \tag{0.5}$$

where

$$A = r(1+\tilde{x}), \quad B = q\tilde{x}, \quad C = \frac{\varepsilon^2\tilde{x}}{2} + \frac{k\eta^2\tilde{y}}{2} + ku\tilde{y}.$$

Then we need that the following relations hold:

$$\exists A_1 > 0, C_1 > 0 : x \leq A_1 V_1(x) + C_1, \ \forall\ x > 0, \tag{0.6}$$

$$\exists B_2 > 0, C_2 > 0 : x \leq B_2 V_2(y) + C_2, \ \forall\ y > 0, \tag{0.7}$$

where A_1, C_1, B_2 e C_2 have to be defined.
Choosing $f(x) = A_1V_1(x) - x$, we find the constants imposing that

$$f(x) = A_1\left(x - \tilde{x} - \tilde{x}\log\frac{x}{\tilde{x}}\right) - x \geq -C1. \tag{0.8}$$

Therefore we find the minimum $\bar{x}$ setting the value of the derivative of $f(x)$ equal to 0:

$$\dot{f}(x) = \frac{A_1\big(x-\tilde{x}\big) - x}{x} = 0 \quad \Rightarrow \quad \bar{x} = \frac{\tilde{x}A_1}{A_1 - 1}. \tag{0.9}$$

Replacing $\bar{x}$ in (0.8)

$$f(x) = A_1\left(\frac{\tilde{x}A_1}{A_1-1} - \tilde{x} - \tilde{x}\log\frac{A_1}{A_1-1}\right) - \frac{\tilde{x}A_1}{A_1-1} \geq -C1, \tag{0.10}$$

hence choosing

$$C_1 = \tilde{x}A_1\log\frac{A_1}{A_1-1}, \ \ A_1 > 1 \tag{0.11}$$

it is satisfied equation (0.6).

Setting

$$A_1 = 2, \quad C_1 = 2\tilde{x}\log 2.$$

In the same way constants B_2 e C_2 are found equal to

$$B_2 = 2, \quad C_2 = 2\tilde{y}\log 2.$$

Considering (0.6) and (0.7), equation (0.5) becomes:

$$\begin{aligned} LV(x,y) &\leq Ax + By + C \\ &\leq A\big(A_1V_1(x) + C_1\big) + B\big(B_2V_2(y) + C_2\big) + C \\ &= \bar{A}V_1(x) + \bar{B}V_2(y) + \bar{C} \\ &\leq \zeta\big(1 + \underbrace{V_1(x) + kV_2(y)}_{V(x,y)}\big), \end{aligned} \tag{0.12}$$

and choosing

$$\begin{aligned} \bar{A} &= AA_1, \quad \bar{B} = BB_2, \\ \bar{C} &= AC_1 + BC_2 + C, \\ \zeta &= \max\Big\{\bar{A}, \frac{\bar{B}}{k}, \bar{C}\Big\}. \end{aligned} \tag{0.13}$$

Hence property *(ii)* of Definition (12.0.4) and Theorem (12.0.5) are proved. □

Given the *generalized Lyapunov function*, it is proved the existence and uniqueness of the solution of system (2.43). Recalling the following relation

Lemma 12.0.6. *[50] (Gronwall). Let u, v be real non-negative function defined on $[\alpha,\beta]$, where u is integrable and v measurable and limited; let $c \in \mathbb{R}^+$ and assuming that*

$$v(t) \leq c + \int_\alpha^t u(s)v(s)ds, \quad \forall\ t \in [\alpha,\beta]. \tag{0.14}$$

Then, $\forall\ t \in [\alpha,\beta]$, we have:

$$v(t) \leq c \ \varepsilon\ e^{\int_\alpha^t u(s)ds}. \tag{0.15}$$

Let $T > 0$, $\mu : D \subset \mathbb{R}^n \to \mathbb{R}^n$, $\sigma : D \subset \mathbb{R}^n \to \mathbb{R}^{n\times m}$ continuous function. Assuming the following hypothesis:

(a) *(σ in D is globally Lipschitz)* $\exists\ L_\sigma > 0$ such that $\forall\ \mathbf{x}_1, \mathbf{x}_2 \in D$, then:

$$\|\sigma(\mathbf{x}_1) - \sigma(\mathbf{x}_2)\| \leq L_\sigma \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

(b) *(μ is locally Lipschitz)* $\forall\ D' \subset D$ limited, exists $L'_\mu > 0$ such that $\forall\ \mathbf{x}_1, \mathbf{x}_2 \in D'$, then:

$$\|\mu(\mathbf{x}_1) - \mu(\mathbf{x}_2)\| \leq L'_\mu \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

(c) Exists a generalized Lyapunov function of system (0.1) associated with μ e σ.

Theorem 12.0.7. *Let $x_0 \in D$ an W standard m-dimensional Wiener process. According to hypothesis* **(a)**, **(b)**, **(c)** *the differential stochastic equation*

$$\begin{cases} dX_t = \mu(X_t)dt + \sigma(X_t)dW_t, & 0 \leq t \leq T \\ X_0 = x_0 \end{cases} \tag{0.16}$$

has unique solution $X_t(\omega)$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$, adapted to the filtration $\{\mathcal{F}_t\}_{t\in[0,T]}$, such that

$$\mathbb{P}(X_t(\omega) \in D,\ \forall\ t \in [0,T]) = 1 \tag{0.17}$$

Remark 12.0.8. *With unique solution we mean that if $X_t^{(1)}, X_t^{(2)}$ satisfy (0.16), then*

$$\mathbb{P}(X_t^{(1)} = X_t^{(2)},\ \forall\, t \in [0,T]) = 1. \tag{0.18}$$

Corollary 12.0.9. *The solution of system (2.43) exists and is unique with strictly positive trajectories for all time.*

Proof. Let $D = [0,+\infty] \times [0,+\infty]$. All the hypothesis of Theorem 12.0.7 are satisfied: the coefficients μ and σ satisfies hypothesis **(a)** and **(b)**, and Theorem 12.0.4 guarantees the existence of the *generalized Lyapunov function* and therefore also **(c)** is satisfied. □

Proof. (Theorem 12.0.7)
Assuming that $X_t^{(1)}$ and $X_t^{(2)}$ satisfies the system of differential stochastic equation (0.16) and satisfies also (0.18). For all $z \geq 0$, defined the set

$$D_z = \{\mathbf{x} \in D : V(\mathbf{x}) \leq z\}. \tag{0.19}$$

According to Definition (12.0.4) of V, D_z is limited set.
In such limited set, according to hypothesis **(b)**, μ has a Lipschitz constant called $L_\mu(z)$.
Given that $V \geq 0$, we set

$$z_t = \max\left\{V\left(X_t^{(1)}\right), V\left(X_t^{(2)}\right)\right\}. \tag{0.20}$$

Since $V(X_t^{(i)}) \leq z_t$, with $i = 1,2$, process $X^{(1)}$ e $X^{(2)}$ take values in the limited set D_z where μ is Lipschitz. We introduce the process

$$\Phi_t = \mathrm{e}^{-\int_0^t \xi_s ds} \left\| X_t^{(1)} - X_t^{(2)} \right\|^2, \tag{0.21}$$

where

$$\xi_t = L_\sigma + 2L_\mu(z_t). \tag{0.22}$$

ξ_t is a positive process adapted to the filtration $\{\mathcal{F}_t\}_{t\in[0,T]}$, in fact

$$L_\mu(z) = \sup\left\{\frac{\|\mu(\mathbf{x}) - \mu(\mathbf{x}')\|}{\|\mathbf{x} - \mathbf{x}'\|} : \mathbf{x}, \mathbf{x}' \in D_z, \mathbf{x} \neq \mathbf{x}'\right\} \tag{0.23}$$

is a measurable function of z, $X_t^{(1)}$ and $X_t^{(2)}$ are $\mathcal{F}_t$-measurable and V is continuous. ξ_s weights squared the norm and $\mathrm{e}^{-\int_0^t \xi_s ds} > 0$, proving that, for all t, $\mathbb{E}[\Phi_t] = 0$, we prove that the solution is unique, in fact given that the functions are measurable

$$\mathbb{P}\left(\bigcap_t \{\omega \in \Omega : X_t^{(1)}(\omega) = X_t^{(2)}(\omega)\}\right) = 1.$$

Considering the stochastic differentialΦ_t:

$$\begin{aligned} d\Phi_t &= d\left(\mathrm{e}^{-\int_0^t \xi_s ds} \left\|X_t^{(1)} - X_t^{(2)}\right\|^2\right) \\ &= d\left(\left\|X_t^{(1)} - X_t^{(2)}\right\|^2\right) \mathrm{e}^{-\int_0^t \xi_s ds} - \xi_t \mathrm{e}^{-\int_0^t \xi_s ds} \left\|X_t^{(1)} - X_t^{(2)}\right\|^2 dt. \end{aligned} \tag{0.24}$$

Set

$$f(X_t^{(1)}, X_t^{(2)}) = \left\|X_t^{(1)} - X_t^{(2)}\right\|^2,$$

then

$$\nabla f = 2\begin{pmatrix} X_t^{(1)} - X_t^{(2)} \\ -\left(X_t^{(1)} - X_t^{(2)}\right) \end{pmatrix} \in \mathbb{R}^{2n}, \quad \nabla^2 f = 2\begin{pmatrix} I_n & -I_n \\ -I_n & I_n \end{pmatrix} \in \mathbb{R}^{2n\times 2n}.$$

Considering the aggregated differential stochastic equation

$$d\begin{pmatrix} X_t^{(1)} \\ X_t^{(2)} \end{pmatrix} = \bar{\mu}(X_t^{(1)}, X_t^{(2)})dt + \bar{\sigma}(X_t^{(1)}, X_t^{(2)})dW_t, \tag{0.25}$$

where

$$\bar{\mu}(X_t^{(1)}, X_t^{(2)}) = \begin{bmatrix} \mu(X_t^{(1)}) \\ \mu(X_t^{(2)}) \end{bmatrix} \in \mathbb{R}^{2n}, \quad \bar{\sigma}(X_t^{(1)}, X_t^{(2)}) = \begin{bmatrix} \sigma(X_t^{(1)}) \\ \sigma(X_t^{(2)}) \end{bmatrix} \in \mathbb{R}^{2n\times 2n},$$

Furthermore,

$$\begin{aligned} \frac{1}{2}\bar{\sigma}^T \nabla^2 f \bar{\sigma} &= \begin{bmatrix} \sigma(X_t^{(1)}), \sigma(X_t^{(2)}) \end{bmatrix} \begin{bmatrix} I_n & -I_n \\ -I_n & I_n \end{bmatrix} \begin{bmatrix} \sigma(X_t^{(1)}) \\ \sigma(X_t^{(2)}) \end{bmatrix} \\ &= \left(\sigma(X_t^{(1)}) - \sigma(X_t^{(2)})\right)^T \left(\sigma(X_t^{(1)}) - \sigma(X_t^{(2)})\right). \end{aligned} \tag{0.26}$$

Using (0.26), we evaluate the stochastic differential $Y_t = \left\|X_t^{(1)} - X_t^{(2)}\right\|^2$ and applying the Ito's formula:

$$\begin{aligned} dY_t = &\Big\{ 2\langle \left(X_t^{(1)} - X_t^{(2)}\right), \left(\mu(X_t^{(1)}) - \mu(X_t^{(2)})\right)\rangle + \\ &+ Tr\left[\left(\sigma(X_t^{(1)}) - \sigma(X_t^{(2)})\right)^T \left(\sigma(X_t^{(1)}) - \sigma(X_t^{(2)})\right)\right] \Big\} dt + \\ &+ 2\left(X_t^{(1)} - X_t^{(2)}\right)^T \left(\sigma(X_t^{(1)}) - \sigma(X_t^{(2)})\right) dW_t. \end{aligned} \tag{0.27}$$

By placing (0.27) in (0.24) and considering the expected value$\Phi_{\ t}$, we have:

$$\mathbb{E}\left[\mathrm{e}^{-\int_0^t \xi_s ds}\ Y_t\right] = \int_0^t \mathbb{E}\Bigg[\mathrm{e}^{-\int_0^s \xi_r dr}\Big(-\xi_s Y_s + 2\langle\big(X_s^{(1)} - X_s^{(2)}\big), \big(\mu(X_s^{(1)}) - \mu(X_s^{(2)})\big)\rangle + \\ + Tr\left[\big(\sigma(X_s^{(1)}) - \sigma(X_s^{(2)})\big)^T \big(\sigma(X_s^{(1)}) - \sigma(X_s^{(2)})\big)\right]\Big)\Bigg] ds, \tag{0.28}$$

using the following relation for the stochastic integral

$$\mathbb{E}\left[\int_0^t \mathrm{e}^{-\int_0^t \xi_s ds}\left(X_s^{(1)} - X_s^{(2)}\right)^T\left(\sigma(X_s^{(1)}) - \sigma(X_s^{(2)})\right) dW_s\right] = 0.$$

From hypothesis **(a)**, the following inequality holds

$$Tr\left[\left(\sigma(X_s^{(1)}) - \sigma(X_s^{(2)})\right)^T\left(\sigma(X_s^{(1)}) - \sigma(X_s^{(2)})\right)\right] \leq L_\sigma \left\|X_s^{(1)} - X_s^{(2)}\right\|^2 = L_\sigma Y_s,$$

from which we can increase (0.28), obtaining

$$\mathbb{E}\left[\mathrm{e}^{-\int_0^t \xi_s ds}\ Y_t\right] \leq \\ \leq \int_0^t \mathbb{E}\left[\mathrm{e}^{-\int_0^s \xi_r dr}\Big(Y_s(L_\sigma - \xi_s) + 2\langle\big(X_s^{(1)} - X_s^{(2)}\big), \big(\mu(X_s^{(1)}) - \mu(X_s^{(2)})\big)\rangle\Big)\right] ds. \tag{0.29}$$

Considering D_z, moreover

$$\langle\left(X_s^{(1)} - X_s^{(2)}\right), \left(\mu(X_s^{(1)}) - \mu(X_s^{(2)})\right)\rangle \leq L_\mu(z_s)\left\|X_s^{(1)} - X_s^{(2)}\right\|^2 = L_\mu(z_s) Y_s. \tag{0.30}$$

From (0.30) and using the process definition of ξ_s given in (0.22), Equation (0.29) becomes

$$\mathbb{E}\left[\mathrm{e}^{-\int_0^t \xi_s ds}\ Y_t\right] \leq \int_0^t \mathbb{E}\left[\mathrm{e}^{-\int_0^s \xi_r dr}\Big(Y_s\big(L_\sigma + 2L_\mu(z_s) - \xi_s\big)\Big)\right] ds = 0, \tag{0.31}$$

from which the thesis is demonstrated.

Now we have to prove that the solution exists, from the definition of D_z we find $D_z \subset D_{2z} \subset D$. It is possible to define the function $\varphi^{(z)} : \mathbb{R}^n \to \mathbb{R}$, with $\varphi^{(z)} \in C^\infty(\mathbb{R}^n)$ and $0 \leq \varphi^{(z)} \leq 1$ such that

$$\varphi^{(z)}(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in D_z \\ 0 & \mathbf{x} \notin D_{2z}, \end{cases} \tag{0.32}$$

Since $\mu^{(z)} = \varphi^{(z)}\mu$ and $\sigma^{(z)} = \varphi^{(z)}\sigma$, a new differential stochastic equation system is defined

$$\begin{aligned} &dX_t^{(z)} = \mu^{(z)}(X_t^{(z)})dt + \sigma^{(z)}(X_t^{(z)})dW_t, \quad 0 \le t \le T \\ &X_0 = x_0, \end{aligned} \tag{0.33}$$

where $X_t^{(z)}$ is the stochastic process that solve (0.33).
The coefficient $\mu^{(z)}$ is globally Lipschitz with constant

$$L_\mu(2z) + \sup_{\mathbb{R}^n} \|\nabla\varphi^{(z)}\| \sup_{D_{2z}} \|\mu\|.$$

Therefore Equation (0.33) satisfy hypothesis of Theorem 12.0.7. We have to prove now that $z \to \infty$, $X_t^{(z)} \to X_t$, is solution of system (0.16).
We define

$$\tau_z = \inf\{t > 0 : X_t^{(z)} \notin D_z\}, \tag{0.34}$$

that is the *exit time* from D_z di $X^{(z)}$.
Since the process $X^{(z)}$ is adapted to filtration $\{\mathcal{F}_t\}_{t\in[0,T]}$, τ_z is the *stopping time.*
Let $t \in [0,T]$, then $\forall\ t < \tau_z$, $X_t^{(z)} \in D_z \subset D$.
Considering the *generalized Lyapunov function*, and considering also (0.2), using the Ito's formula to $V\big(X_t^{(z)}\big)$:

$$dV\big(X_t^{(z)}\big) = L_z V\big(X_t^{(z)}\big)dt + \nabla^T V\big(X_t^{(z)}\big)\sigma^{(z)}\big(X_t^{(z)}\big)dW_t. \tag{0.35}$$

Calling $L_z V$ the operator

$$L_z V = \langle \nabla V(\mathbf{x}), \mu^{(z)}(\mathbf{x})\rangle + \frac{1}{2} Tr\big[\sigma^{(z)}(\mathbf{x})\sigma^{(z)T}\nabla^2 V(\mathbf{x})\big],$$

we find that $L_z = L$ on D_z.
Therefore for $0 \le s \le t \wedge \tau_z$ we have $X_s^{(z)} \in D_z \subset D$ and $V\big(X_{t\wedge\tau_z}^{(z)}\big)$ becomes

$$V\big(X_{t\wedge\tau_z}^{(z)}\big) = V(x_0) + \int_0^{t\wedge\tau_z} LV\big(X_s^{(z)}\big)ds + \int_0^{t\wedge\tau_z} \nabla^T V\big(X_s^{(z)}\big)\sigma\big(X_s^{(z)}\big)dW_s. \tag{0.36}$$

Considering the expected value of the above equation

$$\mathbb{E}\left[V\big(X_{t\wedge\tau_z}^{(z)}\big)\right] = V(x_0) + \mathbb{E}\left[\int_0^{t\wedge\tau_z} LV\big(X_s^{(z)}\big)ds\right], \tag{0.37}$$

since that the expected value of the stochastic integral is null

$$\mathbb{E}\left[\int_0^{t\wedge\tau_z} \nabla^T V\big(X_s^{(z)}\big)\sigma\big(X_s^{(z)}\big)dW_s\right] = 0. \tag{0.38}$$

Using property *(ii)* of the *generalized Lyapunov function* we find

$$\begin{aligned}\mathbb{E}\left[V\big(X^{(z)}_{t\wedge\tau_z}\big)\right]-V(x_0)&=\mathbb{E}\left[\int_0^{t\wedge\tau_z}LV\big(X^{(z)}_s\big)ds\right]\\&\leq\zeta\;\mathbb{E}\left[\int_0^{t\wedge\tau_z}\Big(1+V\big(X^{(z)}_{s\wedge\tau_z}\big)\Big)\,ds\right]\\&\leq\zeta T+\zeta\int_0^t\mathbb{E}\left[V\big(X^{(z)}_{s\wedge\tau_z}\big)\right]ds,\end{aligned}\tag{0.39}$$

that satisfies the hypothesis of (0.14). Therefore we can applies Gronwall Lemma to (0.39) finding

$$\mathbb{E}\left[V\big(X^{(z)}_{t\wedge\tau_z}\big)\right]\leq e^{\zeta T}\left(V(x_0)+\zeta T\right).\tag{0.40}$$

The above relation shows that $\mathbb{E}\left[V\big(X^{(z)}_{t\wedge\tau_z}\big)\right]$ is less than finite amount independent form z.
Hence we find the probability of $\tau_z\leq t$:

$$\begin{aligned}\mathbb{P}(\tau_z\leq t)&=\mathbb{E}\left[\chi_{\tau_z\leq t}\right]\\&=\mathbb{E}\left[\frac{V\big(X^{(z)}_{t\wedge\tau_z}\big)}{z}\chi_{\tau_z\leq t}\right]\\&\leq\frac{1}{z}\mathbb{E}\left[V\big(X^{(z)}_{t\wedge\tau_z}\big)\right]\leq\frac{e^{\zeta T}\left(V(x_0)+\zeta T\right)}{z},\end{aligned}\tag{0.41}$$

where we use (0.40) and the $V\big(X^{(z)}_{t\wedge\tau_z}\big)=V\big(X^{(z)}_{\tau_z}\big)=z$ for $\tau_z\leq t$
Therefore for $z\to\infty$, $\tau_z\to\infty$ in probability and exists a subsequence $z_n\to\infty$ strictly increasing such that $\tau_{z_n}\to\infty$ almost certainly [50]. The process $X^{(z)}_t$ is bounded in D.
We call

$$H^{(n)}_t=X^{(z_n)}_{t\wedge\tau_{z_n}},\quad\text{e}\quad\tau'_n=\tau_{z_n}.$$

Then, $\forall n>m$,

$$H^{(n)}_t=x_0+\int_0^{t\wedge\tau'_n}\mu\big(H^{(n)}_s(s)\big)ds+\int_0^{t\wedge\tau'_n}\sigma\big(H^{(n)}_s(s)\big)dW_s,\tag{0.42}$$

and

$$H^{(m)}_t=x_0+\int_0^{t\wedge\tau'_m}\mu\big(H^{(m)}_s(s)\big)ds+\int_0^{t\wedge\tau'_m}\sigma\big(H^{(m)}_s(s)\big)dW_s.\tag{0.43}$$

Moreover, in D_{z_m},

$$\mu^{(z_n)}=\mu^{(z_m)},\quad\text{e}\quad\sigma^{(z_n)}=\sigma^{(z_m)}.$$

For all $0 \leq t \leq \tau'_m$, $X_t^{(z_n)} = X_t^{(z_m)}$ *almost certainly*. Given that $\tau'_m \leq \tau'_n$, $H_t^{(n)} = H_t^{(m)}$, for all $0 \leq t \leq \tau'_m$, we have *almost certainly*

$$\begin{aligned} H_t^{(n)} &= x_0 + \int_0^{t\wedge\tau'_m} \mu\big(H_s^{(n)}(s)\big)ds + \int_{t\wedge\tau'_m}^{t\wedge\tau'_n} \mu\big(H_s^{(n)}(s)\big)ds + \\ &+ \int_0^{t\wedge\tau'_m} \sigma\big(H_s^{(n)}(s)\big)dW_s + \int_{t\wedge\tau'_m}^{t\wedge\tau'_n} \sigma\big(H_s^{(n)}(s)\big)dW_s. \end{aligned} \tag{0.44}$$

Evaluating $H_t^{(n)} - H_t^{(m)}$, from (0.44) and from (0.43), we find

$$H_t^{(n)} - H_t^{(m)} = \int_{t\wedge\tau'_m}^{t\wedge\tau'_n} \mu\big(H_s^{(n)}(s)\big)ds + \int_{t\wedge\tau'_m}^{t\wedge\tau'_n} \sigma\big(H_s^{(n)}(s)\big)dW_s. \tag{0.45}$$

Let $A = \{\omega \in \Omega : \tau'_n(\omega) \to \infty\}$, $\mathbb{P}(A) = 1$, from what previously said, $\mathbb{P}(A) = 1$. Let $\omega \in A$, e $t \in [0, T]$, then

$$\exists\ \bar{n} : \forall\ n > m > \bar{n},\ \text{segue che}\ \tau'_n \geq \tau'_m > t.$$

Whereby $\forall\ n > m > \bar{n}$,

$$H_t^{(n)}(\omega) = H_t^{(m)}(\omega). \tag{0.46}$$

Then exists $X_t(\omega) \in \mathbb{R}^n$ such that

$$\lim_{n\to\infty} H_t^{(n)}(\omega) = X_t(\omega), \tag{0.47}$$

it is easy to see that $X_t(\omega)$ is a stochastic process adapted to filtration $(\mathcal{F}_t)_{t\in[0,T]}$. Moreover $\forall\ \omega \in A$ and $t \in [0, T]$, for $n \to \infty$, we have $V(X_t(\omega)) \leq z_n$. So $X_t(\omega) \in D$, and according to property *(iii)* of Definition 12.0.4, follows (0.17).

We prove that $X_t(\omega)$ satisfies the differential stochastic equation (0.16). Recalling (0.42)

$$H_t^{(n)} = x_0 + \int_0^{t\wedge\tau'_n} \mu\big(H_s^{(n)}(s)\big)ds + \int_0^{t\wedge\tau'_n} \sigma\big(H_s^{(n)}(s)\big)dW_s.$$

Given $t \in [0, T]$ e $\omega \in A$, $H_t^{(n)}(\omega) = X_t(\omega)$ for all $n > \bar{n}(\omega)$ (so $\tau_n > t$) and $H_s^{(n)}(\omega) = X_s(\omega)$ for all $s < t$ e $n > \bar{n}$, we find that $X_t(\omega)$ verifies (0.16). □

The above result reveals the crucial role that an environmental noise plays in the dynamics. Even a small amount of stochastic noise allows to suppress the deterministic explosion. We also observe that our model is substantially different from the stochastic Lotka-Volterra model studied in Mao [263], [228] which has been derived by stochastically perturbing the parameter in the deterministic version.

Acknowledgements

If I have seen little further, it is by standing on the shoulders of my giant family

References

[1] "Global health expenditure database," 2016. vii, 7, 10, 14, 27, 28

[2] "Oecd website," 2016. vii, 11, 32, 35, 36

[3] "World health organization website," 2016. 6

[4] "The world bank database," 2016. 7

[5] E. Missoni and G. Solimano, "Towards universal health coverage: the chilean experience. world health report 2010 background paper, no.4," 2010. 9

[6] P. Gottret, G. J. Schieber, and H. R. Waters, *Good practices in health financing: Lessons from reforms in low- and middle-income countries.* The World Bank, 2008. 9

[7] WHO, *Health systems financing: the path to universal coverage.* The World Health Report, 2010. 9

[8] P. Prakongsai, W. Patcharanarumol, and V. Tangcharoensathien, "Can earmarking mobilize and sustain resources to the health sector?," *Bulletin of the World Health Organization*, 2008. 10

[9] N. Ravishankar, "Financing of global health: tracking development assistance for health from 1990 to 2007," *Lancet*, 2009. 10

[10] F. Gorunescu, S. I. McClean, and P. H. Millard, "A queueing model for bed-occupancy management and planning of hospital.," *Journal of the Operational Research Society 53: 19-24*, 2002. 11, 123, 128

[11] L. V. Green, "Capacity planning and management in hospital.," *Operational Research and Health Care 70: 15-44*, 2004. 11, 123

[12] P. R. Harper and A. Shahani, "Modelling for the planning and management of bed capacities in hospitals," *Journal of the Operational Research Society 53: 11-18.*, 2002. 11, 80, 128

[13] E. Hans, M. V. Houdenhoven, and P. Hulshof, "A framework for health care planning and control," 2011. 29, 31, 103

[14] S. Glouberman and H. Mintzberg, "Managing the care of health and the cure of disease - part i," *Health Care Manag Rev 26: 56-69*, 2001. 30

[15] S. Glouberman and H. Mintzberg, "Managing the care of health and the cure of disease - part ii," *Health Care Manag Rev 26: 70-84*, 2001. 30

[16] M. Carter, "Diagnosis: mismanagement of resources," *OR/MS today29(2): 26-32*, 2002. 30, 40

[17] K. Khoumbati, M. Themistocleous, and Z. Irani, "Evaluation the adoption of enterprise application integration in health-care organizations," *J Manag Inf Syst, 22(4): 69-108*, 2006. 30

[18] A. V. Roth and R. Dierdonck, "Hospital resource planning," *Prod Oper Manag 4:2-29*, 1995. 30, 31, 40

[19] M. E. Porter and E. O. Teisberg, "How physicians can change the future of health care," *J Am Med Assoc 297: 1103-1111*, 2007. 30, 31

[20] A. Maynard, "Developing the health care market," *Econ J 1001(408): 1277-1286*, 1991. 30

[21] A. Maynard, "Can competition enhance efficiency in health care?," *Soc Sci Med 39(10): 1433-1445*, 1994. 30

[22] D. Rhyne and D. Jupp, "Health care requirements planning: a conceptual framework," *Health Care Manag Rev 13(1): 17-27*, 1988. 31

[23] G. P. Westert, J. S. Burgers, and H. Verkleij, "The netherlands: regulated competition behind the dykes?," *Br Med j 339: 839-842*, 2009. 32

[24] R. E. Wachtel and F. Dexter, "Tactical increases in operating room block time for capacity planning should not be based on utilization," *Anesth Analg 106: 215-226*, 2008. 32

[25] R. N. Anthony, "Planning anc control system: a framework for analysis," *Harvard Business School Division of Research, Boston*, 1965. 32

[26] L. Xaltd, W. C. Benton, and G. K. Leong, "The impact of strategic operations management decisions on community hospital performance," *J Oper Manag 20: 389-408*, 2002. 32

[27] G. Johnson, K. Scholes, and R. Whittington, *Exploring corporate strategy. 8th edn.* Prentice Hall, New Jersey, 2008. 35

[28] P. H. Millard and S. I. McClean, "Modelling hospital resource use: a different approach to the planning and control of health care systems.," *Royal Society of Medicine*, 1994. 35, 80, 112, 124

[29] M. Papi and L. Pontecorvi, "Hospital resource consumption modelling," *Research in Business and Management 1(1).*, 2014. 36, 91

[30] C. Torio and R. Andrews, "National inpatient hospital costs: The most expensive conditions by payer, 2011.," *Agency for Healthcare Research and Quality*, 2013. 37

[31] A. K. Shahani, "Reasonable averages that give wrong answers," *Teaching Statistics 3, 50-54.*, 1991. 37

[32] P. R. Harper, "A framework for operational modelling of hospital resources," *Health Care Management Science 5, 165-173.*, 2002. 37, 103

[33] E. P. C. Kao and G. G. Tung, "Bed allocation in a public health care delivery service.," *Managment Science 27, 507-520*, 1991. 39

[34] X. Huang, "A planing model for requirement of emergency beds.," *IMA Journal of Math. Applied to Medical Biology 12: 345-353*, 1995. 39

[35] R. J. Ruth, "A mixed integer programming model for regional planning of a hospital inpatient service.," *Managment Science 27, 521-533*, 1981. 39

[36] W. T. Lin, "Modelling and forecasting hospital patient movements: Univariate and multiple time series approaches.," *IJF (The Netherlands), 195-208*, 1989. 39

[37] M. B. Dumas, "Simulation modelling for hospital bed planning.," *Simulation 43, 69-78*, 1984. 39

[38] G. Vassilacopoulos, "A simulation model for bed allocation to hospital inpatient departments.," *Simulation 45: 233-241*, 1985. 39

[39] D. Gove, D. Hewett, and A. K. Shahani, "Towards a model for hospital case-load decision support," *IMA Journal of Math. Applied to Medical Biology 12, 329-338*, 1995. 39

[40] I. H. W. et al., "Statistical modelling to predict elective surgery time: Comparison with a computer scheduling system and surgeon provided estimates.," *Anaesthesiology 85: 1235-1245*, 1996. 39

[41] T. G. et al., "Surgical unit time utilisation review: Resource utilisation and management implications.," *Journal of Medical Systems 12, 169-179*, 1998. 39

[42] H. Wolfe and J. P. Young, "Staffing the nursing unit, part i: Controlled variable staffing.," *Nursing Research 14: 236-243*, 1963. 39

[43] E. S. Resonbloom and N. F. Goertzen, "Cyclic nurse scheduling.," *European Journal of Operatinal Research 31: 19-23*, 1987. 39

[44] S. P. Silfred and W. C. Benton, "Workforce staffing and scheduling: Hospital nursing specific models," *European Journal of Operatinal Research 60: 223-246*, 1992. 39

[45] S. P. Silfred and W. C. Benton, "A decision model for shift scheduling of nurses.," *European Journal of Operatinal Research 74: 519-527*, 1994. 39

[46] K. A. Dowsland and J. M. Thompson, "Solving a nurse scheduling problem with knapsacks, networks and tabu search.," *Journal of Operatinal Research Society 51, 825-833*, 2000. 39

[47] J. M. H. Vissers, "Patient flow-based allocation of inpatient resources: a case of study.," *Eur J Oper Res 105: 356-370*, 1998. 40

[48] W. H. M. Zijm, "Towards intelligent manufacturing planning and control systems.," *OR Spectr 22: 313-345*, 2000. 40

[49] Y. A. Ozcan, *Quantitative methods in health care management - techniques and applications.* Jossey-Bass/Wiley, San Francisco, 2009. 40

[50] P. Baldi, *Equazioni differenziali stocastiche e applicazioni.* Pitagora Editrice, 2000. 44, 243, 250, 255

[51] R. Feynman, "Space-time approach to nonrelativistic quantum mechanics," *Rev. Mod. Phys.*, 1948. 46

[52] M. Kac, "On the distribution of certain wiener functionals," *Trans. Amer. Math.*, 1949. 46

[53] D. Duffie and R. Kan, "A yield factor model of interest rates," *Mathematical Finance*, 1996. 47

[54] Q. Dai and K. Singleton, "Term structure dynamics in theory and reality," *The Review of Financial Studies*, 2000. 47

[55] D. Duffie and D. Filipovic, *Affine processes and applications in finance.* Pitagora Editrice, 2009. 48

[56] P. Boyle, W. Tian, and F. Guan, "The riccati equation in mathematical finance," *J. Symbolic Computation*, 2000. 48

[57] F. Black and M. Scholes, "The pricing of options and corporate liabilities," *J.Political Economy: 637-654.*, 1973. 49, 152, 155

[58] R. C. Merton, "theory of rational option pricing," *Bell J Econ Manage Sci 4, 141:183.*, 1973. 49, 152

[59] R. S. Pindyck, "Irreversibility, uncertainty, and investment," *National Bureau of Economic Research Cambridge, Mass, USA.*, 1991. 50, 152

[60] D. R. Williams and P. H. Hammes, "Real option reasoning in healthcare: an integrative approach and synopsis," *J Healthcare Manage 52: 170.*, 2007. 50, 51, 152, 157

[61] S. C. Myers, "Determinats of corporate borrowing," *J Finac Econ 5: 147:175.*, 1977. 50, 152

[62] R. McDonald and D. Siegel, "The value of waiting to invest," *Q J Econ 101: 707-727.*, 1986. 50, 153

[63] A. K. Dixit and R. S. Pindyck, "investment under uncertainty," *Princeton University Press, Princeton*, 1994. 50, 153

[64] L. Trigeorgis, "Real option: Managerial flexibility and strategy in resource allocation," *MIT Press.*, 1996. 50

[65] K. J. Huisman and P. M. Kort, "Strategic technology adoption taking into accountfuture technological improvements: A real options approach," *Eur J operational Res 159: 705-728.*, 2004. 50, 153

[66] H. S. Herath and C. S. Park, "Multi-stage capital investment opportunities as compound real option," *The Eng Economist 47: 1-27*, 2002. 50, 153

[67] P. Kodukula and C. Papudesu, "Project valuation using real option: a practitioner's guide," *J. Ross Publishing*, 2006. 50

[68] E. S. Schwartz and L. Trigeorgis, "Real options and investment under uncertainty: An overview," *MIT Press, pp 1-16.*, 2001. 50

[69] R. C. Merton, "applications of option-pricing theory: twenty-five years later," *Am Econ Rev 88: 323-349.*, 1998. 51

[70] R. Levaggi and M. Moretto, "Investment in hospital care technology under different purchasing rules: a real option approach," *FEEM Working Paper No 7504*, 2004. 51

[71] S. Palmer and P. C. Smith, "Incorporating option valies into the economic evaluation of health care technologies," *J Health Econ 19: 755-766.*, 2000. 51

[72] B. Sengupta and R. E. Kreier, "A dynamic model fo health plan choice from a real options perspective," *Atlantic Econ J 39: 401-419.*, 2011. 51

[73] G. N. Angelou and A. A. Econimides, "A decision analysis framework for prioritizing a portfolio of ict infrastructure projects," *IEEE Trans Eng Manage 55: 479-495*, 2008. 51

[74] G. N. Angelou and A. A. Econimides, "A multi-creiteria game theory and real-option model for irreversible ict investment decisions," *Telecommun Policy 33: 686-705.*, 2009. 51

[75] D. Wyant, "Real option analysis improving project selection in healtcare setting," *J Healthcare Inf Manage 23: 56*, 2009. 51

[76] T. Driffied and P. C. Smith, "A real options approach to watchful waiting: theory and an illustration," *Med Decis Making 27: 178-188.*, 2007. 52

[77] R. Maurer, "Financial valuation of incentive arrangements in managed care contracts: a real options approach," *J Health Care 28: 39.*, 2001. 52

[78] F. T. Magiera and R. A. McLean, "Strategic options in capital budgeting and program selection under fee-for-service and managed care," *Health Care Manage Rev 21: 7-17.*, 1996. 52

[79] P. Bremaud, "Point processes and queues: Martingale dynamics," *Springer-Verlag*, 1981. 66

[80] W. Feller, "two singular diffusion problems," *Ann. Math. 2: 173-182.*, 1951. 68, 210

[81] J. Cox, J. Ingersoll, and S. Ross, "A theory of the term structure of interest rates," *Econometrica*, 1985. 68, 164

[82] J. Cox, J. Ingersoll, and S. Ross, "An intertemporal general equilibrium model of asset prices," *Econometrica*, 1985. 68

[83] M. Fackrell, "Modelling healthcare systems with phase-type distributions," *Health Care Management Science, 12:11*, 2009. 78, 82

[84] C. A. O'Cinneide, "Phase-type distributions: Open problems and a few properties," *School of Industrial Engeneering Purdue University*, 1997. 79

[85] Q.-M. He and H. Zhang, "A note on unicyclic representations of phase type distributions.," *Stochastic Models 21: 465-483*, 2005. 79, 86

[86] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, "Coherent measures of risk," *Mathematical Finance 9, 203-228.*, 1999. 79, 81, 89

[87] P. Cheridito, F. Delbaen, and M. Kupper, "Coherent and convex monetary risk measures for unbounded cádlág processes," *Finance Stochastic 9, 1713-1732.*, 2005. 79, 81, 89

[88] F. Delbaen, "Coherent risk measures on general probability spaces," *In: Advances in Finance and Stochastics, 1-37.*, 2002. 79, 81, 89

[89] A. X. Costa, S. A. Ridley, A. K. Shahani, P. R. Harper, V. Senna, and M. S. Nielsen, "Mathematical modelling and simulation for planning critical care capacity," *Anaesthesia 58, 320-327.*, 2003. 80

[90] C. Vasilakis and A. H. Marshall, "Modelling nationwide hospital length of stay: Opening the black box," *The Journal of the Operational Research Society 56: 862-869.*, 2005. 80, 85, 102

[91] M. Papi, L. Pontecorvi, and R. Setola, "Hospital resource consumption modelling," *Health Care Manag Sci 19(1): 58-69.*, 2016. 80

[92] M. F. Neuts, "phase-type probability distributions," *Encyclopedia of Operations Research and Management Science, 614-616*, 1975. 82

[93] S. M. Ross, *Introduction to Probability Models, Tenth Edition.* Elsevier, 2010. 82

[94] S. M. Ross, *Stochastic Process,2nd Edition.* John Wiley&Sons, 1996. 82

[95] S. Asmussen, O. Nerman, and M. Olsson, "Fitting phase-type distributions via the em algorithm," *Scandinavian Journal of Statistics 23: 419-441.*, 1996. 85, 95

[96] M. J. Faddy and S. I. McClean, "Analysing data on lengths of stay of hospital patients using phase-type distributions.," *Applied Stochastic Models in Business and Industry, 15: 311-317*, 1999. 85, 95

[97] A. Bobbio and A. Cumani, "Ml estimation of the parameters of a ph distribution in triangular canonical form.," *Computer performance evaluation 22, 33-46*, 1992. 85, 95

[98] K. T. Fang and Y. Wang, *Number-Theoretic Methods in Statistics.* Chapman&Hall, London, 1994. 85, 95, 155

[99] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods.* CBMS-NSF Regional Conference Series in Applied Mathematics, 1992. 85, 95, 155, 169

[100] S. I. McClean and P. H. Millard, "Patterns of length of stay after admission in geriatric medicine: an event history approach.," *Statistician 42: 263-274*, 1993. 85, 102

[101] C. A. O'Cinneide, "Patterns of length of stay after admission in geriatric medicine: an event history approach.," *Communications in Statistics. Stochastic Models 15, 731-757.*, 1999. 86, 102

[102] J. W. Jang, "Loan portfolio value," *Risk.*, 2002. 93

[103] E. Palombini, "Factor models and the credit risk of a loan portfolio," *MPRA Paper No. 20107.*, 2009. 93

[104] R. S. Valdez, E. Ramly, and P. F. Brennan, "Industrial and systems engineering and health care: Critical areas of research. final report," *AHRQ Publication No. 10-0079. Rockville, MD: Agency for Healthcare Research and Quality*, 2010. 103

[105] F. Clemente, M. D'Arco, and E. D'Avino, "The use of a conceptual model and related indicators to evaluate quality of healthcare in intensive care units," *Quality Engineering 26*, 2014. 104, 105, 106, 119

[106] HSRG, "Quality of care: what is quality and how can it be measured?," *CMAJ 146: 2153-2158*, 1992. 104, 106, 119

[107] F. Clemente, M. Papi, L. Pontecorvi, and A. Menichetti, "Evaluation of indices for the measurement of quality in health systems," *Int. J. Metrol. Qual. Eng. 7(4).*, 2016. 104

[108] P. Winkel and N. F. Zhang, *Statistical Development of Quality in Medice, p 1-7.* Wiley, 2007. 104, 106, 119

[109] B. Kemper and J. de Mast, "Measurement plans for process flow improvement in services and health care," *Quality Engineering 25: 437-450*, 2013. 104, 119

[110] M. Hensher and B. Keogh, *Quality metrics, p 393-396.* Elsevier, 2009. 104, 106

[111] A. Donabedian, "The quality of care: how can it be assessed?," *JAMA 260: 1743-1748*, 1988. 105

[112] A. Handler, M. Issel, and B. Turnock, "A conceptual framework to measure performance of the public health system," *Am J Public Health 91: 1235-1239*, 2001. 105, 106

[113] C. Herman, W. Karol, A. M. Yip, K. J. Buth, A. Hassan, and J. F. Legare, "Predicting prolonged intensive care unit length of stay in patients undergoing coronary artery bypass surgery development of an entirely preoperative scorecard," *Interactive CardioVasc Thoracic Surgery 9: 654-658*, 2009. 105

[114] L. I. Worthley, "Quality control, audit, adverse events and risk in the intensive care unit," *Critic Care Resusc 2: 304-307*, 2000. 105

[115] D. C. Montgomery, *Statistical Quality Control: A Modern Introduction, 6th Edition.* Wiley, 2009. 105, 107, 120

[116] C. M. Lilly, S. Cody, H. Zhao, K. Landry, S. P. Baker, J. McIlwaine, M. W. Chandler, and R. S. Irwin, "Hospital mortality, length of stay, and preventable complications among critically ill patients before and after tele-icu reengineering of critical care processes.," *JAMA 305: 1777-1785*, 2011. 105

[117] EuropeanCommission, "specific programme implementing horizon 2020 - the framework programme for research and innovation (2014-2020)," *COUNCIL DECISION, Brussels*, 2011. 105

[118] E. A. McGlyn, "identifying, categorizing, and evaluating health care efficiency measures," *AHRQ 8*, 2008. 105

[119] M. DeVos, W. Graafmans, E. L. Keesman, G. Westert, and P. VanDerVoort, "quality measurement at intensive care units: which indicators should we use?," *Journal of Critical Care 22: 267-274*, 2007. 106

[120] S. M. Berenholtz, T. Dorman, N. Koni, and P. J. Pronovost, "Qualitative review of intensive care unit quality indicators.," *Journal of Clinical Care 17: 1-12*, 2002. 106

[121] L. Straney, A. Clements, and J. Alexander, "Quantifying variation of paediatric lenght of stay among intensive care units in australia and new zeland.," *Quality & Safety in Healt Care 19: 1-5*, 2010. 112

[122] F. Roques, P. Michel, A. R. Goldstone, and S. A. Nashef, "The logisitc euroscore.," *European Heart Journal 24: 1-2*, 2003. 116

[123] D. Bertsekas, *Dynamic Programming and Optimal Control, Volume 2 4th Edition.* Athena Scientific, 2011. 122, 130

[124] W. L. Herring, "Prioritizing patients: Stochastic dynamic programming for surgery scheduling and mass casualty incident triage.," *DRUM*, 2011. 123

[125] J. Castaing, A. Cohn, B. Denton, and A. Weizer, "A stochastic programming approach to reduce patient wait times and overtime in an outpatient infusion center.," *IIE Transactions on Healthcare Systems Engineering, 6(3): 111-125*, 2015. 123

[126] A. Ozen, "Stochastic models for capacity planning in healthcare delivery: Case studies in an outpatient, inpatient and surgical setting," *Doctoral Dissertations 2014-current. Paper 125*, 2000. 123

[127] A. Erdelyi and H. Topaloglu, "Approximate dynamic programming for dynamic capacity allocation with multiple priority levels," *IIE Transactions 43(2): 129-142*, 2010. 123

[128] P. Punnakitikashem, J. M. Rosenberger, and D. B. Behan, "Stochastic programming for nurse assignment," *Computational Optimization and Applications, 40(3): 321-349.*, 2008. 123

[129] D. B. Gilleskie, "A dynamic stochastic model of medical care use and work absence," *Econometrica, 66: 1-45.*, 1998. 123

[130] M. Dong and W. Zhao, "Admissions optimisation and premature discharge decisions in intensive care units," *International Journal of Production Research, 53: 7329-7342.*, 2015. 123, 124

[131] G. Xiao, W. van Jaarsveld, M. Dong, and J. van de Klundertc, "Stochastic programming analysis and solutions to schedule overcrowded operating rooms in china," *Computers & Operations Research, 74, : 78-91.*, 2016. 123

[132] F. Dexter and R. D. Traub, "how to schedule elective surgical cases into specific operating rooms to maximize the efficiency of use of operating room time," *Anesthesia and Analgesia, 94: 933-942.*, 2002. 123

[133] I. Ozkarahan, "Allocation of surgeries to operating rooms by goal programming," *Journal of Medical Systems, 24(6): 339-378.*, 2000. 123

[134] F. Dexter, A. Macario, R. D. Traub, M. Hopwood, and D. A. Lubarsky, "an operating room scheduling strategy to maximize the use of operating room block time: Computer simulation of patient scheduling and survey of patients preferences for surgical waiting time," *Anesthesia and Analgesia, 89: 7-20.*, 1999. 123

[135] F. Dexter, J. T. Blake, D. H. Penning, B. Sloan, P. Chung, and D. A. Lubarsky, "use of linear programming to estimate impact of changes in a hospitals operating room time allocation on perioperative variable costs," *Anesthesiology, 96(3): 718-724.*, 2002. 123

[136] J. Belien and E. Demeulemeester, "Building cyclic master surgery schedules with leveled resulting bed occupancy," *European Journal of Operational Research, 176(2): 1185-1204.*, 2007. 123

[137] D. Sier, P. Tobin, and C. McGurk, "Scheduling surgical procedures," *Journal of the Operational Research Society, 48: 884-891.*, 1997. 123

[138] J. Seidel, P. C. Whiting, and D. L. Edbrooke, "The costs of intensive care," *Continuing Education in Anaesthesia, Critical Care & Pain, 6(4): 160-163.*, 2006. 123

[139] M. C. Cavallo, C. Lazzaro, M. Tabacchi, M. Langer, I. Salvo, G. Serra, and C. Taddei, "Il costo del reparto di terapia intensiva in italia.risultati da unindagine empirica su un campione di 12 centri," *Minerva Anestesiologica 67(1-2): 41-53.*, 2001. 123

[140] S. A. M. Nashef, F. Roques, P. Michel, E. Gauducheau, S. Lemeshow, and R. Salomon, "European system for cardiac operative risk evaluation (euroscore).," *Eur J Cardiothorac Surg 16: 512-521*, 1999. 125, 149

[141] T. S. Kurki, U. Hakkinen, J. Lauharanta, J. Ramo, and M. Leijala, "Evaluation of the realtionship between preoperative risk scores, postoperative and total length of stay and hospital costs in coronary bypass surgery," *European Journal of Cardio-thoracic Surgery 20: 1183-1187.*, 2001. 125, 146, 149

[142] T. S. Kurki, U. Hakkinen, J. Lauharanta, J. Ramo, and M. Leijala, "Euroscore predicts health-related quality of life after coronary artery bypass grafting," *Interact Cardiovasc Thorac Surg. 7: 564-568.*, 2008. 126

[143] J. O. McClain, "A model for regional obstetric bed planning," *Heath services research 13(4): 378-394.*, 1978. 128

[144] F. Dexter and A. Macario, "optimal number of beds and occupancy to minimize staffing costs in an obstetrical unit?," *Canadian Journal of Anesthesia 48(3): 295-301.*, 2001. 128

[145] R. A. Harris, "Hospital bed requirements planning," *European Journal of Operational Research 25(1): 121-126.*, 1986. 128

[146] D. K. K. Lee and S. A. Zenios, "Optimal capacity overbooking for the regular treatment of chronic conditions," *Operations Research, 57(4): 852-865*, 2009. 128

[147] G. W. Harrison, "Implications of mixed exponential occupancy distributions and patient flow models for health care planning," *Health Care Management Science, 4(1): 37-45.*, 2001. 128

[148] C. Hober, J. Manry, and L. Connelly, "Simulation development," *Clinical Simulation in Nursing, 5(5): 173-179.*, 2009. 128

[149] G. E. Loyd, C. L. Lake, and R. B. Greenberg, *Practical health care simulations.* Elsevier Mosby, c2004. 1st edition, 2004. 128

[150] J. D. Thompson, R. B. Fetter, C. S. McIntosh, and R. J. Pelletier, "Use of computer simulation techniques in predicting requirements for maternity facilities," *Hospitals, 37-45.*, 1963. 128

[151] D. Sinreich and Y. Marmor., "Emergency department operations: The basis for developing a simulation tool," *IIE Transactions, 37(3): 233-245.*, 2005. 128, 129

[152] F. Ataollahi, M. A. Bahrami, M. abesi, and F. Mobasheri, "A goal programming model for reallocation of hospitals' inpatient beds," *Middle-East Journal of Scientific Research 18 (11): 1537-1543.*, 2013. 129

[153] N. Ravishankar, "A note on the structure of the goal programming model: assessment and evaluation," *International Journal of Operations & Production Management, 16(1): 40-49.*, 1996. 129

[154] L. M. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* New York: Wiley Interscience, 1994. 130

[155] S. Sharifi and K. Saberi, "Capacity planning in hospital management: An overview.," *Indian Journal of Fundamental and Applied Life Sciences 4: 512-521*, 2014. 130

[156] A. H. VanStraten, E. M. Tan, M. A. Hamas, E. J. Martens, and A. A. J. VanZundert, "Evaluation of the euroscore risk scoring model for patients undergoing coronary artery bypass graft surgery: a word of caution.," *Neth Heart J 18: 355359*, 2010. 131

[157] P. Robinson, "The weekly pulse: An analysis of hospital activities by day of the week.," *Insight report. CHKS.*, 2012. 135

[158] P. Pintor, M. Bobbio, S. Colangelo, F. Veglia, R. Marras, and M. Diena, "Can euroscore predict direct costs of cardiac surgery?," *European Journal of Cardio-thorac Surgery 23: 595-598*, 2003. 146, 149

[159] F. J. et. al., "Daily cost of an intensive care unit day: The contribution of mechanical ventilation.," *Crit Care Med Vol. 33, No. 6*, 2005. 146

[160] C. R. Nelson and A. F. Siegel, "parsimonious modelling of yield curves," *The Journal of Business 60(4): 473-489*, 1987. 148

[161] L. E. Svensson, "Estimating and interpreting forward interest rates: Sweden 1992-1994.," *Washington, D.C., National Bureau of Economic Research Working Paper no. 4871, c1994*, 1994. 148

[162] F. Roques, S. A. Nashef, P. Michel, E. Gauducheau, S. Lemeshow, and R. Salamon, "European system for cardiac operative risk evaluation (euroscore).," *European Journal of Cardio-thoracic Surgery 16: 9-13*, 1999. 149

[163] G. Favato, G. Baio, A. Capone, A. Marcelussi, and F. S. Mennini, "A novel method to value real options in health care: The case of a multicohort human papillomavirus vaccination strategy. clinical therapeutics," *Clinical Therapeutics 35(7): 904-914.*, 2013. 157

[164] D. Duffie, *Dynamic Asset Pricing Theory*. Third English Edition, Princeton University Press, 1998. 160, 163

[165] I. Karatzas and E. Shreve, *Methods of Mathematical Finance*. Springer-Verlag, New-York, 1998. 160, 161, 162

[166] J. E. Li and Y. L. Zhu, "Multi-factor financial derivatives on finite domains," *Communications in Math. Sci. 1: 343359*, 2003. 161, 163

[167] J. M. Harrison and D. M. Kreps, "Martingales and arbitrage in multiperiod securities markets," *Journal of Economic Theory 20: 381408.*, 1979. 163

[168] D. Duffie, D. Filipovic, and W. Schachermayer, "Affine processes and applications in finance," *Annals of Applied Probability 13: 9841053.*, 2003. 165

[169] J. I. Marcum, "A statistical theory of target detection by pulsed radar: Mathematical appendix," *RAND Corporation, Santa Monica, CA, Research Memorandum RM-753*, 1947. 165

[170] A. H. Nuttall, "some integrals involving the q.function," *NUSC Technical Report 4297*, 1972. 165

[171] M. V. Kapinas, S. K. Mihos, and G. K. Karagiannis, "On the monotonicity of the generalized marcum and nuttall q-functions," *IEEE Transaction On Information Theory 55: 3701-3710*, 2009. 165

[172] M. Papi, L. Pontecorvi, and R. Setola, "Tax evasion dynamics by fiscal inspection activity," *Submitted to European Journal of Applied Mathematics.*, 2016. 173

[173] M. T. Xiong, J. X. Song, and Q. Z. Wu, "Nonlinear filter algorithm for opto-electronic target tracking," *SPIE 7283 4th International Symposium on Advanced Optical Manufacturing and Testing Technologies: Optical Test and Measurement Technology and Equipment*, 2009. 173

[174] S. Thrun, W. Burgard, and D. Fox, "Probabilistic robotics," *MIT Press*, 2005. 173

[175] Y. K. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with Application to Tracking and Navigation: Theory, Algorithm, and Software*. Wiley, 2001. 173, 180

[176] S. S. Hwang and J. L. Speyer, "Relative gps carrier-phase positioning using particle filters with position samples," *American Control Conference (IEEE ACC09): 4171-4177*, 2009. 173

[177] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering 81: 35-45*, 2009. 173, 178, 190

[178] Y. Zhang and J. Wuxing, "Autonomous navigation of lunar satellite based on the sem orientation information via ekf," *IEEE Chinese Control Conference (IEEE CCC06)*, 2006. 173

[179] S. H. Chung and R. A. Kennedy, "Forward-backward non-linear filtering technique for extracting small biological signals from noise," *Journal of Neuroscience Methods 40: 71-86*, 1991. 173

[180] T. H. Farquhar, G. Chinn, S. C. H. Ciro Kiko Hoh, and E. J. Hoffman, "A nonlinear, image domain filtering method for cardiac pet images," *IEEE Trans. on Nuclear Science 45: 2073-2079*, 1998. 173

[181] C. Wells, *The Kalman Filter in Finance*. Kluwer Academic Publishers, 1996. 173

[182] I. Hoteit, D. T. Pham, G. Triantafyllou, and G. Korres, "A new approximate solution of the optimal non-linear filter for data assimilation in meteorology and oceanography," *AMS Montly Weather Review 136: 317-334*, 2008. 174

[183] F. Carravetta, A. Germani, and M. Raimondi, "Polynomial filtering for linear discrete time non-gaussian systems," *SIAM Journal of Control and Optimization 34: 1666-1690.*, 1996. 178, 180

[184] A. V. Balakrishnan, *Applied Functional Analysis*. Springer-Verlag: New York, 1976. 179

[185] S. J. Julier and J. K. Uhlmann, "A new extension to the kalman filter to nonlinear systems," *AeroSense: The 11th International Symposium on Aerospace/Defence Sensing, Simulation and Controls*, 1997. 180

[186] X. Tian and Y. Bar-Shalom, "Coordinate conversion and tracking for very long range radars," *IEEE Trans. on Aerospace and Electronics Systems 45: 1073-1083.*, 2009. 180

[187] K. Nam and M. J. Tahik, "A second-order stochastic filter involving coordinate transformation," *IEEE Trans. on Automatic Control 44: 603-608*, 1999. 180

[188] A. Germani, C. Manes, and P. Palumbo, "Polynomial extended kalman filter," *IEEE Trans. on Automatic Control 50: 2059- 2064.*, 2005. 180

[189] F. Carravetta, A. Germani, and M. Raimondi, "Polynomial filtering of discrete-time stochastic linear systems with multiplicative state noise," *IEEE Trans. on Automatic Control 42: 1106-1126.*, 1997. 180

[190] T. Moir and M. Grimble, "Optimal self-tuning filtering, prediction, and smoothing for discrete multivariable processes," *IEEE Trans. on Automatic Control 29: 128-137.*, 1984. 180

[191] F. E. Daum, "Nonlinear filters: Beyond the kalman filter," *IEEE Aerospace and Electronics Systems Magazine 20:57-69.*, 2005. 180

[192] F. Schneider, "The increase of the size of the shadow economy of 18 oecd countries: somepreliminary explanations," *IFO Working Paper 306*, 2000. 181, 183

[193] F. Schneider, "The value added of underground activities: Size and measurement of the shadoweconomies and shadow economy labor force all over the world," *World Bank Working Paper. Washington, D.C.*, 2000. 182

[194] L. Bernardi and L. A. Franzoni, "Evasione fiscale e nuove tipologie di accertamento. introduzione all'analisi economica," *Rivista di Diritto Finanziario e Scienza delle Finanze.*, 2003. 182

[195] J. Andreoni, B. Erard, and J. Feinstein, "Tax compliance," *Journal of economic literature XXVI 818-860.*, 1998. 183, 186

[196] D. J. H. et all, "Does deterrence deter? measuring the effect of deterrence in field studies and experimental studies. why people pay taxes: tax compliance and enforcement," *University of Michigan Press.*, 1998. 183

[197] W. W. Pommerehne, A. Hart, and B. S. Frey, "Tax morale, tax evasion and the choice of policy instruments in different political systems," *Public finance (Supplement) 49: 52-69.*, 1994. 183, 185

[198] K. C. dor Government Foundation, "Survey of canadian attitudes toward taxation," *Toronto, KPMG.*, 1994. 183

[199] A. D. Juan, M. Lasheras, and R. Mayo, "Voluntary tax compliance behaviour of spanish income tax payers," *Public finance 49: 90-105.*, 1994. 183

[200] M. Bovi, "The nature of underground economy, some evidences from oecd countries," *Documento di lavoro, 26/02, Roma, Isae.*, 2002. 183

[201] V. Tanzi, *Il sistema tributario italiano: una prospettiva internazionale. La finanza pubblica italiana dopo la svolta del 1992.* Il Mulino Edizioni, 1996. 183

[202] R. Zizza, "Metodologie di stima dell'economia sommersa: un'applicazione al caso italiano," *Temi di discussione del Servizio Studi 463. Banca d'Italia.*, 2002. 183

[203] M. Calzaroni, "L'occupazione come strumento per la stima esaustiva del pil e la misuradel sommerso," *Contributo al seminario dell'ISTAT, Roma, 12-13.*, 2000. 183

[204] M. Bernasconi and A. Marenzi, "Gli effetti redistributivi dell'evasione fiscale in italia. ricerche quantitative per la politica economica," *Convegno CIDE-SADIBA, Banca d'Italia, 1-38.*, 1997. 183, 184

[205] A. Marenzi, "Prime analisi sulla distribuzione dell'evasione dell'irpef per categorie di contribuenti e per livelli di reddito," *Rapporto CNEL, il Mulino, 305-341*, 1996. 183

[206] L. Cannari, V. Ceriani, and G. D'Alessio, "Il recupero degli imponibili sottratti a tassazione," *Ricerche quantitative per la politica economica, sadiba, Banca d'Italia, 3-5 nov.*, 1995. 183, 184

[207] M. G. Allingham and A. Sandmo, "Income tax evasion: a theoretical analysis," *Journal of Public Economics I.*, 1972. 184

[208] G. S. Becker, "Crime and punishment: an economic approach.," *Journal of political economy*, 1968. 184

[209] K. J. Arrow, "Essays in the theory of risk-bearing," *Amsterdam, North Holland.*, 1970. 184

[210] J. Alm, W. H. McClelland, and W. D. Schulze, "Why do people pay taxes?," *Journal of Public Economics 48: 2138.*, 1992. 185, 186

[211] M. J. Graetz and L. L. Wilde, "The economics of tax compliance: Fact and fantasy," *National Tax Journal 38: 355-363.*, 1985. 185

[212] J. Skinner and J. Slemrod, "An economic perspective on tax evasion," *National tax journal 38: 345-353.*, 1994. 185

[213] W. W. Pommerehne and B. S. Frey, "The effects of tax administration on tax morale," *Discussion paper series II, 191, University of Konstanz.*, 1992. 185

[214] B. S. Frey and L. P. Feld, "Deterrence and morale in taxation: An empirical analysis," *CESifo WOrking Paper. 760, CESifo.*, 2002. 185

[215] C. T. Clotfelder, "Tax evasion and tax rates: an analysis of individual returns," *The review of economics and statistics 65: 363-373.*, 1983. 185

[216] S. J. Feinstein, "An econometric analysis of income tax evasion and its detection," *Rand journal of Economics 22: 397-417.*, 1991. 185

[217] J. C. Baldry, "Income tax evasion and the tax schedule: some experimental results," *Public Finance 42: 357-83.*, 1987. 185

[218] M. Bordignon, "A fairness approach to income tax evasion," *Journal of public economics 52: 345-362.*, 1993. 185

[219] B. Erard and J. S. Feinstein, "The role of moral sentiments and audit perceptions in tax compliance," *Public finance (supplement) 49: 70-89.*, 1994. 185

[220] L. Mittone, "Subjective versus objective probability: results from seven experiments on fiscal evasion," *WP 1997-04, Universit di Trento.*, 1997. 186

[221] F. Flatters and W. B. MacLeod, "Administrative corruption and taxation," *International tax and public finance 2: 397-417.*, 1995. 186

[222] E. M. Engel and J. R. Hines, "understanding tax evasion dynamics," *NBER working paper series.*, 1999. 186

[223] S. Klepper and D. Nagin, "Tax compliance and perceptions of risks of detection and criminal prosecution," *Kaw and society review 23: 209-240.*, 1989. 186

[224] C. Alexander and J. Julio Sergio Feinstein, "A microeconometric analysis of income tax evasion and its detection," *Cambridge, Mass., mimeo of Massachussets Institute of Technology.*, 1997. 186

[225] D. H. Boucher, *The Biology of Mutualism.* Groom Helm, London, 1985. 188, 189

[226] X. He and K. Gopalsamy, "Persistence, attractivity, and delay in facultative mutualism.," *J. Math. Anal. Appl. 215: 154-173*, 2005. 189

[227] C. L. Wolin and L. R. Lawlor, "Models of facultative mutualism: density effects.," *Amer. Natural. 124: 843-862*, 1984. 189

[228] X. Mao, S. Sabanis, and E. Renshaw, "Asymptotic behaviour of the stochastic lotkavolterra model," *J. Math. Anal. Appl. 287: 141-156*, 2003. 189, 256

[229] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*. Springer-Verlag, Berlin, 1991. 189, 191, 205

[230] E. Wang and M. Z. Lawlor, "Likelihood ratios and transformation of probability associated with two-parameter wiener processes.," *Probability Theory. 40: 283-308.*, 1975. 189

[231] R. S. Lipster and A. N. Shiryayev, *Statistics of Random Processes I*. Springer-Verlag, Berlin-Heidelberg, New York, 1977. 190

[232] K. Reif, S. Gunther, E. Yaz, and R. Unbehauen, "Stochastic stability of the discrete-time extended kalman filter," *IEEE Trans.Automatic Control.*, 1999. 190

[233] M. Papi, L. Pontecorvi, and C. Donatucci, "Weighted average price in the heston stochastic volatility model," *Submitted to Economics and Finance.*, 2016. 203

[234] J. Hull and A. White, "The pricing of options on assets with stochastic volatilities," *J. Financ. 42(2): 281300*, 1987. 203

[235] E. M. Stein and J. C. Stein, "Stock price distributions with stochastic volatility: an analytic approach," *Rev. Financ. Stud. 4: 23-46.*, 1991. 203

[236] S. L. Heston, "A closed-form solution for options with stochastic volatility with applications to bond and currency options," *Rev. Financ. Stud. 6: 327-343.*, 1993. 203, 204, 219, 227

[237] M. Jex, R. Henderson, and D. Wang, "Pricing exotics under the smile," *Risk Magazine 12(11): 7275*, 1999. 203

[238] A. Lipton, "The vol smile problem," *Risk Magazine 15: 6165*, 1989. 203

[239] Y. Ren, D. Madan, and M. Q. Qian, "Calibrating and pricing with embedded local volatility models," *Risk Magazine, 138143*, 2007. 203

[240] P. H. Labordere, "Calibration of local stochastic volatility models to market smiles: A monte-carlo approach," *Risk Magazine, 112117.*, 2009. 203

[241] A. Sepp, *tochastic local volatility models: Theory and implementation*. University of Leicester, 2010. 203

[242] R. Frey, "Derivative asset analysis in models with level-dependent and stochastic volatility," *Mathematics of finance, Part II. CWI Quarterly 10: 1-34*, 1997. 203

[243] J. P. Fouque, a. R. S. Giulio Papanicolaou, and K. Solna, "Multiscale stochastic volatility for equity, interest rate, and credit derivatives," *Cambridge University Press.*, 2011. 203

[244] A. A. Dragulescu and V. M. Yakovenko, "Probability distribution of returns in the heston model with stochastic volatility," *Quant. Financ. 2: 443-453.*, 2002. 204, 210, 213, 227

[245] C. Costantini, M. Papi, and F. D'Ippoliti, "Singular risk-neutral valuation equations.," *Financ. Stoch 16: 249-274.*, 2012. 206, 214

[246] M. G. Crandall, H. Ishii, and P. L. Lions, "User's guide to viscosity solutions of second order partial differential equations," *Am. Math. Soc.. Bulletin. New Series 27: 1-67.*, 1992. 206

[247] E. Ekstrom and J. Tysk, "the black-scholes equation in stochastic volatility models," *J. Math. Anal. Appl. 368: 498-507.*, 2010. 206

[248] S. M. Ouls-Aly, "Monotonicity of prices in heston model," *HAL Id: hal-00678437.*, 2012. 206

[249] A. L. Lewis, "A simple option formula for general jump-diffusion and other exponential lvy processes," *Envision Financial Systems and OptionCity.net.*, 2001. 207

[250] P. Carr and D. B. Madan, "Option valuation using the fast fourier transform," *J. Comput. Financ. 2: 6173.*, 1999. 207

[251] S. Mikhailov and U. Nogel, "Heston's stochastic volatility model implementation, calibration and some extensions," *Wilmott Magazine 74-79.*, 2003. 208

[252] J. Zhu, *Applications of Fourier Transform to Smile Modelling: Theory and Implementation*. Springer Finance, Berlin-Heidelberg, 2010. 208, 209

[253] C. Chiarella, M. Craddock, and N. El-Hassan, "The calibration of stock option pricing models using inverse problem methodology.," *Research Paper Series 39, Quantitative Finance Research Centre, University of Technology, Sydney.*, 2000. 209

[254] R. Cont, "Recovering volatility from option prices by evolutionary optimization," *J. Comput. Financ., Vol. 8: 4*, 2005. 209

[255] G. Bakshi, C. Cao, and Z. Chen, "Empirical performance of alternative option pricing models," *J. Financ. 52: 2003-2049*, 1997. 209

[256] J. Gatheral, "Lecture 1: Stochastic volatility and local volatility," *Case Studies in Financial Modeling Notes, Courant Institute of Mathematical Sciences.*, 2004. 209

[257] M. Broadie and O. Kaya, "Exact simulation of stochastic volatility and other affine jump diffusion processes," *Oper. Res. 54: 217-231.*, 2006. 214

[258] C. Kahl and P. Jackel, *Not-so-complex logarithms in the Heston model*. Wilmott, 2005. 217

[259] C. Kahl and R. Lord, "Complex logarithms in the heston model," *Math. Financ. 20: 671-694.*, 2010. 217, 221

[260] B. D. Finetti, *Theory of Probability*. John Wiley & Sons, 1994. 240

[261] D. Draper, J. Hodges, C. Mallows, and D. Pregibon, "Exchangeability and data analysis," *Journal of the Royal Statistical Society*, 1993. 240

[262] B. Oksendal, *Stochastic Differential Equations.* Springer-Verlag, 1998. 244

[263] X. Mao, G. Marion, and E. Renshaw, "Environmental brownian noise suppresses explosions in populations dynamics," *Stochastic Process. Appl. 97: 95-110*, 2002. 256