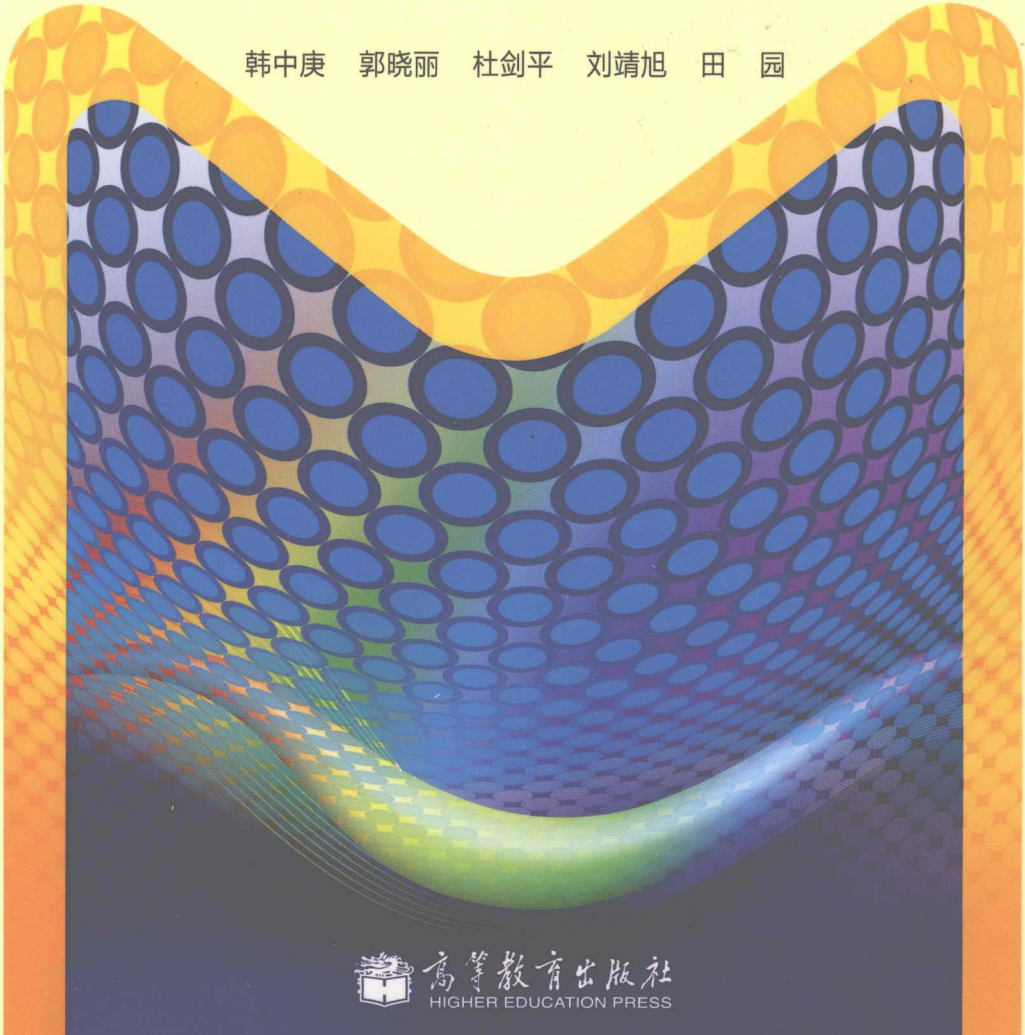
美国MCM/ICM竞赛指导丛书

美国大学生数学建模竞赛题解析与研究

第2辑

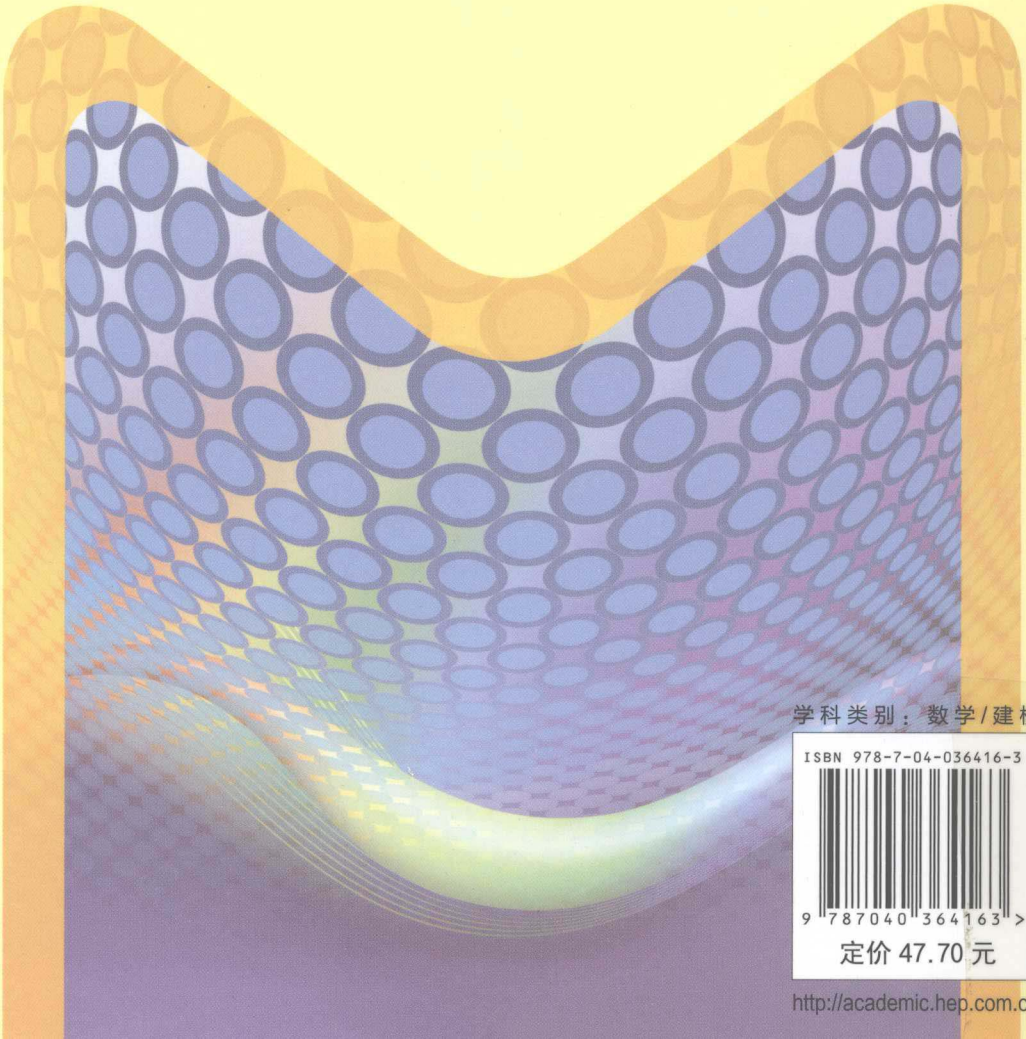
韩中庚　郭晓丽　杜剑平　刘靖旭　田　园



高等教育出版社
HIGHER EDUCATION PRESS

本辑题目

洪水估计问题
高速公路收费亭设置问题
不可再生资源的管理问题
灌溉喷洒系统设置问题
机场轮椅配置问题
艾滋病毒防控资源配置问题



学科类别：数学/建模
ISBN 978-7-04-036416-3
9 787040 364163 >
定价 47.70 元
http://academic.hep.com.cn

美国大学生数学建模竞赛题解析与研究

Meiguo Daxuesheng Shuxue Jianmo Jingsaiti Jiexi yu Yanjiu

第2辑

韩中庚　郭晓丽　杜剑平　刘靖旭　田　园

高等教育出版社·北京
HIGHER EDUCATION PRESS　BEIJING

内容提要

本系列丛书是以美国大学生数学建模竞赛（MCM/ICM）赛题为主要研究对象，结合竞赛特等奖的优秀论文，对相关的问题做深刻细致的解析与研究。本辑针对2005年及2006年MCM/ICM竞赛的6个题目：洪水估计问题、高速公路收费亭设置问题、不可再生资源的管理问题、灌溉喷洒系统设置问题、机场轮椅配置问题以及艾滋病毒防控资源分配问题等进行了解析与研究。

本书内容新颖、实用性强，目前国内尚无同类作品。本书可作为指导学生参加美国大学生数学建模竞赛的主讲教材，也可作为本科生、研究生学习和准备全国大学生、研究生数学建模竞赛的参考书，同时也可供研究相关问题的教师和研究生参考使用。

图书在版编目(CIP)数据

美国大学生数学建模竞赛题解析与研究．第2辑／韩中庚等编著．—北京：高等教育出版社，2013.1

（美国MCM/ICM竞赛指导丛书）

ISBN 978-7-04-036416-3

Ⅰ．①美… Ⅱ．①韩… Ⅲ．①数学模型-竞赛题-研究 Ⅳ．①O141.4-44

中国版本图书馆CIP数据核字（2012）第261616号

策划编辑	刘 英	责任编辑	刘 英	封面设计	李卫青	版式设计	范晓红
责任校对	金 辉	责任印制	朱学忠				

出版发行	高等教育出版社	咨询电话	400-810-0598
社 址	北京市西城区德外大街4号	网 址	http://www.hep.edu.cn
邮政编码	100120		http://www.hep.com.cn
印 刷	涿州市星河印刷有限公司	网上订购	http://www.landraco.com
开 本	787mm×1092mm 1/16		http://www.landraco.com.cn
印 张	23.25	版 次	2013年1月第1版
字 数	440千字	印 次	2013年1月第1次印刷
购书热线	010-58581118	定 价	47.70元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换

版权所有 侵权必究

物料号 36416-00

“美国 MCM/ICM 竞赛指导丛书”
编审委员会

顾　问

Sol Garfunkel	美国数学及应用联合会（COMAP）
Chris Arney	美国西点军校

主　编

王　杰	美国麻省大学罗威尔分校/COMAP 中国合作部
朱　旭	西安交通大学

秘　书

王嘉寅	美国康涅狄格大学/西安交通大学

委　员（按姓氏拼音排序）

Jay Belanger	美国杜鲁门州立大学
陈秀珍	美国乔治华盛顿大学
冯国灿	中山大学
龚维博	美国麻省大学阿默斯特分校
韩中庚	解放军信息工程大学
李向阳	美国伊利诺伊理工大学
杨新宇	西安交通大学
叶正麟	西北工业大学
张存权	美国西弗吉尼亚大学

COMAP 总裁序

美国大学生数学建模竞赛（the Mathematical Contest in Modeling，MCM）已经举办近30年了，时间真是快得难以置信。在此期间，竞赛从最初参赛的90支美国队逐渐发展成为一个国际大赛，今年已有来自世界各地的25个国家超过5 000支队伍参赛。尤其令人感动和鼓舞的是我的中国同事们对竞赛赋予的极大热情以及中国参赛队伍的快速增长。COMAP张开双臂欢迎你们的参与。

COMAP每年举办3个建模竞赛，即MCM，ICM（the Interdisciplinary Contest in Modeling）和HiMCM（the High School Mathematical Contest in Modeling）竞赛。竞赛的目的不仅仅是奖励同学们所作出的努力——无疑这是同样重要的，我们举办各类数学建模竞赛的目的始终是为了推动在世界各国的各级教育体系中增加应用数学及数学建模的比重。建模是人们为了解世间事物的运作规律所做的尝试，数学的使用能够帮助我们建立更好的模型。这不是一个国家的任务，而是所有国家都应该共同关心的问题。COMAP建模竞赛从孕育到现在已经演变成实现这一宏伟目标的有力工具。

我热切地希望同学们通过阅读这套优秀的丛书，对COMAP竞赛有更多的了解，并且学习到更多有关数学建模的方法与过程。我希望同学们尝试自己解决丛书中讨论的所有建模问题，这些都是令人兴奋并且具有实用价值的问题。我希望更多的同学参加MCM/ICM竞赛，并参与推广和普及数学建模的活动，这是很有意义的工作。

萨尔·加芬克尔，博士

COMAP总裁

2012年11月

Forward by Sol Garfunkel

While it is hard for me to believe, the Mathematical Contest in Modeling (MCM) is fast approaching its 30th year. During this time we have grown from 90 US teams to over 5,000 teams representing 25 countries from all across the globe. We have been especially buoyed by the enthusiasm shown by our Chinese colleagues and the rapid growth in Chinese participation. COMAP welcomes your involvement with open arms.

COMAP runs three contests in mathematical modeling; they are MCM, ICM (the Interdisciplinary Contest in Modeling), and HiMCM (the High School Mathematical Contest in Modeling). The purpose of all of these contests has never been simply to reward student efforts—as important as that is. Rather, our objective from the beginning has been to increase the presence of applied mathematics and modeling in education systems at all levels worldwide. Modeling is an attempt to learn how the world works and the use of mathematics can help us produce better models. This is not a job for one country, but for all. The COMAP modeling contests were conceived and evolved to be strong instruments to help achieve this much larger goal.

It is my supreme hope that through this excellent book series the Chinese students will learn more about COMAP contests and more about the process of mathematical modeling. I hope that you will begin to work on the exciting and important problems you see here, and that you will join the MCM/ICM contests and the rewarding work of increasing the awareness of the importance of mathematical modeling.

Sol Garfunkel, PhD
Executive Director
COMAP
November 2012

ICM 竞赛主席序

数学建模的训练与经验能使同学们在解决问题时更有创意，同时也能帮助同学们成为更为优秀的研究生。“美国 MCM/ICM 竞赛指导丛书”的出版，将通过数学建模竞赛题目和概念的解析，帮助同学们掌握数学建模的技能，并为同学们在今后的工作中获得成功打下坚实的基础。

数学建模是一种过程，也是一种理念，或者说是一种哲学。作为过程，学生在理解及使用建模过程或框架时需要指导并积累经验。作为经验，学生需要使用不同的数学方法（离散、连续、线性、非线性、随机、几何及分析）构造数学模型，从中体验不同的细节及复杂程度。作为理念，学生需要发现各种相关的、具有挑战性的及有趣的实际问题，从中培养对数学建模的兴趣，并认识到数学建模在实际生活中的作用。数学建模的主要目的是指导学生用建模的方法解决实际问题。尽管在实际中，有些问题或许可以使用已有的算法和公式来求解，但数学建模的方法比简单使用已有算法和公式能解决更多的问题，特别是解决新的、没有固定答案及没有被解决过的问题。

为了积累经验，同学们应尽早地接受数学建模训练，至少应该在大学低年级时就开始，这样可以在以后的课程学习中进一步强化数学建模能力。由于数学建模的综合与交叉特性，所以各个专业的学生都能够从数学建模活动中受益。

本套丛书从将数学模型作为研究工具的角度出发，介绍模型的构造，分析建模过程，这些都是帮助学生更好地掌握数学建模技能的重要因素。数学建模是充满挑战的高级技能，更重要的是能够帮助学生更好地成长。当今世界需要解决的问题往往很复杂，所以建立的数学模型也很复杂，通常需要通过精细的计算和模拟才能获得解答或对模型结果的分析与检验。由于数据可视化技术的普及，解题方法的增加，所以现在的确是培养更多数学建模高手的最佳时期。

我希望同学们在数学建模探索中取得进步，也希望指导教师在使用这套丛书提供的例子及方法指导学生时取得很好的效果。尽管学生的层次可能不同，但我对你们的忠告是同样的：树立你的信心，发展你的技能，用你的才能解决社会中最具挑战及最重要的问题。祝各位建模好运！

克里斯·阿尼，博士
美国西点军校数学系教授
ICM 竞赛主席
2011 年 10 月

Forward by Chris Arney

Undergraduate students who receive instruction and experiences in mathematical modeling become better and more creative problem solvers and graduate students. This book series is being published to prepare and educate students on the topics and concepts of mathematical modeling to help them establish a problem solving foundation for a successful career.

Mathematical modeling is both a process and a mindset or philosophy. As a process, students need instruction and experience in understanding and using the modeling process or framework. As part of their experience, they need to see various levels of sophistication and complexity, along with various types of mathematical structures (discrete, continuous, linear, nonlinear, deterministic, stochastic, geometric, and analytic). As a mindset, students need to see problems that are relevant, challenging, and interesting so they build a passion for the process and its utility in their lives. A major goal in modeling is for students to want to model problems and find their solutions. Recipes for structured or prescribed problem solving (canned algorithms and formulas) do exist in the real world, but mathematical modelers can do much more than execute recipes or formulas. Modelers are empowered to solve new, open, unsolved problems.

In order to build sufficient experience in modeling, student exposure must begin as early as possible – definitely by the early undergraduate years. Then the modeling process can be reinforced and used throughout their undergraduate program. Since modeling is interdisciplinary, students from all areas of undergraduate study benefit from this experience.

The articles and chapters in this series expose the readers to model construction, model analysis, and modeling as a research tool. All these areas are important and build the students' modeling skills. Modeling is a challenging and advanced skill, but one that is empowering and important in student development. In today's world, models are often complex and require sophisticated computation or simulation to pro-

vide solutions or insights into model behavior. Now is an exciting time to be a skilled modeler since methodology to provide visualization and find solutions are more prevalent and more powerful than ever before.

I wish the students well in their adventure into modeling and I likewise wish faculty well as they use the examples and techniques in this book series to teach the modeling process to their students. My advice to all levels of modelers is to build your confidence and skills and use your talents to solve society's most challenging and important problems. Good luck in modeling!

Chris Arney, PhD
United States Military Academy at West Point
Professor of Mathematics
Director of the Interdisciplinary Contest in Modeling
October, 2011

丛书简介

美国大学生数学建模竞赛（the Mathematical Contest in Modeling，MCM/the Interdisciplinary Contest in Modeling，ICM），即“数学建模竞赛”和“交叉学科建模竞赛”，是一项国际级的竞赛活动，为现今各类数学建模竞赛之鼻祖。

1985 年，在美国教育部的资助下，在美国针对在校大学生创办了一个名为“数学建模竞赛”的竞赛，其宗旨是鼓励大学师生对不同领域的各种实际问题进行阐明、分析并提出解决方案。它是一种完全公开的竞赛，参赛形式为学生三人组成一队，在三天（72 小时）（近年改为四天，即 96 小时）内任选一题，完成数学建模的全过程，并就问题的重述、简化和假设及其合理性的论述、数学模型的建立和求解（及软件）、检验和改进、模型的优缺点及其可能的应用范围的自我评述等内容写出论文。MCM/ICM 非常重视解决方案的原创性、团队合作与交流以及结果的合理性。由专家组成的评阅组进行评阅，评出优秀论文。除了不允许在竞赛期间与团队以外的任何人（包括指导教师）讨论赛题之外，允许使用图书资料、互联网上的资料、任何类型的软件等各种资料和途径，从而为参赛学生提供了广阔的创作空间。第一届竞赛时，只有美国的 158 个队参加，其中只有 90 个队提交了解答论文。2012 年 MCM/ICM 共有 5 026 个队参加，其中 MCM 有 3 697 个队，ICM 有 1 329 个队，遍及五大洲。MCM/ICM 已经成为最著名的国际大学生竞赛之一，影响极其广泛。

近年来，已有越来越多的中国学生组队参加美国大学生数学建模竞赛，其中不乏被评为最优论文（Outstanding Winners）的佼佼者，这充分显示了中国大学生参加 MCM/ICM 竞赛的积极性与实力。学生在准备竞赛的时候，除了在指导教师的帮助下阅读和研究以往竞赛的优胜论文以外，普遍希望能有一些专门针对美国大学生数学建模竞赛的书籍，指导和帮助备赛。

“美国 MCM/ICM 竞赛指导丛书”就是为了满足读者的这一需求而出版的，目的是帮助学生学习从全局出发，不受固定模式的限制，用建模的手段解决开放性问题的研究方法，并提高写作能力。丛书的读者对象包括参赛学生以及对数学建模与算法感兴趣的研究生、专业人员和业余爱好者。

我们邀请到 COMAP 中国合作总监、美国麻省大学罗威尔分校王杰教授担

任丛书主编，他曾为 MCM/ICM 命题，对竞赛具有很多独到的认识。丛书作者来自各高校，他们都是有经验的指导教师或参加过竞赛的优秀成员。丛书包括一本《正确写作美国大学生数学建模竞赛论文》和若干辑《美国大学生数学建模竞赛题解析与研究》，前者为一本指导学生如何正确写作 MCM/ICM 论文的工具书，后者中的每一辑将讨论若干赛题，包括问题的背景、分析技巧、建模与测试方法及算法设计，并引导读者列出进一步研究的课题。目标是培养学生多方面的能力，如数学、编程、写作及课题研究等的训练，提高学生分析问题、解决问题的水平。

丛书的出版计划得到了美国建模专家的广泛支持，COMAP 主席 Sol Garfunkel 博士以及 ICM 主席、美国西点军校数学系教授 Chris Arney 博士受邀担任丛书顾问并为丛书作序。

我们热切希望通过本套丛书的出版，进一步活跃我国大学生参加 MCM/ICM 的积极性，提高他们的自信心，并最终取得满意的成绩。更为重要的，提高学生的研究和解决实际问题的能力。

前言

自从1985年举办了第一届美国大学生数学建模竞赛（Mathematical Contest in Modeling，MCM）以来，到2012年已成功举办了28届，而且参赛队数逐年增加。特别是从1999年开始，又扩展增加了交叉学科的建模竞赛（Interdisciplinary Contest in Modeling，ICM），使得影响力进一步扩大。2011年共有17个国家和地区的3 510个队参赛，其中MCM有2 775个队，ICM有735个队。据不完全统计，来自中国各学校的参赛队占到了80%以上，这也充分反映出中国大学生对数学建模的参与热情。

正是在这个背景下，1992年11月由中国工业与应用数学学会（CSIAM）组织举办了“中国大学生数学建模竞赛”，并决定以后每年举行一次。从1994年开始，中国大学生数学建模竞赛正式由教育部高等教育司直接领导，由CSIAM具体组织，并被教育部确定为在全国推广实施的三大竞赛之一。2011年是中国大学生数学建模竞赛举办20周年，2011年的竞赛有包括香港和澳门在内的33个省市区的1 251所院校的19 490个队参赛，现已成为名副其实的面向全国高校大学生最大规模的科技竞赛活动。同时，2011年也有包括美国、新加坡、伊朗等国家的大学生参赛。

目前，美国大学生数学建模竞赛和中国大学生数学建模竞赛已成为最有影响力的两大赛事。纵览美国的竞赛题和中国的竞赛题，虽然从竞赛形式和要求看基本上是一致的，但从题目的提法和特点来看还是有很大的不同。目前，与中国大学生数学建模竞赛相关的书籍资料众多，但在国内与美国大学生数学建模竞赛直接相关的资料甚少，这也正是编写这套丛书的初衷。希望通过这套丛书的出版，能够对中国大学生参加MCM/ICM有所帮助，使中国的大学生能在竞赛中取得优异的成绩，也有利于加快培养创新型的人才。随着数学建模教学、竞赛和应用研究等活动的不断深入开展，必将对大学的数学教学改革、提高教学质量、促进创新人才培养等工作产生积极的推动作用。我们期待以时代为契机，以需求为牵引，不断推陈出新，一定会迎来数学建模活动的新发展。

本书是这套丛书的2005—2006年分册，包括2005年和2006年MCM/ICM的6个问题的提出、问题的背景、解决问题的数学模型、求解方法与结果分

析，以及进一步可研究拓展的问题等内容。这些赛题所涉及的问题大都来自于工程和社会生活中的实际问题，这些问题的研究对于从事实际工作的工程技术人员也有一定的学习参考价值。由于本书题目均来自美国的实际问题，问题中多采用英制。为了将重点关注内容集中在问题的分析上，故未将其转换为我国通行的国际单位制。

本书的第1章由田园编写、第2章由刘靖旭编写、第3章和第6章由杜剑平编写、第4章由郭晓丽编写、第5章由韩中庚编写，全书由韩中庚修改统稿。在本书的编写过程中，得到了参编院校的各级领导和同事们的大力支持；本丛书主编王杰教授和编委会的各位教授、同仁也都给予了指导与帮助，并提出了很多好的意见；在本书编辑出版过程中，高等教育出版社自然科学学术著作出版事业部刘英编辑从组织策划到编辑出版都付出了辛勤的劳动，使该书得以顺利出版。在此，编者一并表示衷心的感谢。

限于作者水平，书中肯定会有不妥之处，敬请读者批评指正。

目录

COMAP 总裁序
Forward by Sol Garfunkel
ICM 竞赛主席序
Forward by Chris Arney
丛书简介
前言

第1章 洪水估计问题 …… 1

1.1 问题的综述 …… 1
1.1.1 问题的提出 …… 1
1.1.2 问题的背景资料 …… 2
1.1.3 问题的现实意义 …… 3
1.2 逐渐溃坝模型 …… 4
1.2.1 基本模型 …… 5
1.2.2 溃口流量模型 …… 5
1.2.3 各段河道之间的流量模型 …… 7
1.2.4 河道的截取和分析 …… 9
1.2.5 模型的参数估计 …… 11
1.2.6 模型的结果分析 …… 11
1.3 瞬间溃坝模型 …… 13
1.3.1 模型的建立 …… 13
1.3.2 模型的求解 …… 14
1.3.3 模型的结果分析 …… 15
1.4 差分形式的圣维南模型 …… 16
1.4.1 问题分析与假设 …… 16
1.4.2 坝体溃口模型 …… 17
1.4.3 圣维南模型 …… 17

1.4.4 模型的求解 …… 19
1.5 离散网格模型 …… 22
1.5.1 问题的分析 …… 22
1.5.2 模型的假设与符号说明 …… 23
1.5.3 四种溃坝模型 …… 24
1.5.4 下游水流模型 …… 27
1.5.5 模型的模拟求解 …… 32
1.5.6 模型的结果分析 …… 32
1.6 问题的综合分析与进一步研究 …… 34
1.6.1 问题的综合分析 …… 34
1.6.2 进一步研究的问题 …… 37

第2章 高速公路收费亭设置问题 …… 44

2.1 问题的综述 …… 44
2.1.1 问题的提出 …… 44
2.1.2 问题的背景资料 …… 45
2.2 基于宏观交通流的收费站设置模型 …… 47
2.2.1 问题的分析与假设 …… 47
2.2.2 名词解释与符号说明 …… 48
2.2.3 模型的准备 …… 49
2.2.4 模型的建立与求解 …… 50
2.2.5 模型的检验与结果分析 …… 53
2.2.6 模型的推广 …… 55
2.3 收费站的车辆排队模型 …… 55
2.3.1 名词解释与符号说明 …… 55
2.3.2 模型1：单队列多服务台排队系统模型 …… 57
2.3.3 模型2：多个单服务台排队系统模型 …… 62
2.4 基于安全距离的车辆跟驰模型 …… 69
2.4.1 问题的分析与假设 …… 70
2.4.2 名词解释与符号说明 …… 70
2.4.3 基本的跟驰模型 …… 72
2.4.4 单车道车辆跟驰模型 …… 73
2.4.5 多车道车辆行为模型 …… 76
2.4.6 模型参数的确定 …… 77
2.4.7 模型的仿真与结果分析 …… 79
2.5 基于元胞自动机的交通流仿真模型 …… 81

2.5.1 问题的分析与假设 …… 81
2.5.2 名词解释与符号说明 …… 82
2.5.3 交通流元胞自动机模型 …… 83
2.5.4 收费站的元胞自动机模型 …… 85
2.5.5 模型的仿真与结果分析 …… 88
2.5.6 模型的检验与分析 …… 90
2.6 问题的综合分析与进一步研究 …… 91
2.6.1 问题的综合分析 …… 91
2.6.2 进一步研究的问题 …… 93

第3章 不可再生资源的管理问题 …… 99

3.1 问题的综述 …… 99
3.1.1 问题的提出 …… 99
3.1.2 竞赛的基本情况 …… 100
3.2 问题的背景资料 …… 102
3.2.1 不可再生资源概念 …… 102
3.2.2 Hubbert 峰值理论 …… 102
3.2.3 Michael Olinick 理论 …… 104
3.2.4 需求价格弹性理论 …… 105
3.2.5 OPEC 对石油供求关系的影响 …… 106
3.3 问题的分析与假设 …… 107
3.3.1 问题（1）：预测不可再生资源的枯竭日期 …… 107
3.3.2 问题（2）：多因素时不可再生资源耗尽时间的预测 …… 109
3.3.3 问题（3）：资源的公平开采策略 …… 111
3.3.4 问题（4）：资源的可持续开采策略 …… 111
3.3.5 问题（5）：考虑环境因素的资源开采策略 …… 112
3.3.6 问题（6）：存在替代资源的开采策略 …… 112
3.4 石油资源的预测模型 …… 113
3.4.1 模型 1.1：石油资源耗尽时间的预测模型 …… 113
3.4.2 模型 1.2：考虑多因素的石油耗尽时间预测模型 …… 116
3.4.3 模型 1.3：石油资源的可持续开发模型 …… 119
3.4.4 模型 1.4：石油资源的安全开发模型 …… 121
3.4.5 模型 1.5：石油开发对自然环境的危害模型 …… 122
3.4.6 模型 1.6：新能源的替代石油资源的模型 …… 123
3.5 石油资源的消耗与控制模型 …… 126
3.5.1 模型 2.1：石油资源耗尽时间预测模型 …… 126

3.5.2 模型2.2：多因素影响石油消耗模型 …… 129
3.5.3 模型2.3：新能源对石油资源的影响模型 …… 133
3.6 世界水资源的预测与管理模型 …… 137
3.6.1 模型3.1：美国年用水量的预测模型 …… 137
3.6.2 模型3.2：水资源分类消耗的预测模型 …… 139
3.7 问题的综合分析与进一步研究 …… 143
3.7.1 问题的综合分析 …… 143
3.7.2 进一步研究的问题 …… 146

第4章 灌溉喷洒系统设置问题 …… 148

4.1 问题的综述 …… 148
4.1.1 问题的提出 …… 148
4.1.2 问题的背景资料 …… 149
4.1.3 问题的分析 …… 150
4.2 模型1：基于多步迭代算法的农田灌溉优化模型 …… 151
4.2.1 问题的分析与假设 …… 151
4.2.2 模型的建立 …… 152
4.2.3 模型的求解方法 …… 154
4.2.4 模型的结果分析 …… 156
4.2.5 模型的分析与评价 …… 157
4.3 模型2：多管组喷灌系统的优化模型 …… 159
4.3.1 问题的分析与假设 …… 159
4.3.2 模型的建立 …… 160
4.3.3 模型的求解方法 …… 162
4.3.4 模型的结果分析 …… 166
4.3.5 模型的分析与评价 …… 167
4.4 模型3：手动喷灌系统的优化模型 …… 167
4.4.1 问题的分析与假设 …… 167
4.4.2 模型相关参数的确定方法 …… 168
4.4.3 模型的建立与求解 …… 172
4.4.4 模型的模拟求解算法 …… 177
4.4.5 模型的结果分析 …… 180
4.5 模型4：基于不同特性的喷灌系统模型 …… 181
4.5.1 模型的预备知识 …… 181
4.5.2 喷灌作业方案的确定模型 …… 185
4.5.3 模型的分析与评价 …… 187

4.6 问题的综合分析与进一步研究 …… 188
4.6.1 问题的综合分析 …… 188
4.6.2 进一步研究的问题 …… 190

第5章 机场轮椅配置问题 …… 193

5.1 问题的综述 …… 193
5.1.1 问题的提出 …… 193
5.1.2 问题的背景资料 …… 195
5.2 基于多候机厅的配置模型 …… 196
5.2.1 名词解释与符号说明 …… 197
5.2.2 问题的分析与假设 …… 197
5.2.3 多候机厅模型的建立 …… 198
5.2.4 模型的求解实现 …… 202
5.2.5 模型的结果分析 …… 202
5.3 基于航空公司间竞争的配置模型 …… 203
5.3.1 航空公司的市场占有率模型 …… 203
5.3.2 航空公司的乘客流失模型 …… 204
5.3.3 航空公司的利润函数模型 …… 206
5.3.4 基于利润的轮椅配备模型 …… 206
5.3.5 模型的求解实现 …… 207
5.3.6 模型的结果分析 …… 207
5.4 基于最小成本的配置模型 …… 208
5.4.1 问题的分析与假设 …… 208
5.4.2 单项因素的费用模型 …… 209
5.4.3 航空公司的总费用模型 …… 210
5.4.4 模型的求解算法 …… 211
5.4.5 模型的应用与分析 …… 213
5.4.6 模型的结果分析 …… 216
5.5 基于成本效益的配置模型 …… 217
5.5.1 问题的分析 …… 217
5.5.2 模型的建立方法 …… 218
5.5.3 模型的仿真算法 …… 221
5.5.4 运行成本的计算 …… 223
5.5.5 模型的检验与分析 …… 224
5.5.6 模型的结果分析 …… 229
5.6 基于最小费用流的配置模型 …… 230

5.6.1 基本术语与模型假设 …… 230
5.6.2 模型的建立 …… 231
5.6.3 模型的求解算法 …… 234
5.6.4 模型的模拟检验 …… 238
5.6.5 模型的评价 …… 243
5.7 问题的综合分析与进一步研究 …… 244
5.7.1 问题的综合分析 …… 244
5.7.2 进一步研究的问题 …… 246

第6章 艾滋病毒防控资源分配问题 …… 250

6.1 问题的综述 …… 250
6.1.1 问题的提出 …… 250
6.1.2 2006年ICM竞赛的相关情况 …… 258
6.1.3 问题的背景资料 …… 259
6.1.4 常见的传染病数学模型 …… 261
6.1.5 对问题的分析 …… 262
6.2 解决两个基本问题 …… 266
6.2.1 基本问题（1）的解决方案 …… 266
6.2.2 基本问题（2）的解决方案 …… 271
6.3 模型1：基于SI模型的HIV传播模型 …… 273
6.3.1 模型1.1：HIV自由传播模型 …… 273
6.3.2 模型1.2：考虑ARV治疗的传播模型 …… 275
6.3.3 模型1.3：存在HIV疫苗的传播模型 …… 276
6.3.4 模型1.4：考虑ARV治疗和HIV疫苗的传播模型 …… 278
6.3.5 模型1.5：考虑ARV治疗抗药性的传播模型 …… 279
6.3.6 模型1.6：同时考虑ARV治疗抗药性和疫苗的传播模型 …… 280
6.3.7 模型的分析与评价 …… 281
6.4 模型2：基于Logistic模型的HIV传播模型 …… 282
6.4.1 模型2.1：HIV自由传播模型 …… 282
6.4.2 模型2.2：考虑ARV治疗的Logistic传播模型 …… 283
6.4.3 模型2.3：考虑HIV疫苗的Logistic传播模型 …… 284
6.4.4 模型2.4：考虑ARV治疗和HIV疫苗的Logistic传播模型 …… 284
6.4.5 模型2.5：考虑ARV治疗抗药性的Logistic传播模型 …… 285
6.4.6 模型2.6：考虑HIV疫苗和ARV治疗抗药性的Logistic传播模型 …… 285
6.4.7 模型的分析与评价 …… 286

6.5 模型3：考虑出生和死亡的HIV传播模型 …… 287
6.5.1 模型3.1：考虑出生和死亡的HIV自由传播模型 …… 287
6.5.2 模型3.2：考虑ARV治疗的HIV传播模型 …… 291
6.5.3 模型3.3：考虑存在HIV疫苗的传播模型 …… 292
6.5.4 模型3.4：考虑ARV治疗和存在HIV疫苗的传播模型 …… 292
6.5.5 模型3.5：考虑ARV治疗和抗药性的传播模型 …… 293
6.5.6 模型3.6：考虑ARV治疗抗药性和存在HIV疫苗的传播模型 …… 294
6.5.7 模型的分析与评价 …… 296
6.6 模型4：HIV的多级传播模型 …… 296
6.6.1 模型4.1：HIV多级自由传播模型 …… 296
6.6.2 模型4.2：考虑ARV治疗的HIV多级传播模型 …… 299
6.6.3 模型4.3：考虑HIV疫苗的多级传播模型 …… 302
6.6.4 模型4.4：考虑ARV治疗和HIV疫苗的多级传播模型 …… 303
6.6.5 模型4.5：考虑ARV治疗抗药性的多级HIV传播模型 …… 305
6.6.6 模型4.6：考虑ARV抗药性和HIV疫苗的多级HIV传播模型 …… 308
6.6.7 模型的分析与评价 …… 309
6.7 模型5：考虑不同传播途径的HIV传播模型 …… 310
6.7.1 模型5.1：HIV自由传播模型 …… 310
6.7.2 模型5.2：考虑ARV治疗的HIV传播模型 …… 314
6.7.3 模型5.3：考虑HIV疫苗的多传播途径传播模型 …… 317
6.7.4 模型5.4：考虑ARV治疗和HIV疫苗的多途径传播模型 …… 318
6.7.5 模型5.5：考虑ARV治疗抗药性的HIV多途径传播模型 …… 320
6.7.6 模型5.6：考虑ARV治疗与抗药性和HIV疫苗的传播模型 …… 321
6.7.7 模型的分析与评价 …… 323
6.8 模型6：HIV复杂动态网络传播模型 …… 324
6.8.1 模型6.1：HIV自由传播模型 …… 324
6.8.2 模型6.2：考虑ARV治疗的HIV传播仿真模型 …… 328
6.8.3 模型6.3：考虑HIV疫苗的HIV传播仿真模型 …… 329
6.8.4 模型6.4：考虑ARV治疗和HIV疫苗的HIV传播仿真模型 …… 330
6.8.5 模型6.5：考虑ARV抗药性的HIV传播仿真模型 …… 330
6.8.6 模型6.6：考虑ARV抗药性和HIV疫苗的HIV传播仿真模型 …… 331
6.8.7 模型的分析与评价 …… 331
6.9 HIV传播模型仿真求解 …… 332
6.9.1 自由传播模型的Simulink仿真求解 …… 332
6.9.2 考虑ARV治疗传播的Simulink仿真模型 …… 334
6.9.3 考虑ARV治疗以及HIV疫苗的Simulink仿真模型 …… 336

6.9.4 考虑 ARV 抗药性与 HIV 疫苗的 Simulink 仿真模型 …… 338
6.10 HIV/AIDS 的预防控制策略 …… 339
6.10.1 HIV/AIDS 防控资金的最优配置模型 …… 339
6.10.2 给联合国相关部门的建议书 …… 341
6.11 写在后面：ICM 优秀论文评阅标准与启示 …… 342

第1章　洪水估计问题

洪水估计问题是2005年美国大学生数学建模竞赛的A题，研究的是公共安全领域自然灾害预防与控制方面的问题．这个问题的研究在现实生活中有十分重要的意义．特别是在有大型水库和湖泊的地区，由于地震、暴雨造成的山洪和泥石流等自然灾害频发，做好预防性研究工作是非常有现实意义的．

1.1　问题的综述

1.1.1　问题的提出

墨里湖位于美国南卡罗来纳州中部，是由一座用于发电的大型土坝形成，大坝建成于1930年．假设一场灾难性的大地震造成大坝决口，试对由此产生的下游洪水进行建模．有两个值得注意的问题：

（1）罗尔斯河是萨卢达河的一条支流，四季不断流，距大坝下游很近．当大坝决口形成洪水时，罗尔斯河将遭受多大的洪水？洪水会向上游倒灌延伸多远？

（2）洪水会波及南卡罗来纳州议会大厦吗？议会大厦位于康加里河边的一座小山上．

问题的原文如下：

Lake Murray in central South Carolina is formed by a large earthen dam, which was completed in 1930 for power production. Model the flooding downstream in the event there is a catastrophic earthquake that breaches the dam.

Two particular questions:

Rawls Creek is a year-round stream that flows into the Saluda River a short distance downriver from the dam. How much flooding will occur in Rawls Creek from a dam failure, and how far back will it extend?

Could the flood be so massive downstream that water would reach up to the S. C. State Capitol Building, which is on a hill overlooking the Congaree River?

1.1.2 问题的背景资料

1. 萨卢达河（Saluda River）

萨卢达河发源于南卡罗来纳州北部与北卡罗来纳州交界处，由南北两条支流于格林维尔（Greenville）西北约16公里处汇合而成，在哥伦比亚与布罗德河（Broad River）汇合后形成康加里河（Congaree River），总流程约320 km. 最后作为桑蒂河（Santee River）的一部分流入太平洋[1]. 如图1-1所示.

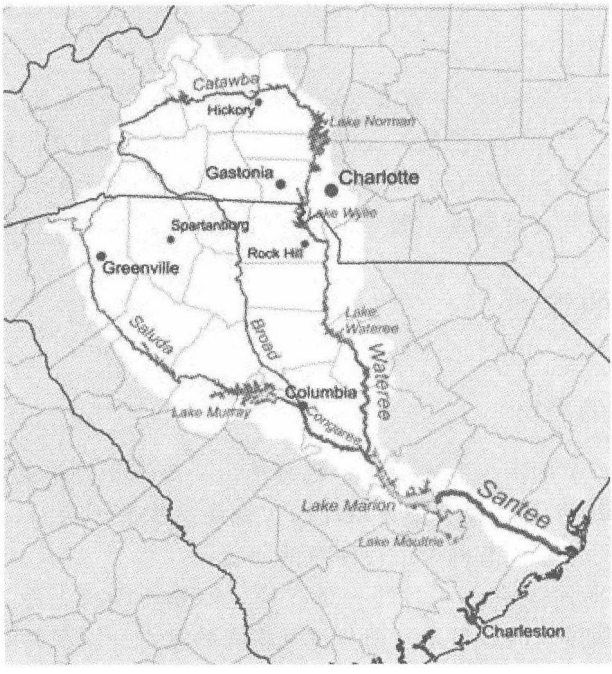


图1-1 萨卢达河流域

截至2011年，萨卢达河上共有10座水坝，其中2座属于大坝（坝高超过15 m），分别是萨卢达大坝和格林伍德大坝（Greenwood Dam），均用于水力发电.

2. 萨卢达大坝（Saluda Dam）

20世纪20年代末，南卡罗来纳州在萨卢达河上修建了一座大坝用以提供水力发电. 萨卢达大坝的官方名称是德里赫沙坝（Dreher Shoals Dam），位于哥伦比亚以西15 km，长约2 400 m，坝高61 m，由当地的红色黏土垒制而成，建造工程始于1927年，于1930年建成，是当时世界上最大的土石坝[2].

3. 萨卢达大坝补救工程

1886年，美国东北部曾发生一次强烈的地震，使南卡罗来纳州的查理斯顿市（Charleston）受到重创，60人丧生，几乎所有建筑物都受到了不同程度

的破坏. 工程师们在对南卡罗来纳州的关键设施的抗震设计评估时，认为这一地区类似震级的地震（估计为里氏 7.1 ~ 7.3 级）在今后 1 000 年的某个时期可能会再次发生，其中墨里湖尤其危险，因为其位置正好位于地震断裂带上[3].

在 20 世纪 80 年代后期对萨卢达大坝进行地震评估时，工程师们认为如果发生 1886 年的地震，有可能使萨卢达大坝溃坝，下游超过 12 万人将处于危险之中，而且还会失去由大坝提供的电力供应，使当地的经济陷入停顿. 为此，工程师们自 1989 年起开始分析大坝溃坝情形并研究对策，同时着手实施一项补救工程. 在对各种可选方案进行评估后，最终决定保持现有建筑物不变，在大坝坝址修建一座备用戗道，即一座应急“干坝”，如果现有大坝溃坝，它将截住坝体所形成的泥石流[4].

备用戗道为碾压混凝土坝，几乎与萨卢达大坝等长等高，其设计目标是防止大坝发生灾难性的洪水泛滥，保证原大坝相关设施能安全关闭.

需要说明的是，尽管该工程的完工时间是 2005 年，但是在 2005 年 1 月 MCM 竞赛时，备用戗道仍在建设中，并未发挥作用，因此命题者和参赛选手均未考虑其影响.

4. 墨里湖（Lake Murray）

虽然墨里湖被称为“湖”，实际上它是修建水坝形成的一座水库，位于美国南卡罗来纳州中部，以萨卢达大坝首席工程师威廉·墨里（William S. Murray）的名字命名. 墨里湖位于北纬 34.06 度西经 81.33 度，距离哥伦比亚市约 20 km. 湖面面积大约 200 km^2，东西方向最大长度 61 km，南北方向最大长度 23 km，最大深度 61 m，湖岸线大约 800 km，储水量约 30 亿 m^3. 1930 年完成第一次蓄水，是当时世界上最大的人造水库[5].

5. 土石坝

土石坝是指由当地土料、石料或土石混合料填筑而成的坝，又称当地材料坝. 坝体材料以土和沙砾为主时，称为土坝；以石渣、卵石、块石为主时，称为堆石坝；土、石料均占有一定比例时，称为土石混合坝. 三者在工作条件、结构形式和施工方法上均有相似之处，通称土石坝. 土石坝在世界上历史最为悠久，应用最为广泛，随着近年来大型土方施工机械、岩土理论和计算技术的发展，放宽了对筑坝材料的使用范围，缩短了工期，也使土石坝成为当今世界坝工建设中发展最快的一种坝型[6]. 在我国已建成的八万多座水坝中，土石坝约占 90%，其中最高的是小浪底土石坝，坝高 154 m.

1.1.3 问题的现实意义

溃坝的主要危害是库内大量水体突然下泄形成的溃坝洪水，其洪峰流量、运动速度、破坏力远远大于一般暴雨洪水或融雪洪水，所到之处荡尽一切，造

成毁灭性的灾害. 溃坝洪水的破坏能力与库蓄水体、坝前上游水深、水头、溃决过程及坝址下游河道的两岸地形有密切关系，坝前库蓄水体越大，坝址水头越高，破坏力越大. 表1-1给出了世界重大溃坝事件和死亡人数[7].

表1-1 世界重大溃坝事件和死亡人数统计

大坝名称	国家	溃坝年份	坝型	坝高 (m)	库容 ($\times10^8 m^3$)	溃坝流量 (m^3/s)	死亡人数
South Fork	美国	1889	土石坝	19.0	0.3		2 209
St. Francis	美国	1928	重力坝	62.5	0.5	14 200	236
Mohne	德国	1943	重力坝	40.0	1.34	8 800	1 200
Malpasset	法国	1959	拱坝	66.5	0.15		421
Vajont	意大利	1963	拱坝	262.0	1.69	漫坝	2 000
Teton	美国	1976	土石坝	93.0	3.08		14
Machhu II	印度	1979	土石坝	43.3	1.01	9 203	3 000
Buffalo Creak	美国	1972	尾矿坝	14.0	0.498		125

我国现有各类水库大坝8.5万多座，其中库容超过1亿 m^3 的大型水库就有400多座，这些大坝一旦发生溃坝，对下游地区的影响是毁灭性的. 根据1954—2000年的统计资料显示，我国已发生的溃坝，大型水库两座，中型水库124座，小（1）型水库668座，小（2）型水库2 668座，总计3 462座，失事率为3.4%. 两座大型水库分别是1975年河南“七五·八”洪水中溃决的板桥和石漫滩水库，溃坝洪峰流量分别达到了78 100 m^3/s 和20 000 m^3/s.

因此，2005年A题所研究的大坝溃决以及洪水影响范围问题，无论是在我国还是在世界范围内，都具有十分重要的现实意义.

1.2 逐渐溃坝模型

土石坝的溃坝有两种可能的情况，一是由于溃口扩大造成的渐进式溃坝，水文学上称其为逐渐溃；二是突发的灾难性事件造成大坝垮塌，称为瞬间溃. 绝大多数的土石坝溃决属于逐渐溃，溃决过程持续时间较长，坝体被洪水逐渐冲溃[8]. 逐渐溃坝模型的机理是：

（1）从水库里通过坝体溃口泻出的水流，通过冲刷和坍塌导致溃口的扩大，这一过程将持续到水库放空或者坝体能抵挡得住水流的进一步冲刷为止.

（2）溃口的形成时间长短主要取决于外泄的水流对筑坝材料的冲刷，与坝高、筑坝材料、材料的密实程度及漫顶泄流状况紧密相关.

（3）溃口在横向、垂向同时发展变化，随着时间的推移，由于斜坡失去

稳定性导致坍塌，从而引起顶部溃口逐渐扩大.

利用逐渐溃坝模型解决洪水估计问题，关键在于估计溃坝洪峰到达罗尔斯河和州议会大厦时的水位高度，假设地形不会因为洪水发生变化，则问题转变为估计洪峰通过上述两个位置时的流量. 这一问题可以分为两个部分进行建模，一是模拟大坝溃口在逐渐扩大的过程中通过溃口的流量变化，二是根据流量的变化，估计洪峰通过下游河道各点的流量[9].

1.2.1 基本模型

逐渐溃坝过程中，针对进入下游的水流量相对变化率，可以利用经典的明渠水力学建立相关的模型. 然后将下游的河道分为数段，运用曼宁方程（Manning formula）和连续方程来描述水流在它们之间的运动. 第一段河道的流量，可以用一个大坝受到溃口腐蚀的模型来给出. （注意本章所说的“流量”，指的是水流的速度，其单位为 m^3/s.）

对任意一段河道，容积的变化速度等于流入的流量与流出的流量之差，即

$$\frac{dV}{dt}=Q_{in}-Q_{out}, \tag{1.1}$$

其中 V 是这段河道的容积，t 是时间，Q_{in} 是流入的流量，Q_{out} 是流出的流量.

从萨卢达大坝到哥伦比亚，萨卢达河没有较大的支流，因此溃坝造成的洪水流量远大于其他支流的流量，可以假设当前时刻流入的流量只来自萨卢达河，也只流出到萨卢达河，即不考虑任何支流的影响. 通过对卫星地图的观察，考虑罗尔斯河出口和州议会大厦的位置，将萨卢达河分为 4 段，根据式(1.1)，每一段的河道有如下关系

$$\frac{dV_i}{dt}=Q_{i-1}-Q_i, \tag{1.2}$$

其中 V_i 是第 i 段河道的容积（从大坝开始往下游编号），$Q_i(i=1, 2, 3, 4)$ 是第 i 段河道流出的流量，Q_0 是从大坝溃口流入第 1 段河道的流量.

1.2.2 溃口流量模型

大坝溃口的被腐蚀过程是水流与坝体材料的相互作用. 溃口一旦形成，不断涌出的水流会冲刷大坝，造成溃口的扩大，而扩大的溃口使得水流速度增大，从而进一步加快冲刷的速度. 如果溃口没有被堵上，这种相互作用会一直持续到水库里的水排空.

这里还需要做一个假设，在溃口出现之前流入和流出水库的流量忽略不计，也就是假设水库的库容水量保持不变. 这是基于两方面的考虑，一是这两者存在相互抵消作用，二是从数量上来看这些流量相对于洪水来说可以忽略不计.

从溃口流出的流量 Q_0 等于水面下降的速度和水面面积 $A_s(H)$ 的乘积. 同时，溃口流量也可以描述为水流的平均速度 u 和溃口的横截面面积 A_b 的乘积，其中 A_b 仅包括溃口有水流通过的部分. 由此可得方程：

$$A_s(H)\frac{\mathrm{d}H}{\mathrm{d}t}=-Q_0=-uA_b. \tag{1.3}$$

实验结果表明，通过溃口的水流量可以通过宽顶堰流量的水力学方程来近似得到[10]，即

$$u=\alpha(H-Z)^{\beta}, \tag{1.4}$$

其中 Z 是溃口底部距离湖底的高度，参数 $\alpha=[(2/3)^3 g]^{1/2}=1.7$ m/s，$\beta=1/2$[11].

进一步假设水库的水面面积 A_s 与高度无关，也就是假设水库的边缘是垂直的. 根据式（1.3）和式（1.4）可得

$$A_s\frac{\mathrm{d}H}{\mathrm{d}t}=-uA_b=-\alpha(H-Z)^{\beta}A_b. \tag{1.5}$$

用一个简单且被证实有效的模型来描述大坝的被腐蚀过程[11]：

$$\frac{\mathrm{d}Z}{\mathrm{d}t}=-\gamma u^{\phi}=-\gamma\alpha^{\phi}(H-Z)^{\phi\beta}, \tag{1.6}$$

其中 γ 和 ϕ 可以由对大坝的材质进行分析得出，u 由式（1.4）给出. 由于大坝的具体数据未知，出于简化计算考虑，不妨假设 $\phi=1$，也就是线性腐蚀模型，同时根据以往的实验数据[11]，令 $\gamma\approx0.01$. 联立式（1.5）和式（1.6），可得水位高度 H 和溃口高度 Z 关于时间 t 的一阶微分方程组.

要得到这一方程组的结果，还需要确定溃口的形状. 大坝的溃口一般是三角形、梯形和矩形，其中以矩形进行近似最为常用，这是因为矩形溃口的常微分方程组（式（1.5）和式（1.6））求解相对容易.

建立矩形溃口的模型，设定宽度 b 为常数，腐蚀只发生在垂直方向上，则水流通过溃口的面积为

$$A_b=b(H-Z). \tag{1.7}$$

将式（1.7）代入式（1.5），将参数 $\phi=1$ 和 $\beta=1/2$ 代入式（1.6），可得

$$\begin{cases}\dfrac{\mathrm{d}H}{\mathrm{d}t}=-\dfrac{\alpha b}{A_s}(H-Z)^{\frac{3}{2}},\\[2ex] \dfrac{\mathrm{d}Z}{\mathrm{d}t}=-\gamma\alpha\,(H-Z)^{\frac{1}{2}}.\end{cases} \tag{1.8}$$

求解方程组（1.8），得

$$\begin{cases}H(Z)=Z+q+Ce^{\left(-\frac{Z_0-Z}{q}\right)},\\[2ex] t(Z)=\dfrac{2\sqrt{q}\arctan\left(\sqrt{\dfrac{H(Z)-Z}{q}}\right)}{\gamma\alpha}-D,\end{cases} \tag{1.9}$$

其中 $q=A_s\gamma/b$，$C=H_0-Z_0-q$，H_0 和 Z_0 是 $t=t_0$ 时刻 H 和 Z 的初值，D 是一个由初始条件决定的常数. 显然，$Z(t)$ 和 $H(t)$ 的数值可以由式（1.9）确定，通过溃口的流量 Q_0 可由式（1.3）和式（1.5）确定，即

$$Q_0=-\alpha\left[q+Ce^{\left(-\frac{Z_0-Z}{q}\right)}\right]^{\frac{1}{2}}A_b.$$

当 $Z(t)=0$ 的时刻 $\tilde{t}$，大坝停止被腐蚀，通过式（1.8）的方程组得到

$$\frac{dH}{dt}=-\frac{\alpha b}{A_s}H^{\frac{3}{2}},$$

求解可得

$$H(t)=\left(\frac{1}{\sqrt{H_0}}+\frac{\alpha b}{2A_s}(t-t_0)\right)^{-2},\quad t\geqslant\tilde{t}.$$

如图 1-2（a）和（b）所示，分别给出了流量 Q_0、溃口高度 Z 与时间 t 的关系，曲线在 $\tilde{t}\approx 2$ 处的拐点就是溃坝过程中两个阶段的交界点.

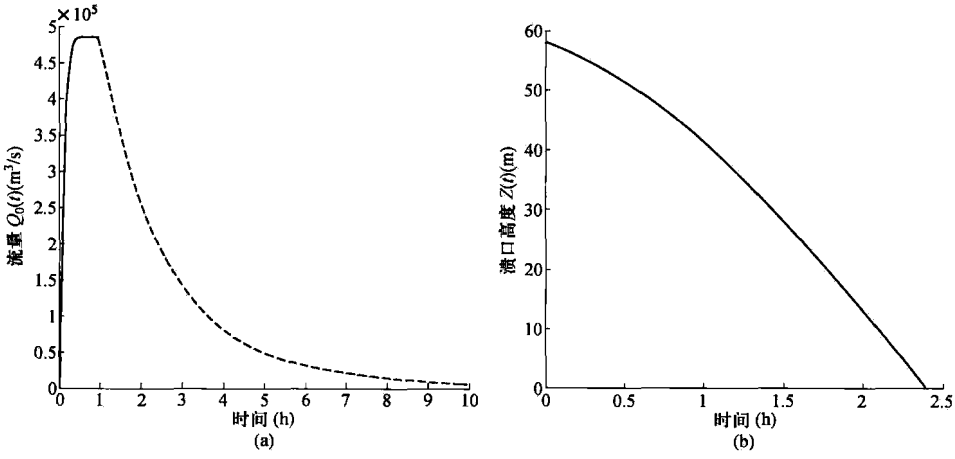


图 1-2 通过大坝的流量和溃口底部高度随时间变化的趋势

1.2.3 各段河道之间的流量模型

接下来用曼宁方程计算各段河道之间的流量. 通过运用美国国家地球物理资料中心（National Geophysical Data Center，NGDC）提供的高分辨率数据和海岸地形模型[12]，可以建立一个三维地形模型. NGDC 提供的地形数据精度在美国本土可以达到 3″，即以经（纬）度每 1°划分为 1 200 等份，网格边长小于 100 m，足以对地形进行简单的估算.

图 1-3 中（a）为墨里湖，（b）为萨卢达河，（c）为罗尔斯河，（d）为州议会大厦[9].

通过这一模型可以估计出每一段河道的容积和下游出口河道的横截面面

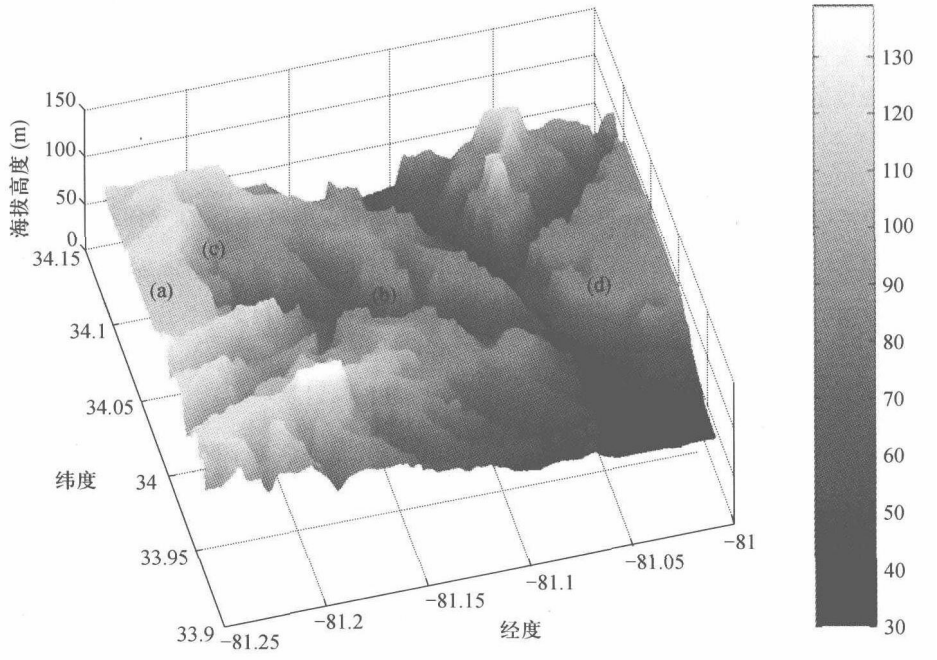


图 1-3　萨卢达河流域地形概貌

积，然后运用曼宁方程和连续方程建立一系列的一阶微分方程组，再对微分方程组求数值解，最后用萨卢达河靠近萨卢达大坝和哥伦比亚的水文观测站测得的正常流量[13]来检验模型.

对于河道分段点的确定，尽量选择相对平直、狭窄的流域. 假设洪水的流量由通过这些河道狭窄处的速度决定，则可将整个河流视为一系列的水池，水流依次流过这些水池.

一般情况下，洪水漫滩可以使用等截面河槽的流量分析方法，也就是曼宁方程[14]：

$$u=\frac{1}{n_m}\left(\frac{A}{P}\right)^{2/3}\sqrt{S}, \tag{1.10}$$

其中 u 是水流的平均速度，n_m 是反映渠道摩阻作用的系数，A 是渠道的横截面面积，P 是湿周长（水流部分横截面的周长），S 是渠道的坡度.

曼宁方程至今没有理论基础，然而大量的实验都证实了其存在性和有效性. 它的主要优势是可以通过大量的信息来估计曼宁糙率 n_m[14]. 本模型可以运用其他河流洪水漫滩的数据来估计萨卢达河的 n_m. 萨卢达河两岸基本都被植被所覆盖，洪水流经灌木丛时 n_m 大约为 0.5，流经林地时 n_m 大约为 1.514[14]. 假设萨卢达河洪水泛滥区域的地貌介于两者之间，取 $n_m=1$.

假设渠道坡度 S 等于河床的坡度，也就是说对于任意一个分段点，水深和流量与水流的方向无关. 坡度的计算方法为两点之间的高程差除以水平距离，

利用 NGDC 的数据估计每段河道流出点的坡度：

$$S_1=\frac{1}{1200},\quad S_2=\frac{1}{800},\quad S_3=\frac{1}{600},\quad S_4=\frac{1}{800}.$$

其中最后一段河道的坡度 S_4 需要谨慎地估计，因为要考虑到最坏的情况，也就是泄洪区域将包括哥伦比亚.

河道的地形数据可以用来估计以上几个数量，包括每一河段水的体积、横截面面积、湿周长以及水面高度. 由此可以建立第 i 段流域横截面面积 A_i 和湿周长 P_i 关于河水体积 V 的函数，记为 $A_i(V)$ 和 $P_i(V)$. 对于给定的渠道横截面，流量 Q 满足

$$Q=uA,$$

其中 u 是水流的平均速度. 将式（1.10）代入可得：

$$Q_i=A_iu_i=\frac{A_i}{n_m}\left(\frac{A_i}{P_i}\right)^{\frac{2}{3}}\sqrt{S_i},$$

其中横截面面积 $A_i=A_i(\rho V)$，湿周长 $P_i=P_i(\rho V)$，河水体积 $V=V(t-\zeta)$.

这里引入了两个参数 ρ 和 ζ 来校正模型，其具体取值可以由实验数据得出. ρ 表示河道的摩擦力和表面特征对水流向下游流动的阻拦作用，ζ 表示水流流经河道所需要的时间，由于每一段河道的长度不变，可以假设 ζ 为常数.

1.2.4 河道的截取和分析

利用 NGDC 提供的三维地形数据[12]建立地形模型. 按照上文的分析，选择适当的位置，用平面将河道横切分成多段. 由此可以确定每一段河道河水的体积 V_i、出口的横截面面积 A_i 和湿周长 P_i. 建立这些变量与水深的数据库，使用式（1.2）进行模拟. 第 4 段河道的地形图和河水体积、横截面面积关于水面高度的变化如图 1-4 所示.

NGDC 提供的数据精度为 0.2 m，即误差不大于 0.2 m. 误差对河水体积的估计没有明显的影响，但是横截面面积和湿周长会产生小幅度的不连续性，图 1-5 中的振动现象就是由此造成的. 如果能提供更精确的观察数据，模型的结果也会得到改善，但是相对于河水体积 V_i 的数量级 10^7 来说，精度过高的数据反而会导致计算的复杂化.

根据上文的分析，对河道分段点的选择需要满足以下的要求：

（1）每一段河道的出（入）口必须比河道的其他部分窄；

（2）每一段河道可以视为是平直、规则的；

（3）洪水流经每一段河道的时间大致相同.

通过观察，要满足上述要求，可以对河道进行如下的分段：

（1）河道分段 1：从萨卢达大坝到位于卡瑞利岛（Correly Island）附近的狭窄河道，长约 6 km；

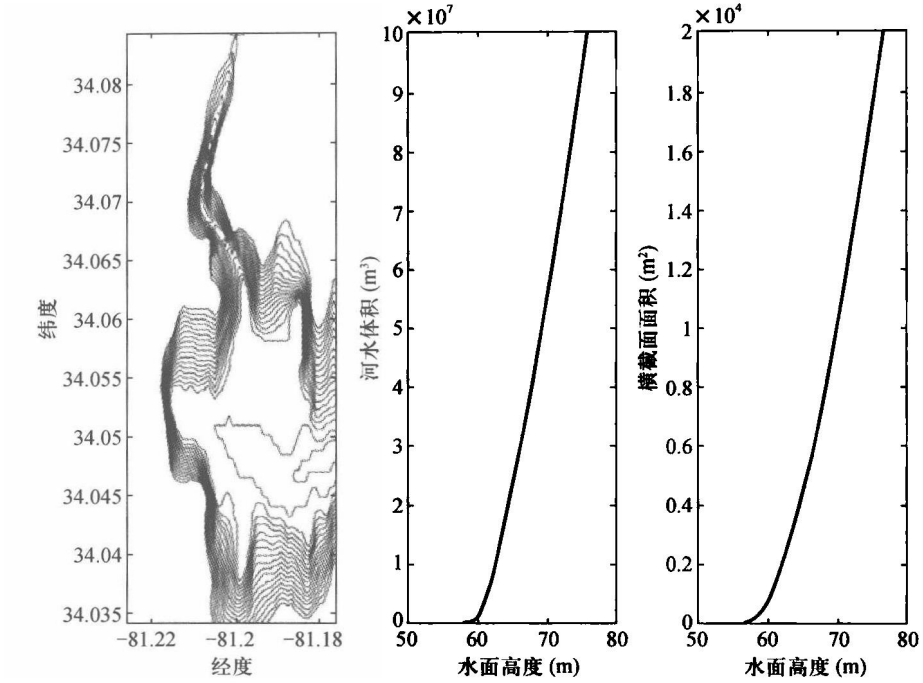


图 1-4 第 4 段河道的地形图、河水体积和出口横截面面积

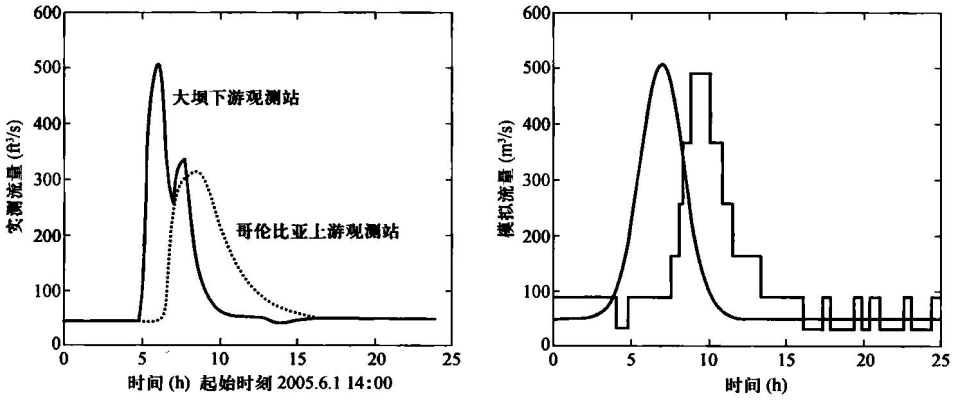


图 1-5 实际流量和模拟得到的流量

（2）河道分段 2：从卡瑞利岛到萨卢达河与 20 号州级公路的交汇处，长约 6 km；

（3）河道分段 3：从 20 号州级公路到萨卢达河与康加里河的交汇处，长约 6 km；

（4）河道分段 4：康加里河从与萨卢达河交汇处向下游延伸的 6 km 河道，包括州议会大厦.

第 1-3 段河道都完全符合给定的要求，康加里河流域的情况比较特殊，河道在此处迅速变宽，没有明显的狭窄处可以利用，而且在洪水来临时会在哥伦比亚西北方向形成漫滩，这使得第 4 段河道无法完全满足要求. 最后，选择康加里河下游一处相对狭窄且满足距离要求的位置，作为所研究河道的终点. 此外，考虑到上游的洪水，选取的河道包括河谷的很大一部分区域.

1.2.5 模型的参数估计

美国地质勘探局（United States Geological Survey，USGS）在萨卢达河设有两个水文观测站，分别位于大坝下游（北纬 34°03′03″西经 81°12′35″）和哥伦比亚上游（北纬 34°00′50″西经 81°05′17″），这两个水文观测站至少会记录 31 天的流量数据[13]. 2005 年 1 月 6 日，萨卢达大坝的水文观测站记录到了一次 30 000 m^3 的浪涌，流量在 5 个小时内由 27 m^3/s 逐渐提高到 700 m^3/s，再逐渐回落. 哥伦比亚观测站记录的水量、流量数据与之类似，在时间上晚了 1.5 小时（如图 1-5）.

首先需要调整参数 ρ，使模型计算出的大坝下游流量处于正常水平 60 m^3/s. 当 $\rho<0.1$ 时，模型的结果稳定，但是有震荡现象. 震荡可以认为是溃口造成的流速突发性变化产生的，也就是由于河道轮廓不精确造成的.

当 $\rho=0.1$ 时，如果有大量的河水进入河道，结果将不稳定. 而模型所针对的洪水正是这种大规模的水量，因此设定 $\rho=0.01$，这一数值成立的前提是，地面致密无渗透，使得停留在地表的总水量距离主河道越远将越多. ρ 的数值较小也符合它的实际意义，表示对流量的阻滞作用.

一旦模型对正常的流量产生稳定的结果，就可以引入一个类似于 1 月 6 日流量变化形式的洪水 Q_0. 接下来调整洪水流经每一段河道的时间 ζ，使得洪水达到第 3 段河道出口处的时间为 1.5 小时. 由于假设 ζ 为常数，显然有 $\zeta=-1.5/3=-0.5$. 由于利用一个流量较小的浪涌来校准用于洪水的模型，因此假设 ζ 与洪水的流量大小无关. 如果要获得更精确的模型参数，就必须利用更大规模的浪涌，由于萨卢达河历史上并没有发生过溃坝洪水，显然无法从水文观测站获取相关数据.

1.2.6 模型的结果分析

利用模型可以预测洪水缓慢流经萨卢达河谷，以及在罗尔斯河（第 1 段河道）和哥伦比亚（第 4 段河道）附近涌上海拔较高位置的过程. 模型结果如图 1-6[9]和表 1-2、表 1-3 所示，洪水在罗尔斯河附近会达到 32 m，而州议会大厦由于位于地势较高的位置，不会受到洪水的波及.

表 1-2 每一段河道洪水的最大体积和对应的海拔高度

分段河道	洪水最大体积（$\times 10^8$ m^3）	洪水最大海拔高度（m）	河流平均海拔高度（m）
1	19	87	58
2	15	79	55
3	3.5	68	50
4	2.4	68	45

表 1-3 重点关注位置的海拔高度

重点关注位置	海拔高度（m）
墨里湖	120
萨卢达河（大坝下方）	60
罗尔斯河（第 1 段河道）	55
萨卢达河（河底）	45
州议会大厦（第 4 段河道）	100

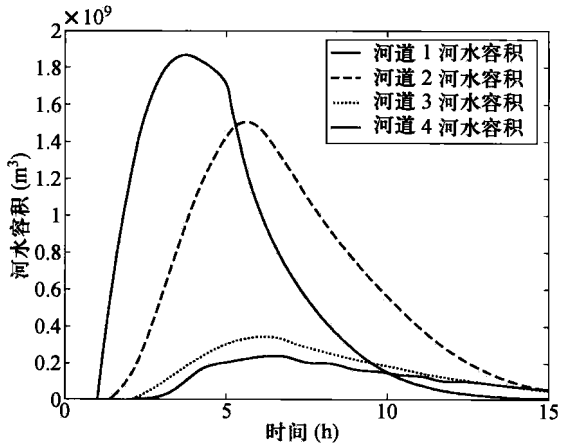


图 1-6 模型所预测的各段河道流量

经过建立模型并计算结果，可知溃坝会对下游造成严重的影响. 罗尔斯河与萨卢达河交汇处的水位会上升 32 m，显然会对罗尔斯河和附近的其他支流造成逆流. 在哥伦比亚附近，洪水会造成水位上升 23 m，不足以影响位于小山上的州议会大厦，但是对哥伦比亚西部低洼地带会造成一定的破坏.

1.3 瞬间溃坝模型

瞬间溃坝一般发生在刚性坝，例如混凝土坝. 对于土石坝，如果地震的强度足够大，有可能出现两种情况：

（1）在第一时间形成较大的溃口，使大坝在极短的时间内迅速溃决；

（2）地震直接导致大坝液化. 这两种情况都属于瞬间溃坝模型.

实际上，于2005年完工的萨卢达大坝备用饯道的修建目的，正是为了防止萨卢达大坝受地震影响短时间内液化，从而造成严重的溃坝洪水.

在实际情况中，严格意义上的瞬间溃坝几乎不能存在，这一方法的意义有两点：一是在大坝的设计和建造过程中，需要考虑瞬间溃决这种极端情况，二是当数据不充足或者条件不允许的情况下，对溃坝洪水进行简化计算.

瞬间溃坝模型与逐渐溃坝模型类似，仍然将问题分为两个部分，模拟大坝溃决和计算下游洪水演进. 其中大坝溃决由于时间较短，可以认为是瞬间完成，不需要做太多的考虑，因此瞬间溃坝模型的主要工作是运用更加精确的模型来计算洪水的演进过程[9].

这种情况需要运用描述水道中渐变不恒定水流运动规律的偏微分方程组——圣维南方程组（Saint-Venant equations）. 洪峰可以看做是圣维南方程组稳定解上的正弦扰动，然后运用动态波动模型[15]确定洪水的漫延过程.

1.3.1 模型的建立

将萨卢达河谷看做是具有矩形横截面的平直渠道，从USGS可以获得规模较小的浪涌数据，利用其估计河道对水流的摩擦系数.

能够造成大坝短时间内完全冲毁的洪水非常汹涌且不稳定，水力学中称其为溃坝洪水波. 大坝的突然垮塌导致水位急剧下降，进而产生倒退的波前（上游水位高于下游水位），也就是反向的浪涌[14]. 这是因为在正常情况下，大坝将水库和河道的水分为有高度差的两个部分，通过交叉点的水流相对缓慢，当溃坝发生时，大量的水在交叉点奔涌而出，通过交叉点的流速就出现了急剧的变化. 由于溃坝形成的洪水具有巨大的冲击力，水库和河道的水无法直接融合成一个浪涌，于是就会产生第二个浪涌，其方向与第一个相反，是正向的.

由正向和反向浪涌形成的水波传播过程，可以用圣维南方程组来描述[16]. 这是一个一阶双曲型偏微分方程组. 圣维南方程组由法国科学家圣维南于1871年提出，用于描述水道中渐变不恒定水流的运动规律：

$$\begin{cases} \dfrac{1}{g}\dfrac{\partial u}{\partial t}+\dfrac{u}{g}\dfrac{\partial u}{\partial x}+\dfrac{\partial d}{\partial x}+(S_f-S_0)=0, \\ \dfrac{\partial d}{\partial t}+\dfrac{\partial du}{\partial x}=\dfrac{\partial d}{\partial t}+d\dfrac{\partial u}{\partial x}+u\dfrac{\partial d}{\partial x}=0. \end{cases} \tag{1.11}$$

其中 u 是水波的平均速度，d 是水深，x 是传播方向，t 是时间，g 是重力加速度，S_f 是摩阻坡度，S_0 是河床坡度.

第一个方程是反映动量守恒律的运动方程. 其中第一部分$\dfrac{1}{g}\dfrac{\partial u}{\partial t}$描述水波处于某个指定位置的局部加速度；第二部分$\dfrac{u}{g}\dfrac{\partial u}{\partial x}$描述由于流速的空间分布不均匀所引起的对流加速度，以上两部分合称为惯性项；第三部分$\dfrac{\partial d}{\partial x}$水深造成的压差，称为水面比降；第四部分 (S_f-S_0) 描述水流内部及边界的摩阻损失，方程给出了重力与压力的联合作用使水流克服惯性力和摩阻引起的能量损失而获得的加速度.

第二个方程是反映质量守恒律的连续方程. 其中第一部分$\dfrac{\partial d}{\partial t}$描述蓄水量的变化，第二部分$\dfrac{\partial du}{\partial x}$描述沿途水量的变化.

圣维南方程组有如下基本假设[16]：

（1）水流是一维的，运动只发生在传播方向上；

（2）不考虑水流垂直方向的交换和垂直加速度，从而可假设水压力呈静水压力分布，即与水深成正比；

（3）水是不能压缩的；

（4）水流的摩阻效果与匀速水流的情况相同，即 $S_f=S_0$.

模型的重点是描述洪水波的递减过程. 假设萨卢达大坝水库中的所有水以一个巨大的浪涌的形式释放出来. 洪峰流量被减弱之后，最后的结果与洪峰初始流量无关，这使得模型对于任意规模的洪水波都可以粗略计算出结果.

1.3.2　模型的求解

在圣维南方程组（1.11）稳定解[17]的基础上，运用正弦扰动来模拟溃坝的效果. 运用频谱分析，可得在点 X 处洪峰流量为

$$q_p=q_{P_0}\exp\left(\frac{-\alpha X}{L_0}\right),$$

其中 $\alpha=\dfrac{2\pi}{n_m^2}\left(\dfrac{L_0d_0B}{V_w}\right)\left[\zeta-\left(\dfrac{C-A}{2}\right)^{\frac{1}{2}}\right]$，$A=\dfrac{1}{F_0^2}-\zeta^2$，$C=(A^2+\zeta^2)^{\frac{1}{2}}$，$\zeta=\dfrac{1}{\sigma F_0^2}$，$\sigma=\left(\dfrac{2\pi}{L}\right)L_0$，$F_0=\dfrac{u_0}{\sqrt{gd_0}}$.

这里的 F_0 为弗劳德数；u_0 为稳定平衡时的平均速度；L 为扰动波长；L_0 为参考河道的长度；d_0 为稳定平衡时的水流深度；B 为河段平均宽度；V_w 为水库蓄水量；g 为重力加速度；n_m 为曼宁糙率. 由此可得单位宽度的流量方程（速度×深度）为

$$q=\frac{N}{N+1}V_{max}d, \tag{1.12}$$

其中 $N=0.4\sqrt{8/f}$，f 是达梳擦因数（Darcy friction factor），V_{max} 是水库最大容积，d 是水流深度.

利用图 1-5 中的经验数据来计算洪水波的速度，并且估计萨卢达河谷的达梳擦因数. 根据（1.12）式还可以估计洪水波的深度.

1.3.3 模型的结果分析

通过模型的估计可以得到大坝两侧的水深（水库深度和萨卢达河在靠近大坝处的水深），以及萨卢达河流域的容积，由此可以得到溃坝波前深度的函数，自变量为与大坝的距离. 模型结果如图 1-7 所示，溃坝波的初始深度大约为 65 m，以指数形式递减，在距离大坝 20 km 处大约降低为 4 m，这个距离与萨卢达大坝到州议会大厦的距离近似相等.

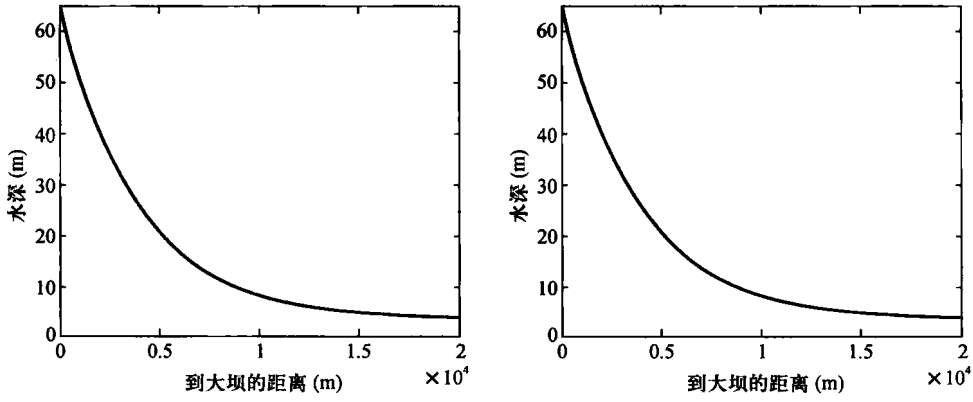


图 1-7 萨卢达河上游和全流域的洪水波最大深度

州议会大厦所在的小山大约比萨卢达河高 50 m，因此根据预测结果州议会大厦受洪水波影响的可能性极小. 事实上，出于简化模型考虑，假设河道是矩形的并且宽度和河床质地不变，而在现实中，萨卢达河的宽度有很大的变化，河床的土质也不尽相同. 这些因素会导致离大坝越远，模型计算出的洪水波深度越高于实际情况，也就是说现实中州议会大厦受洪水影响的可能性还会进一步下降.

对于罗尔斯河，模型结果显示洪峰高度大约为 40 m. 显然瞬间溃坝造成的洪水在罗尔斯河流域会形成明显的倒流.

1.4 差分形式的圣维南模型

圣维南方程组可以用于描述明渠渐变不恒定水流的运动规律，广泛用于洪水估计. 该方程组属于偏微分方程，通常不能获得解析解，而要用数值方法求解. 将偏微分方程组离散化，简化为差分方程组，是一种常用的数值求解方法.

这与直接运用瞬间溃坝模型中的圣维南方程组及其稳定解不同，建立相对完善和精确的圣维南模型包括如下几个步骤[18]：

(1) 针对现实问题做出一系列的简化假设，使其符合圣维南方程组的基本假设（参见1.3.1).

(2) 建立模型估计溃坝洪水的初始流量，作为方程组的初值.

(3) 建立描述洪水漫滩过程的圣维南方程组，并转化为差分方程组.

(4) 测量泛洪区与主河道宽度的比值，用以修正模型参数，提高结果精度.

1.4.1 问题分析与假设

根据问题的要求，假设一场大地震造成了萨卢达大坝损坏. 一般情况下土石坝不会瞬间完全溃决，而是先开始渗漏. 随着时间的推移，水流不断冲刷大坝造成溃口扩大，直到水流和大坝达到平衡或者水库放空. 根据初始溃口的形状和大坝构造的不同，达到平衡状态需要从数分钟到数小时不等的时间[19]. 最终的溃口宽度通常介于坝高的1/2到3倍之间[19]，萨卢达大坝高61 m，长约2 400 m，长约为高度的40倍，这意味着溃口宽度将远小于大坝长度.

在大坝下游，溃坝洪水冲出河道形成泛洪区. 对于较小的支流，例如罗尔斯河，洪水会造成倒灌. 随着距离的推移，泛洪区的宽度会逐渐减小，直至原有河道足以容纳洪水. 在建立模型之前先给出如下的假设：

(1) 忽略余震的影响. 地震包括主震和余震，由于大坝在主震发生时已经受损，则不考虑余震对溃口的扩大作用.

(2) 仅考虑大坝溃口. 地震会影响大坝、大坝内的水和地形. 对水的影响仅发生在地形损害时. 地震有可能会严重影响地形，但是其发生的概率和破坏程度难以预计，为简化问题，这里忽略不计.

(3) 水库的边缘是垂直的，水库的底是平坦的.

(4) 大坝溃口为矩形，溃坝后的沙土与水流相比可以忽略.

(5) 萨卢达河河道宽度不变. 从地图可以看出，11.4 km之后的河道略微变宽，但是为了简化模型，假设宽度不变.

(6) 河床的海拔高度均匀下降，且河道是直的. 弯曲的河道能够减缓水

流速度，但萨卢达河整体方向没有太大的变化，可以把它当做直线.

1.4.2 坝体溃口模型

首先建立一个模型来仿真地震引起的溃口对水库和大坝的影响. 建模的目的是计算在任意时刻通过溃口的水流量和流速，已知信息包括水库容量、湖水表面积和溃口大小.

根据模型的假设，将溃口视为矩形. 假设水从溃口涌出并且能量不变，即势能转换为动能，则有如下方程：

$$\frac{1}{2}mu^2=mgh,$$

将其化简得 $u=\sqrt{2gh}$，其中 u 是水流速，m 是水的质量，g 是重力加速度，h 是水下降的高度，即湖水和溃口底部的高度差.

假设水流始终保持最大速度，可以得到相对的流速方程：

$$u=\sqrt{2g(H-Z)},$$

其中 H 是湖水的高度，Z 是溃口底部的高度.

每个时间段流出溃口的水量，等于溃口横截面面积乘以水流速度，再乘以时间段长度，即

$$v=b(H-Z)ut,$$

其中 v 是水的体积，b 是溃口宽度，t 是时间.

根据假设，水库边缘垂直且底部平坦，所以水面面积是不随水位高度改变的. 湖水的高度和容量、面积关系如下：

$$H=\frac{V}{A_s}, \tag{1.13}$$

其中 V 是库容水体的体积，A_s 是水面面积. 如果放宽假设，把湖当做浅的锥形物，式 (1.13) 中水面面积可以与水量关联，即 $A_s(V)$.

根据假设，溃口形状不变，始终为矩形，但是溃口的高度和宽度会随着水流冲刷而增加，其增量可以表示为通过溃口的水量和流速的函数. 由此可以仿真从溃口冲出的水流造成的大坝侵蚀过程.

1.4.3 圣维南模型

圣维南方程组被广泛地用于模拟水灾的影响，研究对象是非恒定开放河道的水流变化，并考虑了速度、地形、阻力的变化[20,21]. 这使得圣维南方程更适于简单模型，尤其是阻力因素很重要时.

圣维南方程由两个方程构成，分别是水的质量守恒方程

$$\eta\frac{\partial d}{\partial t}+d\frac{\partial u}{\partial x}+u\frac{\partial d}{\partial x}=0,$$

和线性动量方程

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+g\left(\frac{\partial d}{\partial x}+S_f-S_0\right)=0,$$

其中 d 是水的高度，x 是离河的距离，t 是时间，u 是流速，S_0 是河床坡度，η 是泛洪区相对宽度，S_f 是摩阻坡度.

摩阻坡度代表了流水受到的阻力，与速度和水流半径有关，即

$$S_f=n_m^2 k u^2 R^{-\frac{3}{4}},$$

其中 n_m 是糙率，R 是水流半径，由 $R=\dfrac{Bd}{B+2d}$ 确定，其中 B 是河道宽度，$k=1\ s^2/m^{2/3}$，是一个恒定参数.

对于参数 η，模型假设泛洪区周围有高的物体，例如树林、建筑物等，这意味着在所考虑的区域内溢出的水流可以忽略. 但是由于泛洪区的存在使得水面高度下降，所以 $\dfrac{\partial d}{\partial t}$ 由泛洪区宽度 η 和河道宽来衡量. 这样，当水位上升时，水面的实际高度受到 η 的作用而减弱，也就是说有部分水进入泛洪区. 由此，可以通过测量河道各点处泛洪区的宽度来估计 η.

为了使模型可以得到数值解，使用差分方程组来近似偏微分方程组（圣维南方程组）. 对于流体力学的偏微分方程组，需要特别注意算法的稳定性，因此使用 Lax-Wendroff 差分法[22]

$$u_j^{n+1}=u_j^n+\frac{1}{2}\lambda(u_{j+1}^n-u_{j-1}^n)-\frac{1}{2}\lambda^2(u_{j+1}^n-2u_j^n+u_{j-1}^n).$$

其中上标是时间，下标是空间，λ 是时间和空间的比例（模型限制距离和时间段，因此每一段为1），第二项表示减弱效果，代表点与点间的区别.

初步检验发现模型对糙率敏感，这可能是因为糙率是模型中唯一的阻滞因素. 对于普通的大河而言，糙率一般为 0.03，按此计算，水流会出现堆积，最终导致模型无法计算. 但是根据基本假设，河水只在河道中流动，涝地部分的河水假设静止不动. 因此采用一个较小的糙率是合理的，不妨取 $n_m=0.01$.

进一步检验还发现模型对高速流不稳定，这是因为圣维南方程组本质上是针对稳定流的. 通过分段计算水流的平均高度，结果显示并没有太大的影响.

事实上，利用上面所建立的模型可以直接使用圣维南方程组来解决洪水对罗尔斯河的倒灌问题. 假设罗尔斯河与萨卢达河的交汇处存在一个虚拟的大坝，其水位高度为溃坝洪峰到达的高度，调整其他参数使其符合罗尔斯河的实际情况，就可以计算出洪水倒灌的距离. 但是这一方法存在一个难以解决的问题，由于罗尔斯河的流向与萨卢达河几乎垂直，洪峰进入罗尔斯河的初速度难以估计. 为此，需要对模型进行改进，首先通过萨卢达河的圣维南方程组计算得到罗尔斯河出口处的洪峰水位高度，然后利用地质数据寻找上游符合这

一高度的位置. 这样做极大地简化了问题，并且符合模型的假设，即水只在河道中流动，泛洪区的水是瞬时注满的，也就是将罗尔斯河看做是泛洪区的一部分.

1.4.4 模型的求解

要求解上面的模型，需要使用如下的一些参数值：

① 重力加速度 $g=9.8\ m/s^2$；

② 湖水初始水位高度 $H_0=60\ m$；

③ 湖水初始容积 $v_0=3\times10^9\ m^3$；

④ 湖面面积（假设边缘垂直）$A_{s0}=5\times10^7\ m^2$；

⑤ 萨卢达河总长度 $L_r=16\ 200\ m$；

⑥ 与萨卢达大坝的坝基相比，位于大坝正下方的河床高度 $h_u=0$；

⑦ 流入康加里河处的河床高度 $h_d=-10\ m$；

⑧ 河流正常情况下的初始水深 $h_{w0}=1.2\ m$.

为了简化模型，我们假设河流的边缘是垂直的. 在圣维南模型中，水体容积和水位高度更为重要，需要使用真实数据，因此在模型的求解计算中所使用的墨里湖表面积小于实际数据.

由于是洪水，水流不会全部停留在河床，一定会形成泛洪区. 通过查看地形图，可以建立准确的河道和周围泛洪区比例. 河流海拔高度大约 52 m，河道两侧可能的泛洪区海拔高度 61 m（再往外扩张会遇到山地），测量得到宽度大约是 600 m，假设河道与山地之间全是泛洪区，实际宽度随着测量点距大坝的距离线性变化. 按照这个假设建立的模型要比直接假设泛洪区宽度不变要精确得多.

萨卢达大坝宽约 800 m，高约 60 m，理论上讲，溃口可能出现在大坝的任意位置，并且可能是不同形状和大小. 以下分 4 种情况分别讨论.

（1）无溃口

没有出现溃口，就相当于溃口宽度为 0，高度为 0.

首先测试没有溃口的情况，水位基本保持不变，仅有用于水力发电的水进入河道. 由于没有泛洪区，因此除了支流的影响，水面宽度基本保持不变. 如图 1-8 所示，其中两个峰值位置都是支流，第一个是罗尔斯河.

（2）常见溃口

常见溃口，即溃口宽度为 800 m，高度为 10 m 的情况.

最常见的情况是地震造成土质大坝一个宽而浅的溃口. 在这种情况下，洪峰 1.1 h 后到达罗尔斯河，水位相对高度为 7.1 m. 通过观察地图，罗尔斯河倒灌的距离为 2.4 km. 洪峰在 7.5 h 后到达哥伦比亚，水位相对高度为4.5 m，由于国会大厦在 50 m 高处将不会有水灾的危险.

(3) 深溃口

深溃口，即溃口宽度为 133 m，高度为 60 m 的情况.

为了研究溃口形状的影响，考虑一个与常见溃口横截面面积相等但是形状不同的溃口. 因为溃口较深，水速会更高，而且由于湖水的最终高度更低，流出的水量也会更大.

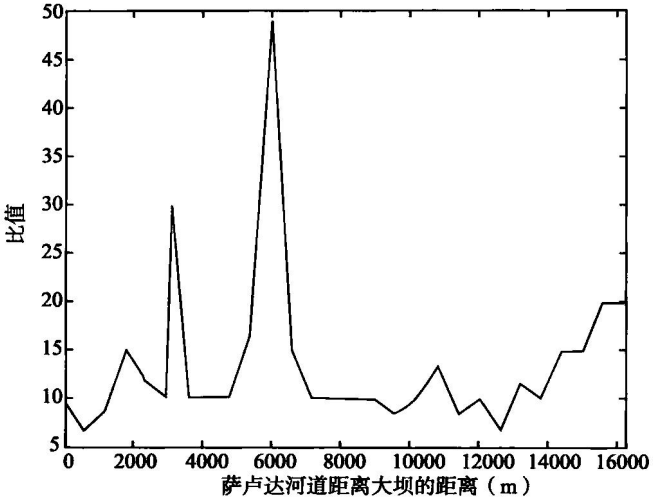


图 1-8 河道宽度与泛洪区宽度的比值

结果显示，洪峰在 1.4 h 后到达罗尔斯河，水位高度达到 9.11 m，倒灌距离扩大到 3 km. 洪峰 7 h 后到达哥伦比亚，水位高度达到 6.23 m. 图 1-9[18]

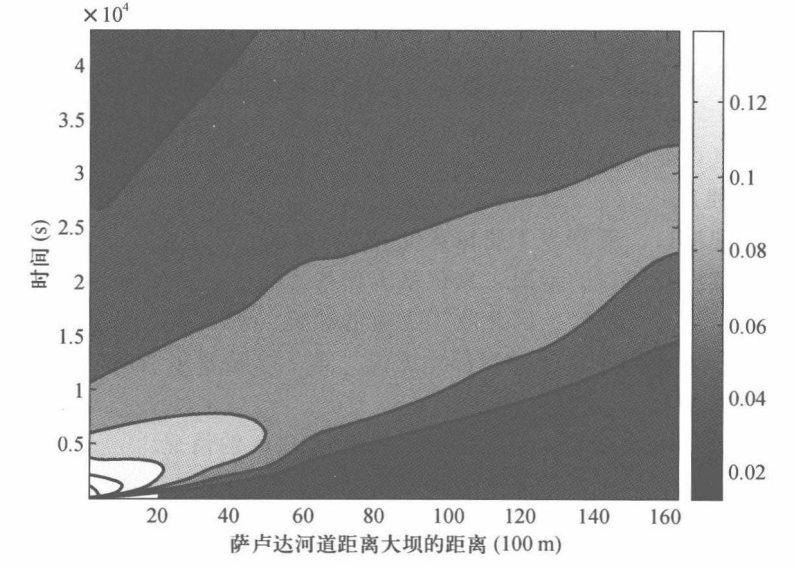


图 1-9 水位等高线图

是对第三种溃口模拟得到的水位高度等高线图，单位为 100 m，其中横轴是观测点距大坝的距离，纵轴是时间.

图 1-10 是罗尔斯河出口处的水位高度，其中横轴是时间，纵轴是水位高度.

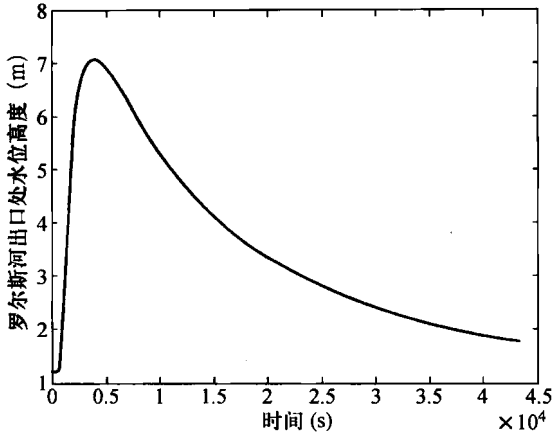


图 1-10 罗尔斯河出口处水位高度

图 1-11 是萨卢达河流入康加里河处的水位高度，其中横轴是时间，纵轴是水位高度.

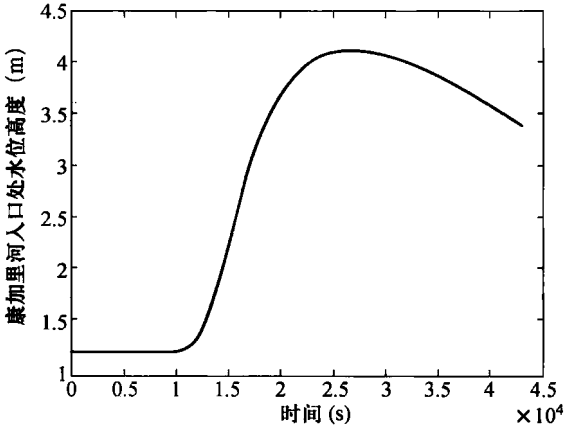


图 1-11 萨卢达河流入康加里河处的水位高度

(4) 最大溃口

最大溃口，是指宽度为 800 m，高度为 60 m 的情况.

假如大坝全部损毁，模型和假设都没有考虑这种情况. 通过将模拟的步长放大，并使用 3 倍的模拟时间，可以得到洪峰对罗尔斯河和哥伦比亚的影响. 但是，由于水位过高，初始设定的泛洪区域无法容纳全部洪水，也就是泛洪区宽度 η 不再有效. 这意味着实际洪峰水位会低于模型的结果.

在大坝完全损坏的情况下，洪峰在0.4 h后到达罗尔斯河，水位高度为34.35 m，倒灌距离为5 km；4.5 h后洪峰到达哥伦比亚，水位高度达到17 m，但是国会大厦（50 m）仍然安全.

通过分析，可以得到以下结论：国会大厦即使在最坏的情况下，仍然是安全的. 如果能建立预警系统，4.5 h的时间也足够疏散哥伦比亚西部低洼地带的居民.

哥伦比亚主要位于山上，预测的洪峰（水位在4 m到17 m）将影响距离河流较近的低洼地区. 但是上游的地区，比如罗尔斯河在半小时内将会达到7 m到34 m，并将会影响罗尔斯河上游2.5 km到5 km的地区，所以早期预警是很重要的.

1.5 离散网格模型

洪水估计问题的离散网格模型是一种典型的连续问题离散化的方法，其特点在于，将从大坝到哥伦比亚的萨卢达河流域分成一系列的网格，每一个网格都包含海拔高度、水位高度和水体容积等信息，将连续的水体流动过程离散化为水体在网格之间的转移过程. 离散网格模型解决问题的过程同样分为溃坝和洪水演进两步.

在溃坝模型中，将水库等效为一个直立圆锥体，在常见的两种溃坝模式的基础上，考虑了四种溃坝情况：瞬间溃坝、逐渐溃坝、管涌溃坝和漫顶溃坝.

在洪水演进模型中，将河道和泛洪区划分为210 m×210 m的网格，从流体力学和直观认知的角度出发，通过假设和简单的力学、几何学推导，可以分别建立作用力模型和下坡模型，并通过模拟结果来分析实际意义[23].

1.5.1 问题的分析

水库的表面面积一般会随着水位高度的降低而减少，这是因为自然形成的水库边缘往往是向内倾斜的. 因此可以将水库看做是一个底部平坦的锥体，进一步简化可以等效为一个直立的圆锥体.

在溃坝模型中，瞬间溃坝和逐渐溃坝是根据溃坝过程的时间长短来分类，在上文中已有介绍，不再重复. 而管涌和漫顶则是大坝溃决仅有的两种表现形式，也就是说后两种溃坝情况是根据溃坝特征进行分类. 综合来看，从两个不同的角度对溃坝进行分类建模，更有利于得到符合实际的结论.

管涌是指坝身或坝基内的土壤颗粒被渗流带走的现象[6]. 当管涌发生时，水面出现翻花，随着持续时间延长，险情不断恶化，使大坝地基土壤骨架破坏，孔道扩大，基土被淘空，最终引起溃坝. 漫顶是指水位上涨，超过坝体高度，造成水流直接越过坝体的现象. 土石坝的漫顶破坏过程主要包括水流冲蚀

引起的连续纵向下切和横向冲蚀，以及溃口边坡失稳坍塌引起的间歇横向扩展[24]. 为了简化模型，仅考虑连续侵蚀过程. 我国有记录的大坝溃决事故中，管涌和漫顶基本上各占一半[24].

在此基础上，结合直立圆锥体的水库模型和四种溃坝模式，可以分别建立模型，计算通过坝体的流量.

洪水演进过程从宏观上看受多种因素的影响，至今没有一个精确且易于计算的方法，但是从微观上看，可以简单地看做在一小段时间里，水从一小片区域转移到另一小片区域的过程，这一过程仅受水位高度、地形海拔高度和初始流速的影响. 因此考虑将整个萨卢达河流域（包括泛洪区）划分为一系列的网格，研究相邻网格之间的水体运动规律，以此建立模型，运用计算机进行仿真计算得到总的结果.

1.5.2 模型的假设与符号说明

1. 模型的假设

① 湖中的正常水位在大坝的溃口之上；

② 河道不会发生季节性的变化，洪水经过时不会改变地形；

③ 土质大坝的材料始终是均匀的，没有人为因素造成大坝损坏；

④ 缺口的宽度大于水坝的高度并且小于高度的三倍；

⑤ 忽略桥梁等建筑对水流的阻力；

⑥ 相对于溃坝造成的洪水，忽略其他源头水流的流入.

2. 符号说明

$A_{s0}=2.02\times10^8$ 表示墨里湖的初始水面面积（单位：m^2）；

$V_w=2.714\times10^9$ 表示正常情况下墨里湖蓄水量（单位：m^3）；

$R=8.019\times10^3$ 表示湖面半径（单位：m）；

Z 表示溃口底部距湖底的距离（单位：m）；

A_z 表示溃口底部对应的水面面积（单位：m^2）；

$E_l=110$ 表示湖面初始海拔高度（单位：m）；

E_b 表示溃口底部海拔高度（单位：m）；

Q_i 表示第 i 种情况下水流通过大坝的流量（单位：m^3/s），$i=1, 2, 3, 4$ 分别表示瞬间溃坝、逐渐溃坝、管涌溃坝和漫顶溃坝的情况；

Q_{max} 表示溃坝过程中水流通过大坝的最大流量（单位：m^3/s）；

t_i 表示第 i 种情况下通过大坝的水流停止时刻（单位：s），$i=1, 2, 3, 4$ 分别表示瞬间溃坝、逐渐溃坝、管涌溃坝和漫顶溃坝的情况；

t_b 表示溃口从出现到稳定所需要的时间（单位：s）；

ΔV 表示墨里湖流出的洪水总体积.

1.5.3　四种溃坝模型

将墨里湖等效为直立圆锥体，如图 1-12 所示.

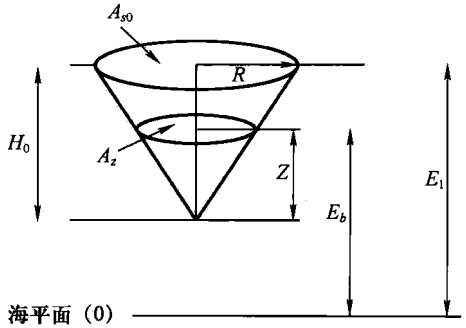


图 1-12　用直立圆锥体等效的墨里湖

根据圆锥体体积公式，可得初始水位高度等效值 $H_0=\dfrac{3V_w}{A_{s0}}\approx 40$ (m). 同理可得 $\Delta V=V_w-\dfrac{1}{3}A_zZ$，其中 $A_z=\pi\left(\dfrac{RZ}{H_0}\right)^2$，则有

$$\Delta V=V_w-\frac{1}{3}\pi Z\left(\frac{RZ}{H_0}\right)^2. \tag{1.14}$$

接下来利用相似的参数将每种类型的溃坝描述为流量随时间变化的函数，则可以分别建立四种溃坝模型.

1. 瞬间溃坝模型

在瞬间溃坝情况下，溃口形成时刻的流量最大，此后随着水位高度的降低，流量逐渐减小. 因此流量关于时间的函数可以简化为一个线性模型，参数是缺口深度和水流的最大流量[19].

$$Q_1(t)=\begin{cases}\dfrac{Q_{\max}(t-t_1)}{t_1}, & t\leqslant t_1;\\ 0, & t>t_1.\end{cases}$$

其中 $t_1=\dfrac{2\Delta V}{Q_{\max}}$，$\Delta V$ 的计算见式 (1.14).

取 $Z=20$，$Q_{\max}=30\ 000$，通过溃口的洪水流量如图 1-13 所示. 即瞬间溃坝的情况洪水流量随时间呈线性关系下降.

2. 逐渐溃坝模型

在逐渐溃坝情况下，溃坝洪水有一半会用来冲蚀大坝[19]，冲蚀完成之后，流量才会达到峰值. 结合瞬间溃坝模型，流量与时间成等比例关系，不妨假设 $t_b=\dfrac{1}{2}t_2$，则可以利用一个等腰三角模型来描述流量的变化，即有

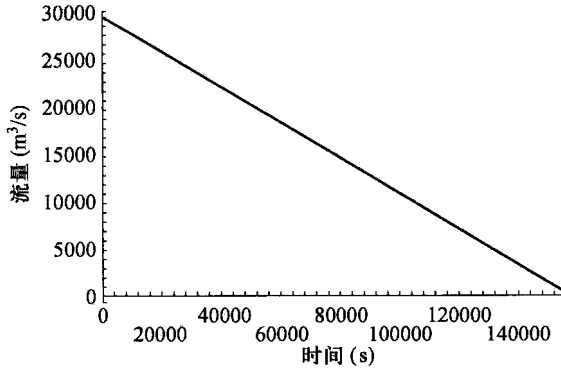


图 1-13 瞬间溃坝的流量随时间变化的关系

$$Q_2(t)=\begin{cases}\dfrac{2Q_{\max}t}{t_2}, & t\leqslant\dfrac{1}{2}t_2;\\ \dfrac{2(t_2-t)Q_{\max}}{t_2}, & \dfrac{1}{2}t_2<t<t_2;\\ 0, & t\geqslant t_2.\end{cases}$$

其中 $t_2=\dfrac{2\Delta V}{Q_{\max}}$，$\Delta V$ 的计算见式 (1.14).

取 $Z=20$，$Q_{\max}=30\ 000$，通过溃口的洪水流量如图 1-14 所示. 即结果显示，随着溃口的逐步扩大，洪水流量随时间成线性形式增加达到最大，然后按线性关系减小.

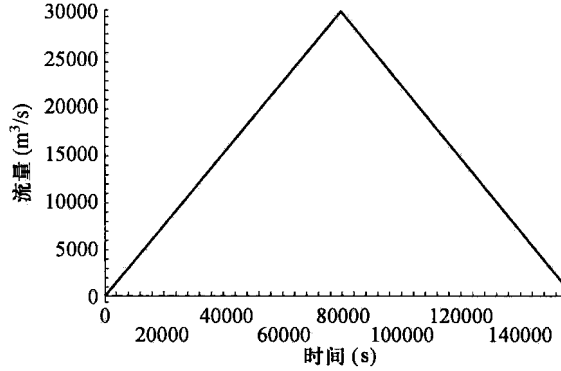


图 1-14 逐渐溃坝的水流量随时间变化的关系

3. 管涌溃坝模型

对于管涌造成大坝损毁的情况，是指从大坝迎水面的中部开始，溃口逐渐扩大直到上方的坝体全部崩溃. 随着溃口的扩大，流量以指数形式增长，并且其峰值出现在溃口上方的坝体完全崩溃时，因此可以利用指数衰减函数来描述这个模型[25].

这一模型有两个要点：一是流量的峰值应该出现在溃口达到稳定的时刻 t_b；二是 t_3 时刻的流量应该小于最大流量的 1%（因为 t_3 以后流量为 0），需要在模型建立和系数确定上加以控制.

$$Q_3(t)=\begin{cases}(Q_{\max}+1)^{\frac{t}{t_b}}-1, & t\leqslant t_b;\\ Q_{\max}\mathrm{e}^{\frac{5(t-t_b)}{t-t_3}}, & t_b<t<t_3;\\ 0, & t\geqslant t_3.\end{cases}$$

其中 $t_3=\Delta V-t_b\left(\dfrac{5\left(\dfrac{2+Q_{\max}}{\ln(Q_{\max}+1)}-1\right)}{Q_{\max}(1-\mathrm{e}^{-5})}+1\right)$.

在这时取 $Z=20$，$Q_{\max}=30\ 000$，$t_b=50\ 000$，通过溃口的洪水流量如图 1-15 所示. t_b 与 t_3 相差 4 个数量级，在溃口形成的过程中流量的变化如图 1-16 所示，表现出了溃口形成之后流量随时间的变化情况，给出短时间内的图形.

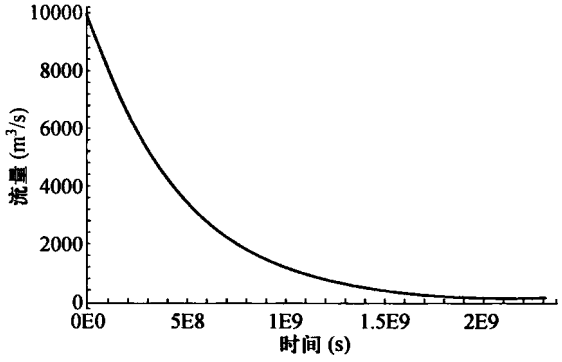


图 1-15　管涌溃坝的流量随时间的变化关系

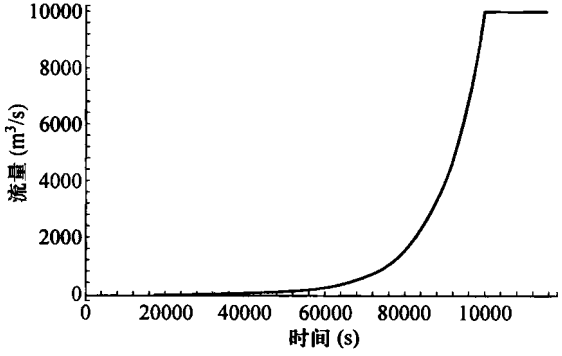


图 1-16　管涌溃坝开始发生时的流量随时间变化的关系

4. 漫顶溃坝模型

在漫顶溃坝过程中，洪水是从溃口上方流过，自上而下侵蚀大坝. 流量的增长过程类似于管涌溃坝，可以采用抛物线来近似表示；在溃口稳定以后，下降的过程类似于瞬间溃坝，则可用直线来近似表示，即有

$$Q_4(t)=\begin{cases}Q_{\max}+15(t^2+2t\cdot t_b-t_b^2)\times10^{-6}, & t\leqslant t_b;\\ \dfrac{Q_{\max}(t-t_4)}{t_b-t_4}, & t_b<t<t_4;\\ 0, & t\geqslant t_4.\end{cases}$$

其中 $t_4=\dfrac{2(\Delta V+5\times10^{-6}t_b^3-t_bQ_{\max})}{Q_{\max}}+t_b$.

这里取 $Z=20$，$Q_{\max}=30\ 000$，$t_b=30\ 000$，则通过溃口的洪水流量与时间的关系如图 1-17 所示. 洪水流量随时间先按抛物线增加，然后按线性关系减小.

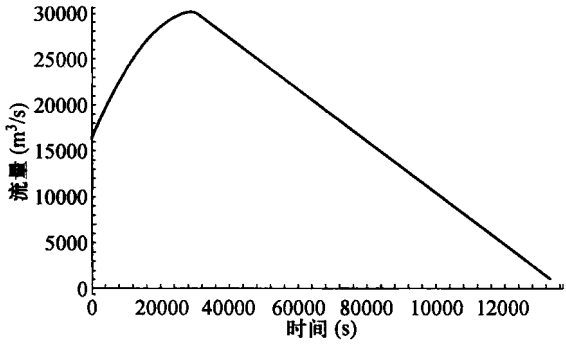


图 1-17 漫顶溃坝的流量随时间的变化关系

1.5.4 下游水流模型

利用离散方式构造两个模型来描述下游地区的洪水演进过程：①水流作用力模型基于伯努利方程（Bernoulli equation）进行物理学分析，②下坡模型可使用更简单、直观的方式模拟水流.

要建立这两个模型，首先要将萨卢达大坝周围的地区分为一系列离散的正方形网格，每个网格覆盖 30 m×30 m 的面积，网格具有固定的海拔高度并且包含一定体积的水（由网格内水的平均深度决定），海拔数据来自美国地质勘探局[13]. 由于原始数据精度和简化计算两方面的原因，将 7×7 个相邻网格作为整体给出海拔高度. 两个模型的主体都是计算水流在网格之间的传播，但在计算相邻网格之间单位时间内所交换水量的方式上有所区别.

1. 作用力模型

这个模型对网格中的水流进行了力学分析. 每一个网格都有海拔高度 e、

平均水深 d 和平均速度 v（包括横向和纵向）. 假设作用于一个网格的力仅由两个因素决定：周围紧邻的 4 个网格给予它的压力和促使水向低处流动的重力.

模型的主要原理包括两个方面：

（1）网格之间水体流动的体积正比于相邻 4 个网格的压力差.

（2）网格之间的压力差正比于平均水深的差.

水可以看做是不可压缩、没有黏滞性的流体，即理想流体. 由于对流域进行网格划分，每一个网格内流速的变化可以忽略，因此水的流动可以看做是定常流动. 因此网格中或网格间的水流过程符合伯努利方程的两个基本条件，有

$$p+\rho g d+\frac{1}{2}\rho v^2=c, \tag{1.15}$$

其中 p 是水压力，ρ 是水密度，g 是重力加速度，d 是平均水深，v 是水流速度，c 是常数. 式（1.15）中的三项分别表示单位体积流体的压力能、重力能和动能，即在沿流线运动过程中总能量是守恒的.

如图 1-18 和图 1-19[23] 所示，假定平均压力为网格中一半水深处的压力，记为

$$p=\frac{1}{2}\rho g d. \tag{1.16}$$

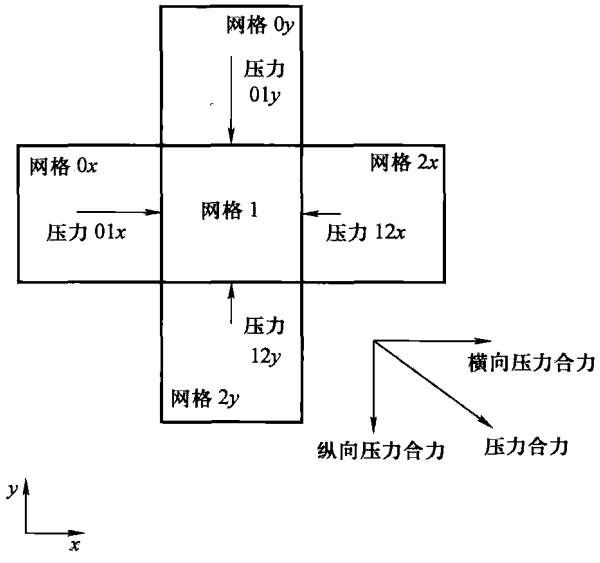


图 1-18　作用在网格上的压力

考虑横向的流动过程. 假设水深的变化是线性的，两个相邻网格之间的压力作用面积等于它们的平均水深乘以网格宽度，并且交界处水深是二者的平均值. 所以相邻网格之间的作用力是它们之间的平均压力乘以交界面积. 联立式

(1.15) 和式 (1.16)，可得压力造成的横向加速度为

$$a_x=\frac{g(d_{0x}^2-d_{2x}^2)}{4wd_1},$$

其中 w 是网格宽度，d_{0x}是网格 $0x$ 的平均水深，d_{2x}是网格 $2x$ 的平均深度，d_1 是网格 1 的平均水深.

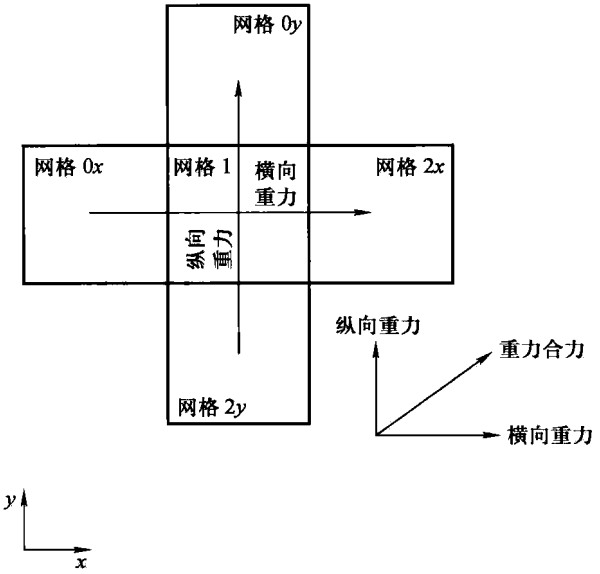


图 1-19 作用在网格上的重力

通过估计当前网格和 4 个相邻网格的地面梯度，可以计算由重力产生的加速度，如图 1-20 所示，根据几何学计算出它的横向分量为

$$a_g=\frac{2\Delta ewg}{4w^2+\Delta e^2},$$

其中 Δe 为相邻网格的地面海拔高度差.

同理可以计算垂直方向的压力和重力加速度.

模型通过大量的步长进行迭代，例如每一步持续时间 1 秒. 在每一步时间段的开始，水被注入溃坝点所在的网格；水量根据上文所述的溃坝模型来确定. 对于每一步，计算每个网格的横向和纵向加速度，区域内每个网格的水流速度根据下式计算：

$$v_{k+1}=v_k+a\Delta t,$$

其中加速度 a 为 x 或 y 方向压力产生的加速度与重力产生的加速度之和.

水流速度的方向由各个网格提供的分量决定，当 $v_x>0$ 时，网格提供的速度向右，当 $v_x<0$ 时，网格提供的速度向左. 而总的水流量与该方向的速度成正比，所以对于一个网格，水深的变化量为

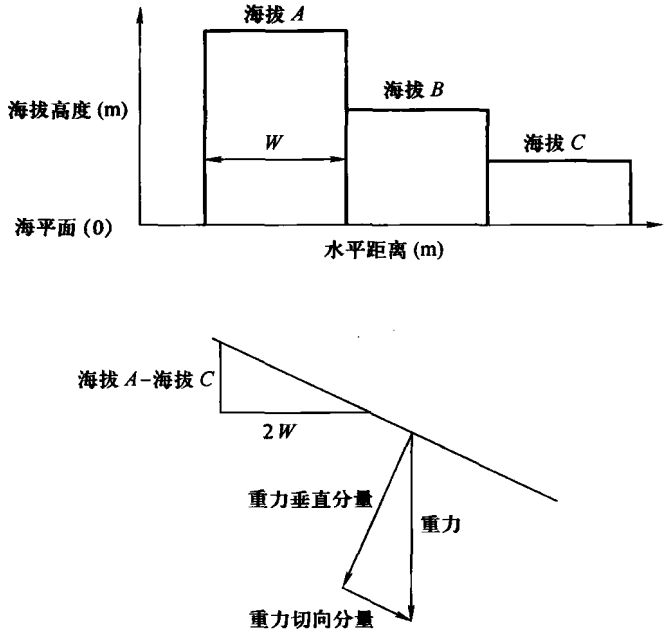


图 1-20 重力切向分量图

$$d_1' = \frac{\bar{d}\Delta t}{2w},$$

其中 $\bar{d}$ 是相邻网格的平均深度.

由于水的流动，相邻网格中的水深也要实时变化，以保证模型中的总水量不变（除了研究区域的边界即溃坝处）.

当水流速度很大时，一个网格能够流出的水量要多于它所包含的水量，特别是当速度乘以时间步长大于网格边长时，流出的水量将会大于网格中现有的水量.

作用力模型的原理比较直观，在水流离散化的基础上，使用一个简单但是广泛应用的物理模型模拟了水的流动. 这个模型计算并保存了速度信息，这使模型能够对洪水泛滥的区域进行精确分析. 例如，这个模型可以预测洪水在哥伦比亚冲击某一座建筑物时的速度，比如州议会大厦.

经过计算和分析发现，模型得出的结论有部分不符合实际，因为含有大量水的网格会具有较小的加速度，这些网格会缓慢地变空，即使它们紧邻空网格；而由于同样的原因，含有少量水的网格倾向于迅速变空. 这样就形成一个棋盘的图案：大的网格越来越大，而它们旁边的小网格越来越小. 这个错误来源于模型的基本假设：一个网格中所有的水都具有同样的速度，一个独立网格中的水不可能向任何一个方向流动. 对于一个简单的地形（比如一条向下游延伸的渠道）来说，可能不会发生这样的问题，但是该模型所建立的网格是

复杂的二维系统，水会同时向两个方向流动，很容易造成上述问题.

模型的另一个不足就是它的复杂性. 模型中对于每个网格的计算包括 7 个步骤：

① 计算横向压力加速度；

② 计算横向重力加速度；

③ 计算横向流动速度；

④ 计算纵向压力加速度；

⑤ 计算纵向重力加速度；

⑥ 计算纵向流动速度；

⑦ 根据流入、流出水量计算新的水深.

计算过程中包含了大量的参数以及多个非线性函数，这意味着进行高精度的计算十分困难.

2. 下坡模型

假设两个网格之间的流量正比于网格质心之间的高度差乘以网格间的作用面积. 这个模型允许水流从一个独立网格的各方向流出，像作用力模型一样，通过划分时间步长的方式来解决问题，每一步增加溃坝点网格的水量.

对于每一步来说，每一个网格都直接同它右方和下方紧邻的两个网格进行交换. 这确保了每一个网格都在一步的时间内与它周围的 4 个网格进行交换. 一般情况下，交换方式为从左方和上方网格流入，从右方和下方网格流出. 网格的水深变化量根据以下的公式进行计算：

$$d_1' = k\,\overline{d}(e_0 - e_2). \tag{1.17}$$

其中，参数 k 的值是基于溃坝处水流速度的峰值为 30 m/s 的假设. 接下来通过 k 值的变化来观察模型的变化.

根据式 (1.17)，由于水的质量守恒，与网格 1 相邻的网格高度变化量即为 d_1' 的负值. 为了保证计算过程的一致性，每一个网格的水深都是在每一步计算结束时才发生改变，也就是先计算所有网格的变化量，再统一将变化量加到原有水深上. 如果一个网格流出了超过其初始包含的水量中允许流出的部分，则将流出量削减到模型允许的最大值. 经过分析和尝试计算发现，大的网格中的水将会完全流动到旁边的空网格中去，而在下一步计算时又会有上面的网格传递回来，所以无论何时总会有一半的网格是空的. 设定参数 $L_{\max}$（单位时间内网格水量的最大流失比例）用以防止这种棋盘式的波动.

对于右方和下方边界处的网格，可以设定它们处于地图外沿的边的交换水量等于对边的水量，即有多少流入就有多少流出，反之亦然. 由于这些网格距离溃坝区域很远，关于它们的计算精度并不重要，所以这一方法的目的仅仅在于是确保水流能够满足物理规律，即连续地流出网格图的边界.

下坡模型使用简单而易于验证的方法提供了一种快速计算方法. 尽管网格

间的水流交换控制方程缺少一些直接的物理推导，但是通过参数的调整，能够得到稳定的、物理性的结论. 这是因为洪水在下坡或者穿过临近的平原时流动速度很快，这一速度基本不会发生变化.

1.5.5 模型的模拟求解

1. 需要的参数取值与说明

$\Delta t=10$：每段时间的长度（单位：s）；

$d_{\min}=10^{-4}$：网格没有水时的深度（单位：m）；

$t_f=10^5$：水库中的水通过溃口完全流出的时间（单位：s）；

$t_b=3\ 600$：溃口达到最大的时刻（单位：s）；

$Q_{\max}=25\ 000$：水流通过大坝的最大流量（单位：m^3/s）；

$Z_{\min}=10$：溃口底部距离坝底的最小深度（单位：m）；

$V_w=2.714\times10^9$：正常情况下墨里湖水体积（单位：m^3）；

$k=0.504$：传递系数（水流在两个网格间流动的调节系数）；

$L_{\max}=0.25$：一个时间段内每个网格水量的最大流失比例.

2. 模拟求解的过程

地形数据可以使用美国地理调查报告中的海拔数据[13]. 这些数据包括一系列按经纬划分的海拔高度数值. 模型所建立的网格为边长为30 m的正方形，为了减少计算量，将7×7个方格一起计算，因此实际使用的网格边长为210 m.

针对4种不同的溃坝模型和溃决持续时间（$t_b=360$，1 800，3 600），将代表溃坝点的网格放置在大坝正前方，对作用力模型和下坡模型进行检测. 出于最安全情形的考虑，使用灾害程度最高的瞬间溃坝模型检测 k 和 $L_{\max}$ 的可靠性.

为了避免在计算某些网格水量较少时出现误差，将水面低于0.000 1米的网格视为空，这种处理在作用力模型中尤为重要，因为当这样的网格位于包含很多水的网格旁边时，将会产生相当大的流速（大于10^6 m/s）.

使用瞬间溃坝模型检测下坡模型，总步数设定为50 000，结果证明这个模型是可行的，但是在40 000步之后计算结果精确性降低，这是因为这时流出地图范围的水流量已经远大于溃坝处洪水水源的流量. 考虑极端情况，以很大的流量（每一步溃坝点网格的水深增加10 m）来检测这个模型，计算只持续了1 000步就变得不再稳定，这个流量几乎是4种溃坝模型中最大流量的30倍.

1.5.6 模型的结果分析

1. 综合结果

模型的数值结果是关于网格的三组数据：0、1取值描述了网格是否受到

了洪水的影响；每个网格中的水深决定了地区受洪水影响的严重程度；每个网格的水深与海拔高度之和，描述了该地区实际水面高度. 从这三组数据和图1-21[23]，可以看出洪水的蔓延情况以及严重程度.

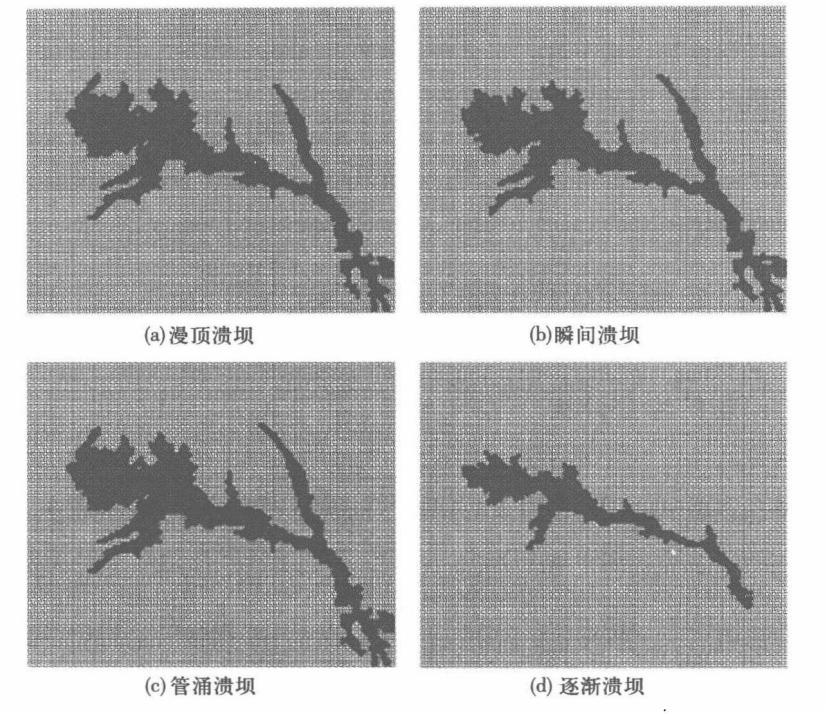


图 1-21 四种溃坝模型 24 小时后洪水泛滥区域示意图

模拟结果显示，洪水的蔓延区域与溃坝模式之间没有显著的联系，不同的溃坝模式只是影响洪水的蔓延速度. 对于瞬间溃坝，洪水最大覆盖面积达到 1 066.5 km^2，在萨卢达河与康加里河交汇处的平坦开阔区域将会出现最大洪峰，而位于这片开阔区域之上的哥伦比亚市，受洪水的影响非常小. 模型并没有模拟洪水向康加里河下游更远处蔓延所产生的后果，但是可以猜测与已经计算过的区域会有一定的相似性.

2. 罗尔斯河和州议会大厦

进入罗尔斯河的洪水会在一定的地域内延伸，但是不会形成向上游流动的洪峰. 尽管很难精确计算出进入罗尔斯河的洪水起止位置，但是模型能够计算出罗尔斯河流域的洪水覆盖范围会达到 1.6 ~ 2.4 km^2，其中洪水流动最快的点位于上游 4.4 km 处.

即使在最坏情况即瞬间溃坝模式下，洪水也不会到达州议会大厦.

3. 参数的检验

下坡模型的求解包括两个重要的参数，传递系数 k 和水量最大流失比例

L_{max}，这两个参数都无法由实际观测得到，因此需要对其合理性、可靠性进行分析.

模型根据参数 k 来计算网格间的水流量，它的值基于溃坝处水流速度最大值为 30 m/s 的假设. 经验证，k 值并不会显著地影响计算. 利用 1.5.5 节的参数取值在瞬间溃坝模式下进行对比计算，当 k 乘以 100 时，在计算 1 800 步之后，受到洪水影响的总面积变化仅为 17%.

参数 L_{max} 用于避免每个网格在计算中提供过多的水量，然而洪水的蔓延范围并不取决于它的值，经过多次尝试性的计算，取其为 0.25 以确保合理、平滑的水量分布.

1.6 问题的综合分析与进一步研究

1.6.1 问题的综合分析

溃坝洪水估计是一个具有十分重要的现实意义的问题. 就萨卢达大坝和墨里湖来说，准确地估计溃坝洪水的影响范围直接影响灾害预警和救援计划的制订，也是大坝维修和建设工程的重要依据. 在南卡罗来纳州应急事务管理部(South Carolina Emergency Management Division) 的网站上可以找到萨卢达流域泛洪区的地图[26]. 事实上早在 20 世纪 80 年代，工程师已经对萨卢达大坝有可能发生的溃坝进行了分析，并设计了补救工程[4].

目前溃坝洪水研究普遍存在的难题是，无法通过实验得到准确的数据，而实际发生过的溃坝事故又很少有数据记录，因此运用数学模型从机理分析的角度出发是解决这一问题的有效手段.

针对建模竞赛，解决溃坝洪水估计问题可以遵循以下两个简单的原则：

(1) 建模思路应该由浅入深. 即首先建立最基本的水流模型，比如可以假设萨卢达大坝瞬间溃决，假设萨卢达河是直线，且宽度、坡度均保持不变. 然后再考虑河道是弯曲的、坡度是变化的、大坝是逐渐溃决的等条件，逐步深入. 渐进式的解决问题方式之所以适用于这一问题，关键在于缺少可以用于验证模型的实际数据，尤其是萨卢达大坝坝体材料的数据，在竞赛期间几乎不可能获得. 如果一开始就考虑所有因素，从物理的角度建立一个理论上完善的模型，那么在计算过程中会因为过多的参数而难以为继.

推广到一般情况，对难以用实际数据验证的问题，建立多个模型，从不同角度进行分析，会获得相对客观的结论；或者在单一模型的基础上，通过调整参数模拟各种不同的现实情况，得到一组综合结果，也能起到类似的作用.

(2) 建模过程应该包括溃坝和洪水演进两方面. 土石坝的溃决有多种形式，逐渐溃坝最为常见，但难以建模和计算；瞬间溃坝计算最为简单，但现实

中很少发生. 题目中只给了一个模糊的假设——地震，然而不同震级的地震对大坝的破坏程度也不同. 在无法确定大坝溃决模式的情况下，针对多种模式分别建模十分必要.

关于洪水演进有一个经典的模型——圣维南方程组，这是一个非线性偏微分方程组，用以描述水道和其他具有自由表面的浅水体中渐变不恒定水流的运动规律，在水文和水利工程领域有着广泛的应用. 圣维南方程组很难用解析方法求解，一般只能通过简化或数值解法得到结果. 此外也可以建立一维或二维的离散模型，将萨卢达河流域划分为数段或数个区域，模拟水流在各段河道或各区域之间运动.

溃坝模型与洪水演进模型相结合才能完整地解决溃坝洪水估计问题，前者解决洪水的源头大小，后者解决洪水的流动过程，两者缺一不可.

以下针对本章介绍的模型进行分析和总结.

1. 逐渐溃坝模型

逐渐溃坝是最为常见的土石坝溃坝模式，1.2 节所建立的模型首先从明渠水力学的角度出发，建立大坝被侵蚀的模型. 然后将下游河道分为 4 段，即一维离散化，假设其平直且截面为矩形，再结合曼宁方程建立洪水演进模型. 最后利用实际观测到的浪涌数据来估计模型参数. 其特点是模型结构相对简单，并且充分利用了流体力学的结论，可以说是用最小的工作量解决了问题.

逐渐溃坝模型有如下三个改进方向：

（1）模型利用萨卢达河小规模浪涌的实测数据来估计参数，因此可以很好地描述小规模浪涌的推进过程. 但是将小规模事件的规律外推到大规模事件往往是不可靠的，对于所研究的溃坝洪水（大规模浪涌），小规模浪涌的参数未必适用. 要想得到更准确的预测结果，模型还需要对萨卢达河或者其他类似河流的大规模浪涌进行验证.

（2）模型对溃坝过程的分析，大量采用了现有结论，包括参数的取值，而设置这些参数的本意是根据大坝具体情况的不同，对方程进行调整，直接使用经典的参数值并不一定符合萨卢达大坝的实际情况. 如果能获得萨卢达大坝的材料和结构的详细信息，对参数进行重新估计，水流对大坝的侵蚀过程建模会更加精确.

（3）模型对河道的分析，使用了最为简单的平直矩形渠道，这一形式是对现实河道的高度简化. 由于一维离散模型本身并不考虑方向和坡度的变化，因此主要矛盾集中在河道的横截面形状上. 现实情况下，河道边缘是自上而下向中央延伸的，如果采用（等边）梯形等更符合实际的河道出口剖面图，可以使得曼宁方程以及整个洪水演进模型的结果更加准确.

2. 瞬间溃坝模型

与逐渐溃坝模型类似，瞬间溃坝模型也采用了流体力学的结论，包括用正

弦波模拟溃坝洪峰，应用圣维南方程组及其稳定解等. 其优点是简化了建模过程，不需要在有限的竞赛时间里做大量的数学推导，可以把精力集中在模型的解释、参数估计和结果分析上. 很多参赛队当时都选择了圣维南方程组，但是普遍都把精力放在了求解上，忽略了对圣维南方程组基本假设的解释和分析，最终尽管有不错的结果，但很难得到评委的青睐. 与此同时，瞬间溃坝模型有以下三点不足：

（1）参数估计所用到的数据来源于萨卢达河较小规模的浪涌，其流量与瞬间溃坝产生的洪峰流量相比，相差2～3个数量级，因此要得到更精确的参数，必须对其他类似的溃坝洪水波进行全面细致的研究.

（2）对曼宁糙率的估计过于简单. 萨卢达河流域从卫星地图上看，基本上被植被覆盖，但是不同类型的植被对应的曼宁糙率有数倍的差距，这一参数的精确与否，直接关系结果的精确程度，即使无法实地考察，也应该假设多个糙率值进行计算.

（3）模型做了很多粗糙的简化假设，包括圣维南方程组的假设也都是比较理想的情况，因此获得的洪峰相关数据只能看做是实际结果的上限. 如果考虑更多的因素，比如洪水推进过程中的方向变化，将圣维南方程组由一维转化为二维，模型的预测结果会更有现实意义.

3. 差分形式的圣维南模型

差分形式的圣维南模型是分析洪水演进的常用方法之一，如上文所述，圣维南方程组需要通过数值解法求解，该模型采用Lax-Wendroff差分法将其转化为差分方程组. 其优点是将经典的圣维南方程组应用解决该问题已经被证实为有效，因此模型从理论上来说是完善的.

1.4节所建立的差分方程模型是经典模型与求解方法相互妥协的结果，将偏微分方程组近似为差分方程组可以求得数值解，且精度易于控制，同时也会产生误差，尤其是在一些极端情况下，模型的解会出现不稳定的情形.

对泛洪区的假设也是一种妥协，为简化模型，假设泛洪区的水流瞬间流入且立即静止，并不需要为泛洪区单独建立模型，只需要认为把河道容纳不下的水“泄入”泛洪区即可. 但这显然是不符合实际的，只能看做是洪峰过后洪水达到稳定时的状态. 如果能将模型扩展为三维，将洪水漫滩过程作为第三个维度，即水流流往下游的过程中不断从河床溢出，逐渐形成泛洪区，模型结果将可能得到极大的改善.

此外，如果希望得到更精确的结果，模型还需要做一些改进：

（1）用梯形溃口来代替矩形溃口.

（2）观测从墨里湖到哥伦比亚的河道宽度，而不是假设河流宽度不变.

（3）按一定的规律观测河道的高度，比如河流会有落差，这会引起洪水的冲击，造成流速的变化.

(4) 将模型从一维直线河流扩展到二维情况，既考虑沿河道方向的变化，又考虑河岸对水流动量的影响.

4. 离散网格模型

离散网络模型属于二维离散模型. 对于4种溃坝模式造成的溃坝洪水流量变化，出于简化模型考虑，所选择的函数都较为简单，但能反映不同溃坝模型的特点. 尤其是管涌溃坝和漫顶溃坝，从溃坝发生的表现形式出发，更加符合机理分析的建模思路. 然后将下游区域，包括河道和泛洪区分为一系列210 m×210 m的网格，通过流体力学的伯努利方程建立作用力模型，通过直观认知建立下坡模型，统称为离散网格模型.

模型的优点有三个方面：

(1) 独立性较好，给出一个地区的海拔数据和根据溃坝得出的水流控制方程，就可以计算出洪水的蔓延情况，而且结果精度随着输入数据的精度提高而提高.

(2) 下坡模型仅仅基于简单的网格间水流流动规律，这使模拟过程很容易通过参数调整来检验模型中存在的问题.

(3) 模拟算法是有效的. 尤其是对下坡模型，其单步计算的过程不包含非线性函数，这使模型能够在短时间内有效地计算不同溃坝类型所带来的不同洪峰，以及受洪水影响的范围.

模型主要的不足包括两个方面：

(1) 水流在洪水蔓延区域存在波动的趋势，即高水位网格中的水会流入相邻的低水位网格，在下一个时间段内又会回流. 这就需要引入最大失水比例来控制波动程度.

(2) 时间步长的取值较为随意. 下坡模型中唯一与时间直接相关的只有传递系数，传递系数本身并不具有物理意义，因此模型的时间步长取值与洪水蔓延的范围没有直接关系. 如果需要计算精确的洪水蔓延过程，例如溃坝发生后多长时间洪峰到达某个地点，还需要对时间刻度通过一些模拟计算进行校正，例如通过检测水的速度来确定时间步长.

1.6.2 进一步研究的问题

1. 溃坝模型

如上文所述，溃坝可以分为多种形式，需要分别建立模型. 不同的溃坝模型产生的洪峰有很大的区别，但是都可以用大坝的水流量关于时间的函数来表示，因此溃坝模型的核心问题是构建这个函数以及估计相关的参数.

数学建模的基本方法不外乎两种，机理分析和测试分析. 本章所介绍的几个溃坝模型都属于机理分析的范畴，即根据大坝溃决的物理过程，寻找可以利用的内部规律进行建模. 从理论上讲，测试分析方法也可以用来建立溃坝模

型. 因为溃坝模型的本质是建立溃坝洪水流量与坝高、溃口高度、溃口形状、库容水体等因素之间的函数关系，因此可以在实际数据的基础上，通过回归分析等测试分析方法给出近似的函数关系. 也可以先用量纲分析建立各物理量之间的关系，再通过实际数据做参数估计. 但是研究溃坝问题最主要的困难就是实际数据难以获取，而回归分析高度依赖输入数据的准确度，因此并不适用，命题人对这一类方法也持否定态度[27].

回到机理分析方法. 目前的研究结果表明，土石坝的溃决模型主要分为两大类. 第一类称为基于参数的模型，主要利用一些关键参数（如溃口最终宽度、溃口形成历时等），通过简单的时变过程（如溃口尺寸的线性发展理论）模拟溃口的发展. 总的来说，这类模型较简单，对数据输入要求较少，使用方便，适合数学建模竞赛这种要求短时间内建立模型并计算结果的情形. 但是这类模型未涉及实际溃坝机理，准确度不够，计算结果也不太稳定，在实际应用中仅用于初步计算. 本章先后建立了6个溃坝子模型，不难看出与实际的溃坝过程相比，其假设都过于理想化，尤其是完全没有考虑溃坝泥沙造成的冲刷效果.

第二类模型称为基于物理过程的模型，通过综合水力学、泥沙、土力学等学科知识构建一个时变过程以模拟实际溃坝过程和溃坝洪水过程线，比较典型的有 P-T 模型，BEED 模型和 BREACH 模型等[8].

上述模型的共同点在于都采用了牵引力分析以及泥沙转移理论来模拟水流对坝体的冲刷. 与第一类模型相比，第二类模型更加注重对溃坝的物理过程进行模拟，其结果精确度更高，而且曾多次应用于真实发生的溃坝事故. 但是这些模型的运用需要大量的输入参数，尤其是坝体材料方面的数据几乎是必需的，这些数据很难在短时间里获得.

对于第二类模型，在个别参数无法获取的情况下，一是可以直接使用权威文献的数据，二是可以针对某个参数设定多个数值，分别计算，给出一个综合的结果.

综合来看，针对土石坝溃决的建模，现有的方法无论是简单还是复杂，都是基于溃坝的物理过程，从其他角度进行分析的方法几乎没有. 究其原因是因为用以校正参数的实际数据难以获取，只能通过尽可能精确地模拟物理过程来提高精度. 然而，对于土石坝筑坝材料的不稳定性决定了在溃坝过程中必然存在间歇性的垮塌，要模拟这种不连续的溃决过程只能依靠不断提高模型的微观程度，这使得模型结构越来越复杂，规模越来越庞大.

实际上，由于实验和观测技术的进步[28]，以及历史数据的积累，完全可以尝试采用模糊综合评判、灰色关联分析等方法先对影响溃坝洪水流量的各个因素进行分析，确定主要因素，排除次要因素. 然后在第二类模型的基础上进行简化处理，这样既保持了第二类模型的精确度，又具有第一类模型简单易行

的特点.

2. 连续型的洪水演进模型

由于数学建模竞赛时间有限，运用圣维南方程组建模的参赛队普遍采取一维模型，即忽略河道方向改变造成的影响，假设水流始终沿河道主流方向前进. 从萨卢达河河道的实际情况来看，整体流向没有太大的变化，而且河道宽度在大部分流域保持一致，所以针对这一问题，使用一维圣维南方程组可以得到相对满意的结果. 但是根据一维圣维南方程组的基本假设（见 1.3.1 节），有两点明显不符合实际，一是河道存在一定的弯曲，水流在转向过程中会消耗一部分动能，也就是说弯曲的河道可能对洪水形成阻滞；二是在萨卢达河与康加里河的交汇处，河道突然变宽，且流向有较大的变化，具有明显的二维特征. 虽然这部分河道占整个流域的比例不大，但是位置紧邻哥伦比亚，这是问题研究的重点.

需要说明的是，以上两点不会都对洪水造成阻滞作用，也就是说如果发生溃坝，在其他条件不变的情况下，哥伦比亚的受灾程度会小于由一维圣维南方程组的计算结果. 这看似是一种稳妥的做法，实则不然，现实中模型的结果直接关系到洪水预警等级和居民疏散计划的制订. 过高地估计洪水的影响，会导致这些计划浪费大量的公共资源.

因此有必要研究二维浅水方程组在这一问题中的应用. 与一维方程组类似，二维方程组也由连续方程和运动方程构成[29]：

$$\begin{cases} \dfrac{\partial d}{\partial t}+u\dfrac{\partial d}{\partial x}+v\dfrac{\partial d}{\partial y}=0, \\ \dfrac{\partial u}{\partial t}+u\dfrac{\partial u}{\partial x}+v\dfrac{\partial u}{\partial y}+g\dfrac{u\sqrt{u^2+v^2}}{c_0^2 d}+g\dfrac{\partial d}{\partial x}=0, \\ \dfrac{\partial v}{\partial t}+u\dfrac{\partial v}{\partial x}+v\dfrac{\partial v}{\partial y}+g\dfrac{v\sqrt{u^2+v^2}}{c_0^2 d}+g\dfrac{\partial d}{\partial y}=0. \end{cases}$$

其中第一个方程为连续方程，后两个是运动方程，d 是水深，u、v 分别是 x、y 方向的流速，g 是重力加速度，c_0 是泄洪系数，t 是时间.

二维浅水方程组同样需要用数值方法求解，且已经有一些成熟的算法[29,30].

3. 离散型的洪水演进模型

离散的洪水演进模型同样可以分为一维和二维两类. 1.2.1 节中的河道分段法就是一个简单的一维离散模型，水在每段河道内不改变流向. 1.5.4 节中的离散网格模型则是二维离散模型，水可以向横向和纵向两个方向移动.

需要注意的是本章所涉及的离散模型，严格地说属于连续问题离散化的模型. 洪水演进过程本身是连续的，可以认为客观存在连续的物理和数学规律，建立离散模型的目的在于用离散化的模型近似连续的洪水演进规律. 其建立过

程有如下几个要点：

（1）对研究对象进行合理的划分. 无论是一维还是二维模型，都涉及将原有的整体对象拆分为多个规则的线段或者方格. 拆分的原则可以参考积分的定义，用多个矩形近似一个曲边梯形，每一部分越小，得到的结果精确度越高. 因此离散模型对研究对象的划分首先应该足够小，才能保证计算的精度.

但是在实际操作过程中，并不一定越小越好. 划分得越小，需要计算的对象数量越大，尤其是对二维模型，边长每缩小一个数量级，方格数量增长两个数量级. 过大的数量带来过大的计算量，不利于快速得到模拟结果，也不利于对模型进行调整和验证.

因此对研究对象的划分，需要在划分精度和计算量之间寻求平衡. 常用的方法是先粗后细，先用粗略的划分验证模型和算法，再提高划分精度，在可接受的时间内计算精确的结果.

（2）时间与空间的关系. 洪水演进过程是与时间相关的，因此在离散模型中必须考虑时间步长和线段的长度或者方格的边长之间的关系. 以一维模型为例，一个典型的问题是，如果参数设置不当，单位时间内水流运动的距离超过分段河道的长度，那么每一段河道每一步的计算都需要涉及不相邻的河道段，使得建模过程过于复杂，也失去了离散模型简化问题的优势. 这时就需要缩小时间或者空间的单位，具体的操作取决于两者在模型中的作用以及缩小后产生的计算量.

（3）各部分之间的联系. 由于离散模型将空间和时间的单位缩小，因此可以做一些较强的假设，如某些参数在单位时间、单位空间内是近似不变的，以此简化洪水演进的物理过程，用相对简单的物理（流体力学）公式来构建模型. 例如一维模型中每段河道出口处的流量，可以忽略河道内部地形的影响，直接建立出口横截面的流量公式；二维模型中相邻网格之间的相互作用，可以运用流体力学中广泛运用的伯努利方程，甚至可以认为水流的转移量与水位高度差成正比.

这些函数或者公式的建立，一般情况下应选择线性函数，不能选择线性函数时则应尽可能使用多项式函数，尽量避免使用指数函数等计算复杂度较高的函数.

在目前的数学建模竞赛论文中，无论是洪水估计问题还是其他类似的可离散化的连续问题，一个共同的特点是普遍采用了标准的划分方式. 首先是一维模型，尽管在分段长度上有不同的方法和划分结果，但是无一例外地使用平直的河道进行近似. 事实上，对于数公里长的河道，不可能没有方向的变化，但是，如果考虑方向的变化又会变成二维模型. 那么有没有办法用一维模型模拟河道的变化呢？

观察卫星地图可以发现，萨卢达河在所研究的流域，是由几个较平缓的弯

道构成的，因此不妨使用简单的曲线来近似分段河道，例如抛物线. 在河道的分段点，可以假设河道沿曲线的切线方向是平直的，并不影响使用曼宁公式. 虽然用曲线近似的河道不能直接求得一维的坡度，但是可以通过伯努利方程计算水流坡度来近似.

对于二维模型，也并不一定需要标准的矩形划分方式. 虽然矩形网格生成简单、省时，但是最大的缺点是生成的锯齿状边界会造成流动失真. 现实中的泛洪区往往具有复杂的三维几何外形，而且铁路、公路、堤防等设施都会对洪水的演进过程产生阻滞作用. 矩形网格在复杂地形下提高精度非常困难，因此可以考虑使用无结构不规则网格[31]. 这一网格划分方式可以根据地形地物，将网格设计成三边形、四边形或五边形，在网格中心点计算出水位高度，在网格周边通道的中点处计算出水流量.

需要说明的是，不仅无结构不规则网格可以单独使用，建立二维离散模型，还可以作为有限单元法的一部分，用以求解圣维南方程组.

综合上述分析，可以建立一个连续和离散统一的模型，即首先建立一维或二维的圣维南方程组，然后将所研究的萨卢达河流域离散为曲线的河道（一维）或者无结构不规则网格（二维），最后通过有限单元法求解. 这样既保持了连续模型的理论优势，又具有离散模型易于计算的优点.

参考文献

[1] Wikipedia. Saluda River - Wikipedia, the free encyclopedia [G/OL]. [2011-03-16]. http://en.wikipedia.org/wiki/Saluda_River.

[2] Wikipedia. Saluda Dam - Wikipedia, the free encyclopedia [G/OL]. [2011-03-16]. http://en.wikipedia.org/wiki/Saluda_Dam.

[3] SCIWAY. South Carolina SC-Detailed County Maps[G/OL]. [2011-03-17]. http://www.sciway.net/maps/cnty/.

[4] 保罗 C. 里才等. 塞卢达坝补救工程 [J]. 水利水电快报，2003，24（16）：4-6.

[5] Wikipedia. Lake Murray (South Carolina) - Wikipedia, the free encyclopedia [G/OL]. [2011-03-16]. http://en.wikipedia.org/wiki/Lake_Murray_(South_Carolina).

[6] 汝乃华，牛云光. 大坝事故与安全：土石坝 [M]. 北京：中国水利水电出版社，2001.

[7] 邹双凤. 土坝溃坝数值模型和溃坝洪水演进研究 [D]. 南宁：广西大学，2009.

[8] 朱勇辉，廖鸿志，吴中如. 土坝溃决模型及其发展 [J]. 水力发电学报，2003（2）：31-38.

[9] Ryan Bressler, Christina Polwarth, Braxton Osting. Analysis of Dam Failure in the Saluda River Vally [J]. UMAP Journal, 2005, 26 (3): 263-278.

[10] Chow V T. Open-Channel Hydraulic [M]. New York: McGraw-Hill, 1959.

[11] Singh P. Vijay. Dam Breach Modeling Technology [M]. Dordrecht, Holland: Kluwer Academic Publishers, 1996.

[12] NGDC. National Geophysical Data Center Coastal Relief Model [DB/OL]. [2011-03-17]. http://www.ngdc.noaa.gov/mgg/gdas/gd-designagrid.htm.

[13] USGS USGS Real-Time Data for the Nation-Streamflow [DB/OL]. [2011-02-17]. http://waterdata.usgs.gov/nwis/current/? type=flow.

[14] Chanson Herbert. The Hydraulics of Open Channel Flow [M]: An Introduction. 2nd ed. Amsterdam, Holland: Elsevier, 2004.

[15] Ponce V M, Simons D B. Shallow wave propagation in open channel flow [J]. Journal of the Hydraulics Division, 1977, 103 (12): 1461-1476.

[16] Fread D L. Transient hydraulic simulation: Breached earth dams [D]. University of Missouri-Rolla. 1971.

[17] Ponce V M, Taher-shamsi A, Shetty A V. Dam-breach flood wave propagation using dimensionless parameters [J]. Journal of the Hydraulics Divi-sion, 2003, 129 (10): 777-782.

[18] Clay Hambrick, Katie Lewis, Lori Thomas. From Lake Murray to a Dam Slurry [J]. UMAP Journal, 2005, 26 (3): 219-278.

[19] US Army Corps of Engineers. Engineer Manual 1110-2-1420. Engineering and Design—Hydrologic Engineering Requirements for Reservoirs [M/OL]. [2011-03-17]. http://www.usace.army.mil/inet/usace-docs/eng-manuals/em1110-2-1420/. 1997.

[20] 赵克玉. 天然河道一维非恒定流数学模型 [J]. 水资源与水工程学报, 2004, 15 (1): 38-41.

[21] Moussa, Roger, and Claude Bocquillon. Approximation zones of the saint-venant equations for flood routing with overbank flow [J]. Hydrology and Earth System Sciences. 2000, 4 (2): 251-261.

[22] 陆金甫, 关治. 偏微分方程数值解法 [M]. 北京: 清华大学出版社, 1987.

[23] Jennifer Kohlenberg, Michael Barnett, Scott Wood. Through the Breach: Modeling Flooding from a Dam Failure in South Carolina [J]. UMAP Journal, 2005, 26 (3): 245-261.

[24] 陈生水, 钟启明, 任强. 土石坝漫顶破坏溃口发展数值模型研究 [J]. 水利水运工程学报, 2009 (4): 53-58.

[25] Sedimentation and River Hydraulics Group. Comparison between the methods used in MIKE11, FLDWAV 1.0, and HEC-RAS 3.1.1 to compute flows through a dam breach [R]. US Department of the Interior. 2004.

[26] SCEMD. evacuation map [EB/OL]. [2011-03-17]. http://www.scemd.org/evacuation_map.pdf.

[27] Daniel Zwillinger. Judge's Commentary: The Outstanding Flood Planning Papers [J]. UMAP Journal, 2005, 26 (3): 279-281.

[28] 沈登乐. 土坝溃坝试验研究与应用 [J]. 江淮水利科技, 2009 (5).

[29] 黄金池，何晓燕. 溃坝洪水的统一二维数学模型 [J]. 水利学报，2006，37（2）：222-226.

[30] 李大鸣，林毅，徐亚男，周志华. 河道、泄洪区洪水演进数学模型 [J]. 天津大学学报，2009，42（1）：47-55.

[31] 范玉，陈建，李大鸣. 一、二维洪水演进数学模型在滞洪区的应用 [J]. 华北水利水电学院学报，2009，30（4）：12-15.

第2章　高速公路收费亭设置问题

高速公路收费亭的设置问题是2005年美国大学生数学建模竞赛的B题，研究的是交通管理、工程设计和优化管理等领域的一个非常实际的问题．实际中，每一条高速公路都设置有若干个收费站．经常可以看到经过收费站的车辆排长队的现象，但如果收费亭设置过多，也可能会造成资源的浪费．因此，对于每一个收费站而言，依据实际的车流量合理设置收费亭是一个值得研究的问题．

2.1　问题的综述

2.1.1　问题的提出

随着人们生活节奏的加快，高速公路网的交通也日趋繁忙，对于收费的高速公路更是如此．诸如美国新泽西州的风景区干道、95号州际公路等．这些公路是有多个车道的交通干线，在公路上每隔一段距离会设置一个收费站．由于收费亭设置的不合理，经常会影响或限制车辆通行，甚至造成交通阻塞，因此收费站对于驾驶员来说，一般是不受欢迎的．如何使得收费站的设置更合理，以使驾车员的不满意程度最小化，提高交通效率，成为高速公路优化设计的重要问题．通常，一个收费站内都会设置多个收费亭，其数目往往会远多于进入收费站的高速公路车道数．车流进入收费站时，成扇形散开进入收费亭；当离开收费站时，车流又会汇合到与进入收费站之前车道数相同的高速公路上．因此，在交通繁忙时，车辆会在离开收费站时出现拥塞；更严重的时候，收费站的入口也会由于车辆多付费所花时间过长而排长队．

试建立数学模型以确定收费站的最佳收费亭数目，并考虑为每个高速公路车道仅设置一个收费亭的情况．将你制订的方案与现有的方案比较，说明在什么情况下你的方案更有效或是无效？注意：“最优”是由你自己来定义的．

问题的原文如下：

Tollbooths

Heavily-traveled toll roads such as the Garden State Park way in New Jersey, Interstate 95, and so forth, are multilane divided highways that are interrupted at intervals by toll plazas. Because collecting tolls is usually unpopular, it is desirable to minimize motorist annoyance by limiting the amount of traffic disruption caused by the toll plazas. Commonly, a much larger number of tollbooths is provided than the number of travel lanes entering the toll plaza. Upon entering the toll plaza, the flow of vehicles fans out to the larger number of tollbooths, and when leaving the toll plaza, the flow of vehicles is required to squeeze back down to a number of travel lanes equal to the number of travel lanes before the toll plaza. Consequently, when traffic is heavy, congestion increases upon departure from the toll plaza. When traffic is very heavy, congestion also builds at the entry to the toll plaza because of the time required for each vehicle to pay the toll.

Make a model to help you determine the optimal number of tollbooths to deploy in a barrier-toll plaza. Explicitly consider the scenario where there is exactly one tollbooth per incoming travel lane. Under what conditions is this more or less effective than the current practice? Note that the definition of "optimal" is up to you to determine.

注：原文取自文献[1].

2.1.2 问题的背景资料

社会经济的飞速发展和交通建设的相对滞后，已经成为困扰全球的世界性问题，交通拥挤成为困扰各国大中城市发展的“瓶颈”之一. 早在20世纪90年代，在美国主要城市周围，高度密集的车流使得交通高峰期间高速公路上的车速降到56 km/h以下，过大的车流造成时间的耽搁总计达到每年50亿小时以上[2,3]. 交通拥堵造成了巨大的经济损失，德克萨斯运输研究所对39个主要城市的研究表明：美国每年因交通拥堵造成的经济损失以千亿美元之计，欧洲每年因交通拥堵所造成的经济损失为数千亿欧元[4]. 交通拥挤导致车辆只能在低速态行驶，频繁停车和启动不仅增加了汽车的能源消耗，而且也增加了尾气排放量，加剧了噪音污染.

到2008年末，我国高速公路通车总里程已达到6.03万公里. 目前我国已经成为全世界拥有最多收费道路的国家，也已成为全世界拥有高速公路通车里程第二的国家[5].

然而随着高速公路的建设和车流量的变化，出现了一系列的矛盾[6]. 高速公路收费排队交费的场景，如图2-1. 已有研究表明车辆在高速公路上的延误时间中有36%是由于停车收费原因造成的，收费站是公路交通流的瓶颈.

世界银行报告提出[7]：根据国际经验，在有效经营的情况下，数量适中、设备较为简单的收费设施的运营成本与收入比为5% ~10%，如果收费设施的数量大于所需数量时，运营成本与收入比增大到10% ~15%，如果经营不善，则可能高达30%以上. 此外，某些地区公路收费站设置的过于密集，严重地影响了出行效率，例如，某些地区公路收费站的平均间距仅为18.143 km.

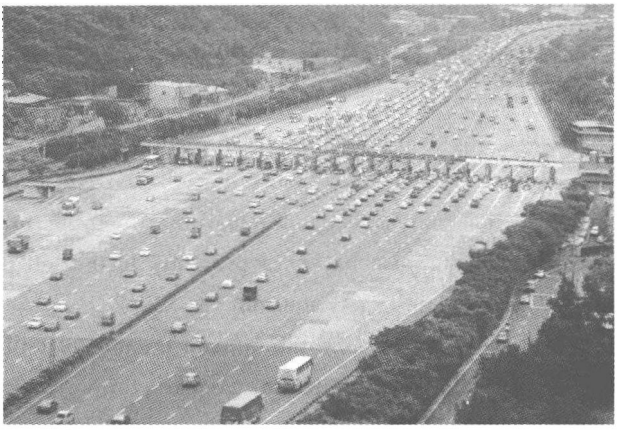


图2-1　高速公路收费站排队交费的场景

为缓解交通拥堵问题，世界各国都相继投入了大量的人力、物力和财力进行交通问题的研究. 在提高收费站的通行能力方面，研究比较多的是电子收费系统（Electronic Toll Collection，ETC），驾驶员不需停车就可缴费通行. 由于其具有消除收费站附近车辆延迟、节省人工及简化收费手续等优点，已经逐渐在欧美各国的收费公路上布设使用. ETC已经成为美国回收公路投资和养护费用的高效率手段，最著名的联网运行电子不停车收费系统是E. Zpass系统[8]. 1997年7月，“E. Zpass”工程的最终运行方案开始付诸实施和运行. 截止1998年12月，仅经过一年半的时间，共计23条专用ETC车道的电子不停车收费网络就承担了整个月平均交易量的43%，高峰时段甚至达到55% ~60%. 网络化运行的电子不停车收费系统的效益和吸引力充分体现出来. 葡萄牙的Via Varde电子收费系统可以算作欧洲具有代表意义的联网电子收费系统之一，根据运营报表统计数据，人工收费（Manual Toll Collection，MTC）车道的平均通行能力为200辆/h，电子收费车道的平均通行能力为1 500辆/h，1条ETC车道的通行能力是MTC车道通行能力的7倍多.

ETC最早引进我国是在20世纪90年代中期. 那时，我国部分经济发达地区的高速公路车流量激增，导致了收费口交通堵塞、环境污染等一系列问题. 据ISO/TC204中国委员会统计资料，当年仅广州地区停车等待交费损失的车时就达数百万车小时，由此导致的汽油浪费以亿元计算[9].

先进的交通流理论的应用、先进的交通工程与技术可以产生重大的经济和

社会效益. 20 世纪 60 年代，纽约市政府拟新增一条从曼哈顿（Manhattan）连接新泽西（New Jersey）的隧道，但经过交通流建模和分析，调整交通控制和管理系统，可以使交通流处于高流量的亚稳态，通行能力增加了 20%，从而取消了新隧道的修建[10]. 1994 年 Los Alamos 实验室的科学家实施了交通分析与仿真系统项目研究，基于元胞自动机理论开发了交通分析软件，并成功用于多个城市的交通分析和规划[11]. 2000 年，德国政府支持物理学家 Helbing 领衔，用流体动力学模型研制了宏观流量仿真与道路预测软件包，并对德国与荷兰之间的一条高速公路成功地进行了交通预测，预测与实测结果基本相符[12]. 因此开展交通流的研究，对于揭示各种交通现象的特性本质，为交通规划和交通管理策略的制定提供可靠的理论依据. 2006 年 8 月，我国科技部批准了关于城市交通的首个国家重点基础研究发展计划项目（973 计划）“大城市交通拥堵瓶颈的基础科学问题研究”，这标志着我国也开始逐步进入科学管理交通的行列.

2.2 基于宏观交通流的收费站设置模型

宏观交通流模型将交通流视为由大量车辆构成的可压缩连续流体，引入车辆密度、平均速度、流量等概念来刻画交通流，用流体动力学理论建立交通流的连续性流体动力学方程. 1955 年，英国学者莱特希尔（Lighthill）和惠特汉（Whitham）建立了交通流的一维连续性运动方程[13,14]，考虑路段没有车辆进出时，交通流密度与交通流流量之间满足流量守恒定律时的运动规律. 他们的工作被视为交通流理论研究的重要里程碑. 宏观交通流模型既描述了交通流沿道路空间的分布，又反映了其随时间变化的规律，能较准确地描述交通流的真实行为，因而在实际交通控制中得到了广泛应用.

2.2.1 问题的分析与假设

宏观交通流建模方法起始于对交通行为的总体观察，并将某些大的区域或长时间的交通行为近似看做连续流. 考虑到高速公路主干线交通流密度较大，因此近似地将高速公路交通流视为连续流进行研究. 对于高速公路上的收费站系统的交通流，如果系统运行时间足够长，驾驶员之间及车辆之间的属性差异就可以被忽略，即可以认为接近于平衡状态.

在宏观交通流理论中，建立关于密度和速度的偏微分方程或方程组模型，通过求解这个模型可以研究交通流动力学行为. 为了研究收费站交通流问题，需要结合其交通流特点，构建相应的理论模型. 在宏观交通流理论中流体满足两个基本假设：一是流量守恒，二是速度与密度存在对应关系. 第一个假设使用守恒方程或连续性方程来表示；第二个假设使用速度与密度的关系式来表

示. 将这些方程和基本交通流方程相结合，可以获得某一路段任何时间内的速度、流量和密度. 因此，通过建立相应的模型，了解交通系统.

在收费站中主要存在两个限制系统流量的瓶颈：一是收费亭，二是车辆合流点. 由于通常收费站内的车道数多于高速公路的车道数，车辆离开收费亭回到高速公路时，在合流点可能会出现拥堵. 由于流速受制于容量最小的瓶颈，因此，吞吐量由收费亭和合流点两处的相对流速决定，从而问题简化为研究“问题区域”的吞吐量，即对这两点的交通行为进行建模，而不需要考虑其间的交通行为.

为了防止由于瓶颈带来的拥堵，同时又使收费站系统保持较大的吞吐量，设置最佳的收费亭数量使得在不出现排队情况下系统能够达到最大流量. 对于“最优”的评价，考虑使用两个标准：总的车流量和车辆平均通过时间. 为了使模型更符合实际情况，对收费站和车辆行为做出以下假设：

(1) 车辆通过收费系统是指车辆从 m 车道的高速公路分流到 $n(\geqslant m)$ 车道的收费站，然后再合流回 m 车道的高速公路的过程.

(2) 每个收费亭对应一个收费车道，且任意时间至多只能为一辆车服务.

(3) 车辆驶出收费站的时间不计.

(4) 驾驶员在保证安全的情况下，尽可能快地通过收费站.

(5) 在收费站，对于给定的车流密度，车辆以安全的最快速度行驶.

2.2.2 名词解释与符号说明

分流区域：指收费站的高速公路入口到收费亭之间的 n 条车道的区域.

合流区域：指收费亭与收费站的高速公路出口之间的 n 条车道的区域.

单位时间内的车流量 q：指在单位时间内通过道路上某一点、某一截面或某一条车道的车辆数. 通常通过测量过去一段时间 T 内通过的车辆数 N 来估计当前时刻 t 的车流量 $q(t)$，即 $q \approx N/T$.

车流密度 $\rho(t)$：指单位长度的道路上，某时刻 t 的车辆数.

车流速度 v：指某段时间内通过道路上某一点的多辆车的车速的平均值.

车辆限速 $v_{\max}$：指车辆在高速公路上的最高限制行驶速度.

收费亭的服务时间 τ_i：车辆驶入收费亭 i、付费且驶出收费亭所需要的平均时间.

队列长度 $B(t)$：在时刻 t 等待进入收费亭的排队车辆数或等待驶出收费站的排队车辆数.

用 q_{in} 和 q_{out} 分别表示单位时间内进入和驶出收费站的车辆数；q_{in_i} 和 q_{out_i} 分别表示单位时间内进入和驶出车道 i 的车辆数；q_t 为收费亭的总车流量，q_{t_i} 为收费亭 i 的车流量；q_b 为某瓶颈 b 的最大车流量；$q_{\max}$ 为合流点单车道的最大车流量.

2.2.3 模型的准备

1. 车流连续性守恒方程

假设在单向连续的道路上，沿车流方向依次有断面 1 和断面 2，其间距为 dx，在此路段内没有出口和进口（即路段内没有交通流的增加和减少）. 设车流量 q 为单位时间（每秒）通过垂直于高速公路所有车道的横断面的车辆数，则车流量满足：

$$q=\rho v,$$

其中 v 是车辆平均行驶速度，ρ 为平均车流密度. 该方程被称为基本交通流方程.

取相邻的先后两个时间点 t_1 和 t_2，其时间间隔为 dt. 在 t_1 时刻两断面间的车流密度为 ρ，则在 t_2 时刻车流密度为 $(\rho+d\rho)$. 在 dt 时间间隔内，车流在断面 1 处的车流量为 q，在断面 2 处的车流量为 $(q+dq)$. 根据质量守恒定律，则有

$$\frac{\partial \rho}{\partial t}+\frac{\partial q}{\partial x}=0.$$

这就是交通流守恒方程，此方程表明在没有出入匝道的路段上，车辆数目是守恒的.

2. 交通流的瓶颈约束

收费站的车辆通行能力大小，主要取决于收费站瓶颈处的车流量，即通过系统中的车流量受路径中容量最小的瓶颈的限制.

事实上，如果假设在道路某点 x 处的车流量大于它前面所有瓶颈的车流量，即有

$$q(x)>\max_k\{q_{b_1},\cdots,q_{b_k}\},$$

其中，q_{b_i}是 x 前面第 i 个瓶颈 b_i 的流量. 然而，在稳态模型中道路上所有的点具有相同的车流量，即意味着

$$\max_k\{q_{b_1},\cdots,q_{b_k}\}=q(x),$$

说明 $q(x)\leqslant\max\{q_{b_1},\cdots,q_{b_k}\}$，与假设矛盾，因此假设不成立.

进一步，为了防止由于瓶颈带来的拥塞，同时保持最大的系统吞吐量，则应满足：

$$q_a\leqslant q_{d_{\max}},$$

其中，q_a 是进入系统的车流量（辆/s），$q_{d_{\max}}$是最小容量瓶颈处的车流量，即最大系统吞吐量.

为了说明上面不等式是成立的，考虑系统中车辆队列（积压的车辆）的变化情况. 依据车辆连续性守恒方程，可以得到队列长度 $B(t)$ 的增长速度为

$$\frac{dB(t)}{dt}=\begin{cases}q_a-q_d, & q_a>q_d;\\ 0, & q_a\leqslant q_d,\end{cases}$$

其中，q_d 是离开该系统的车流量. 系统中的车辆队列可能会出现以下几种情形.

情形一：当 $q_a>q_{d_{max}}$ 时，由车辆的增长速度公式可知，系统中的车辆数的增长速度为 $q_a-q_{d_{max}}>0$.

情形二：当 $q_a<q_{d_{max}}$ 时，由车辆的增长速度公式可知，系统中的车辆增长速度为0，系统中的车流量为 q_a.

情况三：当 $q_a=q_{d_{max}}$ 时，类似于情形二，队列不增长，系统的车流量为 $q_{d_{max}}$. 这时的系统流量 q_a 是避免在瓶颈堵塞的最大值，因此，也是最佳的情形.

综上所述，系统中的最大流量取决于流量最小的瓶颈，为了避免堵塞，进入系统的车流量应不大于该瓶颈处的车流量.

2.2.4　模型的建立与求解

1. 最优收费亭数目的确定

收费站系统的两个瓶颈分别为收费亭和合流点，通过对这两点交通行为的建模分析，可以确定在什么时候出现排队现象，并找出在给定的时间内这些瓶颈点的车辆数. 通过对高速公路车流特点的分析，可以得到如下结论：

（1）通过高速公路横截面 dx 的最大车流量 q_{max} 与该点之前的道路结构无关. 换句话说，该截面的最大车流量仅由该点的车道数决定，与分流到该点或合流到该点的车道数无关.

（2）可以改变最大车流量的唯一横截面 dx 在收费亭. 这是通过改变收费亭的数量实现的，因为它直接影响单位时间内通过车辆数的变化.

基于此，将收费站系统内的路段分为两段：第一段是区间 $(-\infty, x_t)$，第二段是区间 (x_t, x_m)，其中 x_t 是公路上收费亭所在的点，x_m 是车辆合流回到 m 车道高速公路的位置点. 对于第一段路段，车辆的到达率 q_a 等于收费站系统的车辆到达率 q_{in}，车辆的离开率 q_d 等于收费亭的总车流量 q_t；对于第二段路段，车辆的到达率 $q_a=q_t$，车辆的离开率 q_d 等于离开收费站的车流量 q_{out}.

由于 $q_t=n/\tau$，因此，收费亭的数目和单个收费亭的平均服务时间 τ 决定了进入车辆合流区域的车流量. 另外，根据上面的结论（1），合流点的最大车流量 $q_{out_{max}}$ 独立于 n 和 q_t，它仅仅是 m 车道高速公路的特性. 下面分析 q_t 如何影响通过合流区域的车流. 依据结论（2），当 $q_t>q_{out}$ 时，车辆队列的增加速度为

$$\frac{dB(t)}{dt}=q_t-q_{out}.$$

此时，车辆队列将不断增加，直到整个合流区被填满. 直到 q_{in} 减少到低于 $q_{out_{max}}$ 这个拥堵增加的过程才会消失，即进入车辆的车流率低于最小的瓶颈容量，高峰时间结束.

为了防止出现这种情况，要求 $q_t \leqslant q_{out}$，使车流经过合流区域而不会造成阻塞. 然而，当 $q_t < q_{out}$ 时，最大车流量不会出现在车辆合流点处. 因此，取 $q_t = q_{out}$ 为最优的车流量，实际上这是 q_t 一个较低的边界值. 由此可得

$$q_{out} = q_{out_{max}} = q_t = \frac{n^*}{\tau},$$

其中 n^* 是 m 车道的高速公路收费站最优的收费亭数目. 通过求解可得

$$n^* = \tau q_{out_{max}}.$$

设 q_{max} 为合流点单车道的最大车流量，则

$$n^* = m\tau q_{max}.$$

2. 收费亭数目的性能分析

当 $n > n^*$，且 $q_{in} > q_{t_{max}}$ 时，合流区域的车辆拥堵的增加速度为 $q_{t_{max}} - q_{out}$. 收费亭继续以其最大服务率 $q_{t_{max}} = n/\tau$ 进行收费，直到整个合流区被填满. 而等待收费的车辆将以速率 $q_{in} - q_{t_{max}}$ 增长. 这个积累过程直到 q_{in} 减少到小于 $q_{out_{max}}$ 才会消失. 因此，收费站系统中总的车辆拥堵量的增长速度是各瓶颈的增长速度之和，即有

$$\frac{dB_a}{dt} = \frac{dB_t}{dt} + \frac{dB_o}{dt} = q_{in} - q_{t_{max}} + q_{t_{max}} - q_{out} = q_{in} - q_{out}.$$

这一结果意味着，该系统的总拥堵量完全取决于车辆的输入速度和它们可以离开的最大速率（即瓶颈处的通行能力）. 因此，只要收费亭不限制系统的车辆通行能力，收费亭处理车辆的速率不影响系统的车流量. 由这种推理可以得到结论：m 条车道的高速公路最佳的收费亭数量 n 应大于或等于 $n^* = \tau q_{out_{max}}$.

3. 高速公路最大车流量的确定

自 20 世纪 50 年代以来，人们提出了多种静态关系模型. 在这些模型中，一般假设速度和流量之间都存在着特定的对应关系，然后结合特定的交通参数值（稀疏自由态、堵塞态和最大流态等特殊状态对应的参数值），通过连续或者分段的函数关系式进行拟合描述. 常用的模型有格林希尔兹（Greenshields）提出的线性关系模型[15]、格林伯（Greenberg）提出的对数关系模型[16]等.

设 v_{max} 为汽车在高速公路上的最大行驶速度（即车辆的最高限速值），ρ_{max} 为高速公路上单位区域的最大车辆数，车流密度为 ρ，车辆行驶速度为 v. 线性关系模型假定速度和流量之间具有如下的线性关系：

$$v = v_{max}\left(1 - \frac{\rho}{\rho_{max}}\right).$$

因此，结合基本交通流方程可以得到：

$$q=\rho v_{\max}\left(1-\frac{\rho}{\rho_{\max}}\right).$$

线性关系模型的适用范围较宽，特别适用于交通流密度适中的情况.

对数关系模型假设速度和流量之间具有如下的对数关系：

$$v=v_{\max}\ln\left(\frac{\rho_{\max}}{\rho}\right).$$

类似地，可以得到：

$$q=\rho v_{\max}\ln\left(\frac{\rho_{\max}}{\rho}\right).$$

该模型比较适合交通流密度较大的情况.

考虑到交通要道的收费站车辆密度较大，这里采用对数关系模型. 为了粗略地估计合流点的最大流量，对对数关系模型进行微分，并令其为0，即

$$\frac{\mathrm{d}}{\mathrm{d}\rho}q(\rho)=v_{\max}\ln\left(\frac{\rho_{\max}}{\rho}\right)-v_{\max}=0,$$

求解得到 $\rho=\rho_{\max}/\mathrm{e}$. 因此当车流密度为 $\rho_{\max}/\mathrm{e}$ 时，车流量为最大，此时有

$$q_{\max}=\frac{\rho_{\max}v_{\max}}{\mathrm{e}}.$$

这里需要说明的是，公式中的 $\rho_{\max}$ 指高速公路单车道的车流密度，则收费站合流点的最大流量 $q_{out_{\max}}=mq_{\max}$.

4. 模型的进一步改进

在前面的模型中，假设车流量在所有车道上都是均匀分布的，即每条车道的平均车流量相同. 然而在实际收费系统中，这种假设并不完全适用. 例如，在新泽西花园州道路上，收费亭前后的车道被分开，车辆被限定在不同的车道上. 此外，收费站的收费通道也可能会被分为低速通道（大卡车）和高速通道（电子收费、摩托车、公共汽车等）. 因此，为了使模型更具一般性，将通过收费站的总流入量和总流出量再划分到每个车道上，以考虑不同车道的车流量差异.

定义 q_{in_i} 为每秒钟进入车道 i 的车辆数，q_{t_i} 为收费亭 i 的车流量，于是有

$$q_t=\sum_{i=1}^{n}q_{t_i},\quad q_{out}=\sum_{j=1}^{m}q_{out_j}.$$

考虑由两车道或多车道合并成一个车道时合流点的车流速度，合流点的平均车流量不可能超过单车道的最大车流量 $q_{\max}$. 当两个车道合流后的车流量超过车道瓶颈处的车流量时，可为每条车道提供一个降低的车流量，以满足两车道之和等于瓶颈处的通过能力. 一种简单的方法是按比例分配：

$$q_1'=q_{\max}\frac{q_1}{q_1+q_2},\quad q_2'=q_{\max}\frac{q_2}{q_1+q_2}.$$

因此，当两个车道的车流量超出它们合流后的车流量时，每条车道的 q_{out_i} 将是

合流后车道总车流量的百分比，并且不超过单车道的最大车流量 q_{max}.

通过这种简单的比例分配，可以将多种类型的车流状况考虑进来. 例如，电子收费车道和其他类型车道的引入将大大简化.

2.2.5 模型的检验与结果分析

宏观交通流模型是将车流进行了连续近似. 为了验证理论模型的合理性，使用元胞自动机模型进行了交通流的微观仿真. 元胞自动机模型将单个车辆看做依明确定义的简单规则行动的独立实体. 由于微观模型可以描述系统中的个体行为，更符合交通流的实际情况. 因此，如果理论模型与仿真模型结果相一致，则将说明理论模型的合理性和正确性.

1. 元胞自动机模型

将收费站用一个二维的网格点（元胞）来表示，每个网格代表的面积比汽车平均宽度和长度略大. 通常元胞有四种类型：空闲元胞表示未被占用的区域，障碍元胞表示收费站的边界点，以及被收费亭占用的元胞和被车辆占用的元胞. 每种类型的元胞都有其规则集，当时间从 t 变化到 $t+\Delta t$ 时（Δt 为仿真时间步长），元胞状态变化的规则如下：

（1）位置更新：所有车辆以定常速度前进，每个时间步长前进 1 个元胞.

（2）变道规则：如果车辆占用元胞的旁边有一个开放的相邻元胞，且在相应的对角线上有开放的元胞，则车辆以给定的概率随机变换车道且向前移动，变道概率的大小受对目标车道期望大小的影响. 如果车辆前面的元胞是空闲的，或者它在收费亭前且收费已经完成，则车辆保持在同一车道上，无需变道.

（3）车辆到达：随着收费站入口处的元胞被释放，车辆以一定概率 p 到达这些元胞. 假设车辆进入收费站系统的过程是参数为 λ 的泊松流，则概率 p 取决于参数 λ.

假设每辆车在收费亭收费所花的时间 x 服从于以 $\tau=3$ 为中心的分布：

$$P(x)=\begin{cases}0.25, & x\in\{2,\ 4\};\\ 0.5, & x\in\{3\}.\end{cases}$$

2. 仿真参数的设置

为了将仿真收集的数据与模型的预测数据结果进行比较，将收费站的几何形状和车辆进入某车道的概率作为仿真的参数. 对相关参数给出以下假设：车辆的平均长度加上与前面车辆的间隔距离大约为 15 ft；在高车流密度情况下，车辆行驶在合流区域的平均速度为 15 mph.

单次仿真运行 1 000 个时间步长，结果返回车辆的总通过量、总等待时间、平均等待时间、系统中每个部分（即收费站的前、中、后）的车辆密度等统计量. 时间步长设为 1 s，则单次仿真运行时间相当于大约 17 min 的交通

运行过程.

将高速公路车道数 m 从1到8之间变化，收费亭数目 n 从 m 到 $4m$ 之间变化（对于较小的 m），或者从 m 到 $3m$ 之间变化（对于较大的 m）. 对于不同规模的高速公路分别对不同的车辆密度进行仿真检验，每种情况重复运行5次，统计量取多次运行结果的算术平均值.

3. 模型预测结果与仿真结果的比较分析

为了考察不同交通状况对收费亭设置的影响，对不同的车流密度进行仿真分析. 低密度、中密度、高密度车流的车流密度分别对应最大流量的15%、50%和80%. 不同交通状况下的模型预测结果和仿真所得结果如表2-1所示，其中 n^* 是再增加收费车道系统性能不会显著变化的最小车道数.

从表2-1可以看出，高速公路车道数与最优的收费亭数目之间大概是1比4的关系. 对于高密度车流的情形 $p=0.85$，两个模型所得的结果基本是一致的，则说明宏观交通流模型作为收费站车辆运行的长期预测值是准确的. 而对于低车流密度情况，由于一般不会达到最大车流量限值 q_{max}，因此在合流区域堵塞的情况几乎不会出现，所以我们的模型并不适用于这种情况.

基于以上分析，我们可以得出以下结论：

（1）在中车流密度情形下，模型预测的结果较好. 当车流密度不低于50%时，模型是适用的. 当车流密度较低时，可以关闭部分收费通道.

表2-1 理论结果与仿真结果的比较

高速公路车道数	不同车流情况的仿真结果			模型结果［n^*］
	15%	50%	85%	
1	1	4	5	3
2	2	6	10	6
3	3	9	13	9
4	4	10	16	12
5	5	13	18	15
6	6	14	21	18
7	7	16	25	21

（2）模型提供的收费通道数是一个较低的上限值，如果实际有更多的收费通道，则不妨碍车辆的通过量.

（3）对于低车流密度的情况，取 $m=n$ 就足够了. 对于中高车流密度的情况，这种设置的性能不是最优的.

4. 仿真精度的检验分析

（1）参数的敏感性分析. 针对系统参数：平均服务时间 τ、收费车道长

度、合流区域长度等，通过改变参数的取值来检验结果对参数的敏感性. 结果表明：当合流区域的长度减少到一个非常低的值时，它的变化对最优车道数影响是显著的，而其他参数的影响并不明显.

(2) 仿真的逼真程度分析. 使用元胞自动机仿真来验证交通流模型的有效性，前提是元胞自动机是现实世界中的收费站交通流的真实描述. 为了验证仿真模型的逼真程度，对 Garden State Parkway 的 Union toll plaza 几个高峰点 7 个小时的数据进行统计，得到车流的通过量为 2 393 辆/h，且此时 $m=13$，$n=5$. 模型的预测结果是 2 530 辆/h，可见对于高车流密度情形下，模拟结果与实际数据基本吻合. 根据仿真结果，该收费站收费亭的设置是不太合理的，仿真模型的最优设置值为 $m=20$，这样可以使平均通过量增加近两倍之多.

2.2.6 模型的推广

在电子付费系统（如快速通道和电子通道）中，可以为车辆安装一个电子装置，当通过收费通道时，通过自动扫描完成付费，在这种情形下车辆只需要稍减速或根本不用减速. 交通流模型和 CA 仿真都可以引入这种“快速通道”. 在交通流理论模型中，快速通道仅被看做是通过收费亭的车流量更高的车道. 因为合流区域的车流量依据车流率的百分比将 $q_{\max}$ 进行分配. 假设快速通道的车流量为 q_f，其他车道的通行量为 q_o，则通过快速通道驶离收费站的车流量为

$$q_f' = q_{out}\frac{q_f}{q_f+q_o}$$

当快速通道的车流量 q_f 远大于其他非快速通道的车流量 q_o 时，快速通道的车流量将接近于系统的流出量. 因此，使用快速通道的车辆应该能够比其他通道的车辆更快地通过收费站.

2.3 收费站的车辆排队模型

排队论通过研究各种服务系统在排队等待中的概率特性，解决系统的最优设计和最优控制问题. 高速公路上的车辆陆续到达收费站，依次接受收费服务，然后离开收费站. 如果到达的车辆不能及时得到服务，就产生了排队现象. 高速公路收费系统是一个典型的排队系统，运用排队理论建立收费站的车辆排队模型，是研究收费亭合理配置问题的方法之一.

2.3.1 名词解释与符号说明

1. 收费站的车辆排队系统

收费站内的车辆排队现象主要有两种情况：如果没有足够的收费亭，车辆

将会在收费亭前形成等待交费的队列；如果有较多的收费亭，当车流量大时车辆驶回车道较少的高速公路时，也会出现车辆排队现象. 因此，可以将整个收费系统 S 分为两部分：收费子系统 S_1（toll system）和车辆合流子系统 S_2（merging system），其结构如图 2-2 所示. 收费子系统是从车辆到达收费站进入交费等候队列，到收费完成后离开收费亭的过程；车辆合流子系统是从车辆离开收费亭到回到高速公路上的过程.

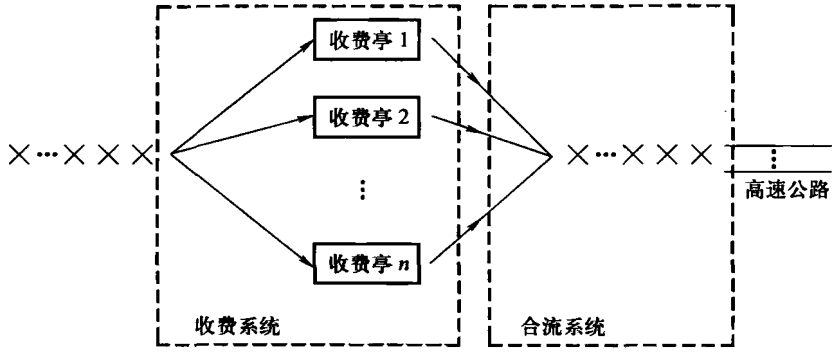


图 2-2　收费站车辆排队系统示意图

2. 排队模型

排队系统主要由三个部分组成：输入过程、排队规则和服务方式，这几部分的差异形成了不同的排队模型. 为了后续描述方便，这里给出排队模型中一些常用名词的解释和符号的说明.

（1）泊松（Poisson）分布. 当顾客的到达满足无后效性、平稳性和普遍性等条件时，顾客的到达形成泊松分布. 一般地，在时间段 $[t_0, t_0+t)$ 内有 $n(n\geqslant 0)$ 个顾客到达的概率 $P_n(t)$ 满足：

$$P_n(t)=\frac{(\lambda t)^n}{n!}e^{-\lambda t},\ n=0,\ 1,\ 2,\ \cdots;\ t>0,$$

其中 λ 为单位时间内平均到达的顾客数. 当顾客到达为泊松流时，两个顾客相继到达的时间间隔服从负指数分布，其均值为 $1/\lambda$.

（2）负指数分布. 当顾客的服务时间 s 服从负指数分布时，其密度函数为

$$f_s(t)=\begin{cases}\mu e^{-\mu t}, & t\geqslant 0,\\ 0, & t<0,\end{cases}$$

其中，μ 为单位时间内服务完成的顾客数，而 $1/\mu$ 表示每个顾客的平均服务时间.

（3）服务强度 ρ. 单位时间内服务完成的顾客数与请求服务的顾客数之比，称为服务强度. 对于单服务台排队系统有 $\rho=\dfrac{\lambda}{\mu}$，对于 n 个并行服务台的排队系统有 $\rho=\dfrac{\lambda}{n\mu}$.

(4) 在 t 时刻系统中有 n 个顾客的概率为 $P_n(t)$. 当系统达到稳态时, $P_n(t)$ 与时间 t 无关, 记做 P_n.

(5) 逗留时间是指一个顾客在系统中的停留时间, 其期望值记做 W.

(6) 等待时间是指一个顾客在系统中的排队等待时间, 其期望值记做 W_q.

2.3.2 模型 1: 单队列多服务台排队系统模型

1. 问题的分析与假设

收费亭数目的多少将会影响车辆通过收费站的时间、收费站的建设和运营成本等. 假设不考虑收费亭的运营成本, 以车辆通过收费站的平均时间作为收费亭优化的目标, 分析不同的收费亭设置时的车辆平均逗留时间, 包括车辆在付费队列中的等待时间、车辆接受服务的时间和付费后离开收费站的时间.

车辆进入收费站后, 将从 n 个收费亭中选择一个接受服务. 如果有空闲的收费亭, 则到达车辆直接进入收费亭进行付费; 如果没有空闲的收费亭, 车辆倾向于到队列最短的收费亭, 假设进入收费站的车辆都会选择最早空闲的收费亭排队等候. 因此, 为了简化收费亭前的排队, 假设所有车辆形成一个队列. 此外, 假设离开收费亭的车辆会因为车辆合流而减缓车速, 但不会影响到收费亭的运行, 即不足以将收费亭出口堵塞. 因此, 将车辆进入收费站、付费、离开收费亭的过程看成是一个单队列多服务台的排队系统.

假设所有车道具有相同的车辆到达率, 车辆到达收费站是一个泊松过程, m 车道高速公路的车辆平均到达率为 λ; 收费亭的服务时间服从相同的概率分布, 即参数为 μ 的指数分布. 因此, 用 M/M/n 排队模型描述车辆集中在收费亭前并交费的过程, 如图 2-3 所示, 用排队模型预测车辆在收费站的等待时间和接受服务的时间.

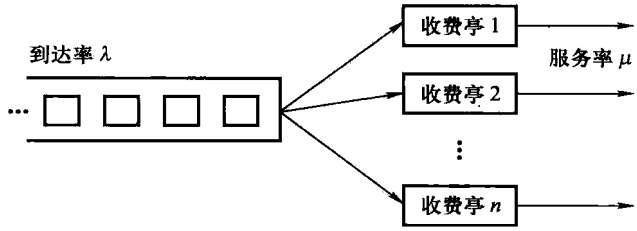


图 2-3 单队列多服务台排队系统

下面确定车辆通过合流子系统的时间. 假设车辆回到高速公路上恢复到最大限速, 则根据交通流理论模型可以确定车辆的密度, 然后依据宏观交通流的流量守恒定律确定车辆通过合流子系统的速度, 进而计算车辆通过合流子系统的时间. 由于 M/M/n 系统的输出过程与输入过程具有相同的分布, 因此, 在服务强度 $\rho<1$ 的情形下, 车辆合流系统的输入服从平均到达率为 λ 的指数

分布.

另外，为了检验理论计算的结果，用仿真的方法模拟真实交通流通过收费站的情况，统计车辆在系统中的时间延迟. 通过理论计算结果与仿真结果的比较，来检验理论模型的正确性.

2. 收费系统的多服务台排队模型

根据排队理论，当时间 $t\to\infty$ 时，系统中有 k 个顾客的概率 $P_k(t)$ 与时间 t 无关，即系统达到平稳状态，此时概率 $P_k(t)$ 记作 P_k. M/M/n 排队系统在平稳状态下，满足下面的平衡方程：

$$\begin{cases}\mu P_1=\lambda P_0,\\ (k+1)\mu P_{k+1}+\lambda P_{k-1}=(\lambda+k\mu)P_k, & 1\leqslant k<n,\\ n\mu P_{k+1}+\lambda P_{k-1}=(\lambda+n\mu)P_k, & k\geqslant n.\end{cases}$$

其中，$\sum\limits_{k=0}^{\infty}P_k=1$，且服务强度 $\rho=\dfrac{\lambda}{n\mu}<1$. 由递推关系可以求得系统状态概率为

$$P_0=\left[\sum_{k=0}^{n-1}\frac{(n\rho)^k}{k!}+\frac{1}{n!}\frac{(n\rho)^n}{1-\rho}\right]^{-1},\quad P_k=\begin{cases}\dfrac{1}{k!}(n\rho)^kP_0, & 0\leqslant k<n,\\ \dfrac{1}{n!}\dfrac{1}{n^{k-n}}(n\rho)^kP_0, & k\geqslant n.\end{cases}$$

当 M/M/n 排队系统达到平稳后，就可以求得系统平均队长、平均等待队长、平均占用服务台数、系统通过能力等几个重要数量指标.

设 W_q 为在系统平稳条件 $\rho<1$ 下单辆车的等待时间，则 W_q 的分布函数为

$$F_{W_q}(t)=1-\frac{P_n}{1-\rho}e^{-n\mu(1-\rho)t},\quad t\geqslant 0.$$

等待时间 W_q 的数学期望为

$$E(W_q)=\frac{\rho}{\lambda(1-\rho)^2}\cdot P_n.$$

设车辆在系统中的逗留时间为随机变量 W，它包括车辆在队列中的等待时间和接受服务的时间. 在平稳状态下，逗留时间 W 的分布函数为

$$F_W(t)=\begin{cases}1-e^{-\mu t}\left[1+\dfrac{P_n}{1-\rho}\mu t\right], & n\mu=\lambda+u,\ t\geqslant 0,\\ 1-e^{-\mu t}-\dfrac{P_n}{(n-n\rho-1)(1-\rho)}\left[e^{-\mu t}-e^{-n\mu(1-\rho)t}\right], & n\mu\neq\lambda+u,\ t\geqslant 0.\end{cases}$$

逗留时间 W 的期望 $E(W)$ 为

$$E(W)=\frac{\rho}{\lambda(1-\rho)^2}\cdot P_n+\frac{1}{\mu}.$$

上面的公式给出了车辆在系统中逗留的时间与收费亭数目（即服务台的个数）n 的关系.

3. 车辆合流系统的交通流模型

格林希尔兹（Greenshields）通过对观测数据研究发现，交通流的速度和车流密度呈线性关系，提出了最早的速度-密度线性关系模型：

$$v=v_f\left(1-\frac{\rho}{\rho_f}\right),$$

其中，v 为车流流速，ρ 为车流密度，v_f 为理论最高速度（即自由流速），ρ_f 为阻塞车流密度.

设 q 为单位时间内通过车道某截面的车辆数，根据车流量与车流速度、车流密度的关系，可以得到

$$q=\lambda=rv\rho,$$

其中，r 为车道数，λ 为单位时间内平均到达的车辆数. 需要说明的是，这里的 ρ 为单车道车流密度，λ 为单位时间内多个车道的总的车辆到达数.

当交通流通过合流区域时，如图 2-4 所示，交通流的速度会发生改变，从而影响到车辆通过过渡区的时间. 假设车辆在过渡区的平均速度为 v_m，过渡区前后交通流的速度、密度、流量分别为 v_1、ρ_1、q_1 和 v_2、ρ_2、q_2，过渡区前后的车道数分别为 r_1 和 r_2.

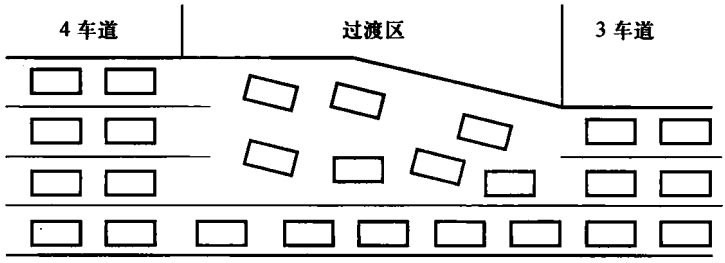


图 2-4 收费站合流区的交通状况示意图

因为总车流量不变，驶入过渡区的车流量与驶出过渡区的车流量相等，即

$$(v_1-v_m)r_1\rho_1=(v_2-v_m)r_2\rho_2.$$

因此，可以得到

$$v_m=\frac{v_1r_1\rho_1-v_2r_2\rho_2}{r_1\rho_1-r_2\rho_2}.$$

对于 m 车道的高速公路，考虑设置有 n 个收费亭的收费站，则有 $r_1=n$ 和 $r_2=m$. 对于 M/M/n 排队系统，车辆离开系统具有无记忆性，离开时间间隔相互独立[17,18]. 系统的输出过程与输入过程具有相同的分布，离开时间间隔服从参数为 λ 的指数分布. 定义 D 为时间间隔 Δt 内离开收费亭的车辆数，于是在时间 Δt 内有 d 辆车离开的概率为

$$P\{D=d\}=\frac{e^{-\lambda\Delta t}(\lambda\Delta t)^k}{d!},$$

其中 λ 为在单位时间内到达收费站的平均车辆数. 由于车辆离开收费系统即

进入车辆合流系统，合流系统的车辆到达率为 λ，因此，$r_1 v_1 \rho_1 = \lambda$.

在给定车辆到达率为 λ 的情况下，对于给定的车流速度，可以确定相应的车流量. 假设车辆离开收费亭进入过渡区之前以舒适加速度 α 匀加速前进，则过渡区前的车道长度 l_{line} 决定了车流速度 v_1，即

$$v_1 = \sqrt{2\alpha l_{\text{line}}}.$$

假设车辆回到高速公路上恢复到公路最大限速速度，即 v_2 等于公路的最大限速，则依据速度-密度的线性关系模型可以确定车流密度 ρ_2.

综上，可以确定车辆通过过渡区的平均速度 v_w. 假设过渡区域的长度为 l_{rec}，则车辆通过合流区域的时间近似为 l_{rec}/v_w.

因此，车辆离开收费亭回到高速公路的平均时间为

$$\sqrt{2l_{\text{line}}/\alpha} + l_{\text{rec}}/v_w.$$

4. 模型的求解与结果分析

根据统计数据，可以计算出收费系统的平均到达率 λ. 不同类型的收费亭具有不同的平均服务率 μ，电子收费亭的平均服务率为每车道 1 200 辆/h，人工收费亭为 350 辆/h，投币收费亭为 550 辆/h[19]. 不同类型的公路具有不同的自由行驶速度 v_f，通常美国平原区域的最高限速为 129 km/h 到 139 km/h，山区为 80 km/h 到 96 km/h. 另外，设道路阻塞的车辆密度 ρ_f 为每车道 402 辆/mile①. 对于 4 车道的高速公路，在设置不同的收费亭情况下，平均等待时间计算结果如图 2-5 所示.

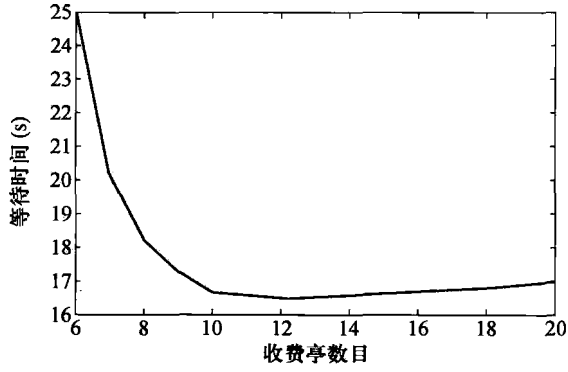


图 2-5　不同收费亭数目与平均等待时间的关系

从图中可以看出，平均等待时间开始时随着收费亭数量的增加而快速减少，但经过最优点之后，又有缓慢的增加. 这与前面的分析是一致的，如果设置太多的收费亭，车辆经过合流点回到高速公路时会出现堵塞情况.

将模型应用于 Garden State Parkway 的收费站，得到不同收费站不同时段

① 1 mile = 1.609 344 km.

的最佳收费亭数目，如表 2-2 所示. 从表中可以看出，在不同时段由于交通流密度的变化，所需要的收费亭数量也随之变化. 因此，可以通过收费亭工作车道的动态调整来减少收费亭的运行费用.

表 2-2 不同时段的最佳收费亭数目

收费亭 \ 时间段		6—7am	7—8am	8—9am	9—10am	3—4pm	4—5pm	5—6pm	6—7pm
Raritan	收费亭数	20	20	20	17	10	11	11	9
	等待时间(s)	16.79	17.00	16.72	16.51	16.53	16.53	16.52	16.53
Union	收费亭数	13	17	15	13	12	14	14	11
	等待时间(s)	16.52	16.51	16.51	16.52	16.52	16.51	16.51	16.52
Essex	收费亭数	10	14	13	10	12	12	12	11
	等待时间(s)	16.53	16.52	16.52	16.53	16.52	16.52	16.52	16.52
Bergin	收费亭数	5	10	11	9	9	9	10	9
	等待时间(s)	16.65	16.51	16.51	16.53	16.53	16.53	16.52	16.53
Hillsdale	收费亭数	4	7	9	5	5	7	8	7
	等待时间(s)	16.73	16.56	16.54	16.65	16.67	16.60	16.55	16.58

综合以上分时段的求解结果，对不同收费站的最佳收费亭设置提出建议，如表 2-3 所示. 考虑到 Raritan 和 Union 两个收费站高峰时段车流量较大，因而应设置较多的收费亭. 因此，也建议对这两个收费亭进行改造，引入效率较高的其他形式的收费亭，如 ETC 等.

表 2-3 不同时段的最佳收费亭数目

	高速公路车道数	最佳收费亭数目	平均等待时间(s)
Raritan	6	16	19.42
Union	10	17	23.52
Essex	6	14	18.52
Bergin	6	14	18.54
Hillsdale	6	9	15.16

5. 模型的检验

理论模型的计算是在稳态假设下得到的结果，需要进一步检验其合理性. 一般可以考虑将理论模型与仿真模型的结果进行比较. 对于排队系统理论模型的检验，通常有两类方法：

（1）通过对排队系统的仿真，将被仿真系统的状态变化过程与实际的交通状况进行比较，或将仿真的统计结果与实际交通流的统计数据进行比较.

（2）通过交通流的微观仿真，将基于理论模型的求解结果与基于微观仿真得到的结果进行比较. 通过比较分析，一方面检验理论模型中假设、参数设定等的合理性，另一方面可以检验求解结果的一致性.

这两种思路和方法分别在多个单服务台排队系统模型、宏观交通流模型中都有所体现，此处不再详述.

2.3.3　模型2：多个单服务台排队系统模型

1. 问题的分析与假设

在理想情况下，车辆进入收费站时会选择排队最短的收费亭. 然而，实际情况是如果公路上车辆较多，车辆不能变换太多的车道进入最短的队列，而是被限制到所在车道直接对着的队列或邻近的队列，如图2-6所示. 图中不同图案的方块代表单辆车使用该收费亭的可能性，方块的大小表示使用该图案收费亭的可能性的大小.

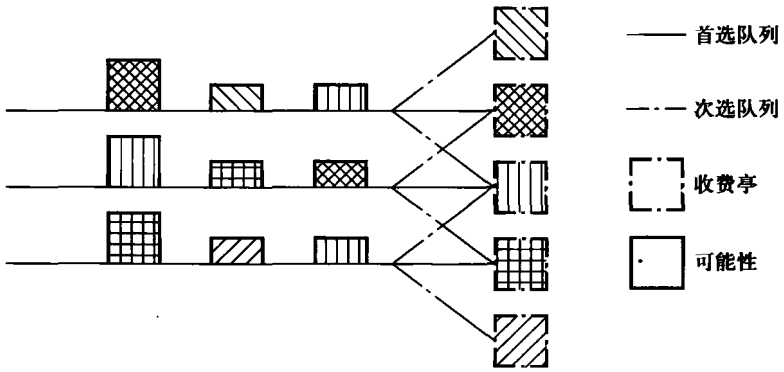


图2-6　车辆随机选择收费亭的排队示意图

根据分析，假设驾驶员多数会使用正前方的收费亭，少数会以相等概率选择左边或右边邻近的收费亭，以较小的概率使用较远的收费亭. 在所有队列的长度差异不大的情况下，车辆最可能径直进入它正前方的队列，这种行为类似于单服务台排队系统. 因此，把收费站的收费子系统看作是由几个并行的单服务台队列组成的多服务台的排队系统. 类似地，车辆合流子系统也可以看做是同类型的排队系统. 车道合流子系统是从车辆离开收费亭到回到高速公路上，并恢复到进入收费站前的行驶速度的过程.

假设车辆的到达服从泊松流，车辆的到达时间间隔为独立的指数随机变量；驾驶员倾向于选择加入邻近最短的队列，因此，在正常运行中，每个收费亭的队列长度是一样的；车辆的服务时间服从负指数分布，采用先到先服务的服务规则. 假设高速公路的车道数为 m，且不区分不同车道；收费亭的个数为

n，且单个收费亭的服务率为 μ_1. 在稳定运行的情况下，车辆的服务率应大于车辆的到达率，否则在一段时间后将会出现无限长的队列.

从图 2-6 可以看出，对于不同的车道，车辆通过收费亭的概率之和是一样的. 因此，从长期运行的情况看，可以看做是 n 个收费亭的服务能力均分到不同的车道上，收费亭对每个车道的服务率为 $\mu_1 \cdot n/m$. 因此，每个高速公路车道可以看做是拥有一个“超级服务台”的队列，该服务台的服务率为 $\mu_1 \cdot n/m$. 类似地，可以用同样的方法确定车辆合流子系统中服务台的服务率.

在建立排队模型的基础上，基于排队理论和马尔可夫过程可以确定排队系统的运行指标. 这里以车辆在通过收费站所消耗的时间最小为优化目标，以确定最优的收费亭数目.

另外，考虑到理论模型的计算结果是在稳态的假设下得到的，需要进一步检验理论模型的合理性. 基于 Matlab 的 Simulink 工具箱实现排队系统的随机仿真，并比较仿真系统中的交通流与实际交通流的一致性.

2. 多个单服台排队模型的建立

（1）收费子系统的平均到达率和服务率

假设车辆到达收费子系统的时间间隔服从于负指数分布，在不区分不同车道的情况下，设每车道单位时间内平均到达的车辆数为 λ_1. 收费时间为车辆反应时间与服务时间之和. 车辆反应时间为驾驶员踩油门并行驶到收费亭前的这段时间，服务时间是指从车辆停车接受服务至车辆开动的这段时间. 假设每个收费亭完成一次收费的平均时间为一个固定的数值 t，则单个收费亭的服务率为 $\mu_1 = 1/t$. 考虑到在交通繁忙时，车辆存在从停止状态启动行驶到收费亭前的过程，因此将车辆反应时间 r 引入服务时间，得到

$$\mu_1 = \frac{1}{t+r}.$$

将 n 个收费亭的服务能力平均到单个高速公路车道，可得收费亭对每个高速公路车道的平均处理率为

$$\frac{\mu_1 \cdot n}{m}.$$

（2）车辆合流子系统的平均到达率和服务率

由于车辆离开收费子系统 S_1 后立即加入到车辆合流子系统 S_2 中，因此子系统 S_2 的平均到达率等于 S_1 的离开率. 对于 m 条车道的高速公路上有 n 个收费亭的收费站，如果 $m\lambda_1 \geqslant n\mu_1$，则所有 n 个收费亭都处于忙的状态，此时，收费子系统总的离开率为 $n\mu_1$. 假设收费站之后的高速公路的车道数为 k，由于车辆从收费子系统 S_1 出来必然进入 k 车道的高速公路，因此，子系统 S_2 每车道的平均到达率为 λ_2，满足：

$$k\lambda_2 = n\mu_1 \quad \Rightarrow \quad \lambda_2 = \frac{n\mu_1}{k}.$$

反之，如果 $m\lambda_1<n\mu_1$，则总的离开率为 $m\lambda_1$，此时

$$k\lambda_2=m\lambda_1 \quad \Rightarrow \quad \lambda_2=\frac{m\lambda_1}{k}.$$

假设车辆在高速路上的行驶速度为 v，v 与收费站无关，而与高速公路上的交通状况有关. 车辆离开收费子系统是从速度 0 到速度 v，这里使用这段时间的平均速度. 用平均速度除以车辆之间的安全距离，则为单位时间内通过的车辆数，即平均服务率 μ_2. 假设车辆的安全距离为 l（l 为相邻两辆车的车头间距），且 l 的大小依赖于车辆行驶速度，因为车辆高速行驶时所需要的安全距离远大于低速时的安全距离. 因此，平均服务率 μ_2 为

$$\mu_2=\frac{v}{2l}.$$

因此，车辆合流子系统也可以看做是 k 个单服务台排队系统组成，每个服务台的车辆到达率为 λ_2，平均服务率为 μ_2. 考虑到模型的适用性，这里的 k 可以等于或不等于 m.

（3）车辆在系统中的逗留时间

基于排队理论，系统的输入过程是泊松流，服务时间服从负指数分布，顾客源无限的单服务台排队系统，顾客在系统中逗留时间满足：

$$W=\frac{1}{\mu-\lambda},$$

其中 λ 和 μ 分别为系统中顾客的平均到达率和服务台的平均服务率.

车辆在系统中的逗留时间 W 是通过两个系统的时间之和，即 $W=W_1+W_2$. 则

$$W=W_1+W_2=\frac{1}{\frac{n\mu_1}{m}-\lambda_1}+\frac{1}{\frac{v}{2l}-\lambda_2}.$$

综上可得，如果 $m\lambda_1<n\mu_1$，则

$$W=W_1+W_2=\frac{1}{\frac{n\mu_1}{m}-\lambda_1}+\frac{1}{\frac{v}{2l}-\frac{m\lambda_1}{k}}.$$

由于在多数情况下高速公路收费站前后的车道数相同，即 $k=m$，因此，上式可以简化为

$$W=W_1+W_2=\frac{1}{\frac{n\mu_1}{m}-\lambda_1}+\frac{1}{\frac{v}{2l}-\lambda_1}.$$

在系统长时间的运行中，通常车辆服务率大于车辆的到达率，否则将会出现无限长的队列. 因此，上面 W_1 和 W_2 的表达式中分母应大于零，即

$$\frac{n\mu_1}{m}>\lambda_1，且\frac{v}{2l}>\lambda_1.$$

如果 $m\lambda_1 \geqslant n\mu_1$，则 $\lambda_2=\dfrac{n\mu_1}{k}$，此时为了避免合流区域出现无限长的队列，需要满足合流子系统的到达率小于系统的处理能力，即

$$\lambda_2=\frac{n\mu_1}{m}<\mu_2=\frac{v}{2l}$$

因此，可以确定收费亭数目 n 的可行域为

$$\left(\frac{\lambda_1 m}{\mu_1},\ \frac{mv}{2\mu_1 l}\right).$$

也就是说，收费亭的处理能力要大于收费站系统车辆的到达率，且小于收费站系统合流区域的通行能力.

3. 模型的求解及结果分析

为了考察不同交通状况下最佳收费亭的设置，将交通流划分为低密度、中密度和高密度 3 个区间. 对于中、低密度交通流，收费亭的服务时间忽略车辆的反应时间 r，而对于高密度交通流，通过收费亭的时间则包括 r. 取收费亭平均单次收费时间 $t=5.5$ s，车辆反应时间 $r=2.5$ s. 考虑到在不同车流密度情况下车辆通过车辆合流区域的速度会受到影响，且安全车距也随着车辆行驶速度的降低而减少. 因此，对于不同的交通状况，设置相应的参数. 不同交通状况下模型参数的具体取值如表 2-4 所示.

表 2-4 模型的基本参数取值

交通流密度	车辆到达率 λ_1 (辆/min)	收费亭服务率 μ_1 (辆/min)	车辆行驶速度 v (km/h)	安全车距 l (m)
低密度	0.5 ~ 5	11	60	24
中密度	5 ~ 10	11	46	18
高密度	10 ~ 15	7.5	32	12

当收费子系统和合流子系统都满足车辆到达率小于系统的服务率时，从车辆在系统中的逗留时间 W 的计算公式可以看出，在模型的基本参数确定之后，对于特定车道的高速公路，逗留时间 W 随着收费亭数目 n 的增加而降低. 如果用最小逗留时间作为优化目标，则得到的最佳收费亭数目是 n 的上边界. 因此，这里采用收费亭数目增加所带来的“边际效用”来衡量收费亭增加所带来的效果，即考察每增加一个收费亭所带来的车辆在系统中逗留时间的减少量. 当这个减少量小于某个阈值时，则认为没有必要再增加新的收费亭. 这里设阈值为 2.5 s.

在不同交通流、不同高速公路车道数的情形下，通过公式计算车辆在系统中的平均逗留时间. 在其他条件不变的前提下，通过变换收费亭的数目，寻找

最优的数目 n. 对于不同的交通状况和不同的高速公路车道数，最优的收费亭数目如表 2-5 所示.

表 2-5　不同情形下的最佳收费亭数目

	高速公路车道数	3	4	5	6
不同车流密度时的收费亭数	低密度	4	7	9	10
	中密度	5	8	9	11
	高密度	7	8	12	15

从表 2-5 可以看出，不管处于什么样的交通流密度情况，最优的收费亭数目总是大于高速公路车道数. 在低密度车流的情况下，最优的收费亭设置的平均等待时间与 $n=m$ 时的平均等待时间仅仅相差 2 s. 在中密度车流的情况下，这个差异大约等于 15 s，因此，设置多于高速公路车道数的收费亭个数是合理的. 在高密度车流情况下，每个车道上只设置一个收费亭将会导致无限长的等待队列.

为了分析求解结果对交通流量的灵敏性，以 6 车道高速公路为例，分析最优收费亭设置随交通流量的变化情况. 在三种交通流密度情况下，求解车辆的平均等待时间 W 随收费亭数目 n 的变化关系的图表，即对于每个确定的 λ，计算不同收费亭数目时的平均等待时间 W，通过线性分段插值得到 W 随 n 变化的曲线. 高密度交通流状况下的求解结果如图 2-7 所示.

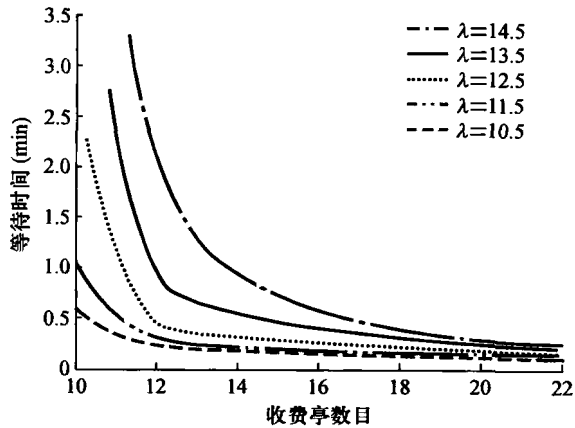


图 2-7　平均逗留时间与收费亭数目间的关系

计算结果表明，随着交通流密度的逐渐增大，收费亭数目 n 的可行域逐渐变小. 这是由于太少的收费亭将会带来 S_1 系统中无限长的等待. 另外，由于所考察的车流密度并没有超过 S_2 系统的通行能力，所以对 n 的制约作用并没有体现出来. 在低密度交通流情况下，取 $n=5$ 和 18 时的平均等待时间的差异仅有2 s；在中密度的交通状况下，收费亭数目的变化会带来较高的等待时间

的变化.

但是，设置太多的收费亭会造成不必要的浪费. 为此建议对于经常使用的公路，收费亭的数目应该取算法得到的所有最优数目的最大值，并且随交通状况的变化调整合适的收费亭数目. 对于低密度的交通流情况，每车道设一个收费亭可以减少收费站的建设和运行成本，因为减少的等待时间不能证明需要更多的收费亭.

4. 模型的仿真检验

(1) 仿真流程设计

由于收费排队系统的状态（即系统中的车辆数）只有在车辆到达、离开收费子系统和合流子系统等时刻发生变化，通常采用事件扫描法来实现仿真过程的推进. 采用事件步长法构造仿真模型，以预定的事件出现作为确定仿真步长的依据，事件不出现，系统不更新. 基于事件扫描的系统仿真流程如图 2-8 所示. 收费站排队系统的状态变化主要有三类事件：车辆的到达事件、车辆开始接受服务事件和车辆离开事件. 通常用事件表来记录各种事件的发生，不同类型事件的处理分别由相应的子程序来完成. 在初始子程序中设置系统的初始状态，包括收费亭空闲、确定第一辆车到达系统的时间，并将该到达事件放入事件表中.

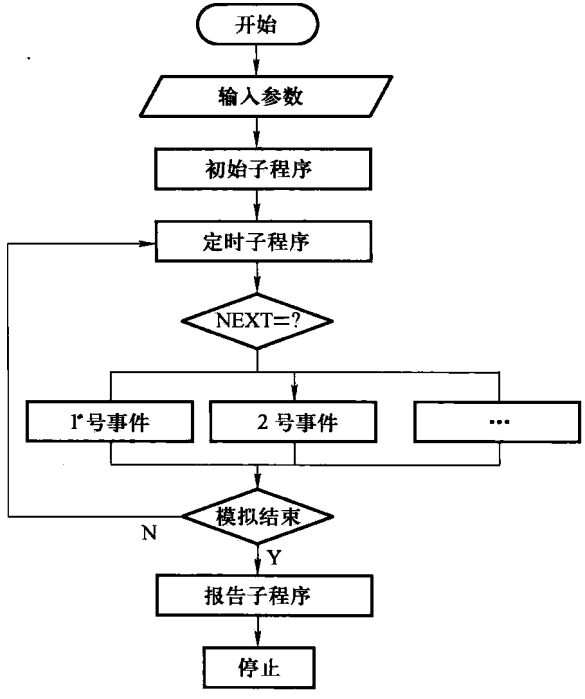


图 2-8 面向事件的系统仿真流程图

车辆到达事件子程序完成车辆到达事件的处理. 首先，根据系统的当前状态，安排下一个到达事件；然后，根据收费亭的状态确定操作方式. 如果收费亭当前状态是忙，则队列长度 L 加 1，返回主程序；如果收费亭当前状态是空闲，则进行收费服务，并记录服务时间，安排服务结束事件，返回主程序.

车辆离开事件子程序完成收费服务结束事件的处理. 若车辆队长 L 为零，则把收费亭的状态置为空闲，返回主程序；否则，安排下一车辆离开事件，返回主程序.

(2) 仿真模型的建立

基于以上仿真流程，可以采用 Matlab、C++等编程语言完成仿真程序的编写和调试. 这里采用一种便捷的方式，使用 Matlab 中的 Simulink 工具箱来完成仿真模型的实现. Simulink 是一个用来对动态系统进行建模、仿真和分析的软件包，它提供了一种图形化的交互环境，不需要编写代码. 通过使用 Simulink 的专用功能模块程序包，用户可以迅速地对系统进行建模、仿真与分析.

根据上面的车辆收费子系统排队模型，车辆到达该系统的过程服从泊松流，该系统是由 m 个并行的单服务台组成的排队系统，该单服务台排队系统的平均到达率为 λ_1，平均服务率为 $n\mu_1/m$；车辆合流子系统也是由 m 个并行的单服务台组成的排队系统，该单服务台排队系统的平均到达率为 λ_1，平均服务率为 $\mu_2 = v/2l$.

考虑到多个并行单服务台排队系统的差异不大，且车流是均匀分布在不同的单服务台排队系统中，为了便于进行仿真计算，可以用一个两服务台串联系统来实现系统的仿真. 图 2-9 是两个串联的 M/M/1 排队系统的 Simulink 模型. 通过对两个串联 M/M/1 系统的分析，获得系统整体变化过程.

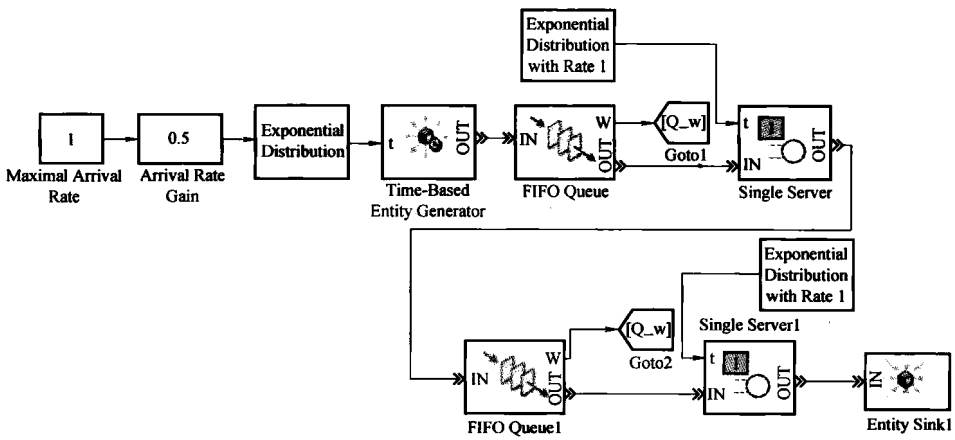


图 2-9　串联的 M/M/1 排队系统仿真模型

(3) 仿真结果分析

为了得到车辆收费系统在稳态时的结果，不考虑仿真前 10 分钟的数据结果，使用系统达到稳态后的数据. 以 6 车道高速公路、系统车辆到达率 $\lambda = 10.5$ 的情况为例，考察车辆的平均等待时间随收费亭个数的变化，理论计算和仿真结果如图 2-10 所示.

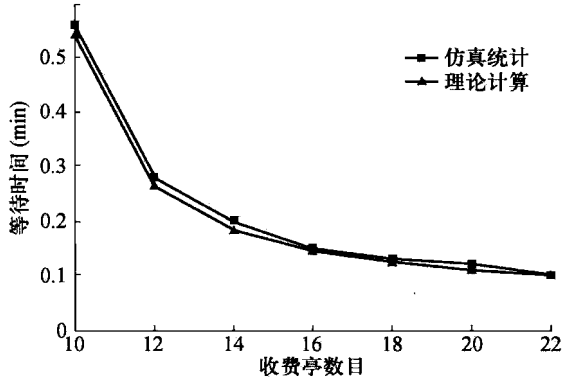


图 2-10 理论模型与仿真模型的结果比较

从图中可看出，理论模型与仿真模型的结果几乎是相同的，仿真模型的车辆平均等待时间略小于理论模型的结果. 考虑到在实际中并不是所有的车辆都会服从先到先服务的假设，因而会导致实际的排队等待的时间会大于理论值.

2.4 基于安全距离的车辆跟驰模型

当道路上车流密度很大时，车辆间距很小，任一辆车的车速都会受到前面车辆速度的影响. 驾驶员只能按前车的速度来控制自己的车速，这种状态被称为非自由行驶状态，也叫跟驰行驶. 跟驰理论运用系统动力学的方法，研究在受限制的单车道上行驶的车队中前车的运行变化引起后车的反应问题，并应用数学模型表达跟驰过程中所发生的各种状态，它是微观交通流仿真的重要理论基础.

自 20 世纪 50 年代以来，研究者基于对驾驶员行为不同角度的认识，提出了多种跟驰模型. 一般将跟驰模型分为以下几类：刺激-反应模型 (Stimulus-Response Model)、安全距离模型 (Safety Distance Model)、驾驶心理模型 (Psycho-Physical Model) 以及基于人工智能的模型 (Artificial Intelligence based Model) 等. 刺激-反应模型的基本原理是驾驶员试图与前车保持一致，即只要有前车的刺激，后车就会对此做出反应. 安全距离模型以避免碰撞为目标，建立速度与距离变化之间的函数关系. 驾驶心理模型也称反应点模型 (Action Point Model)，使用一系列阈值和期望距离体现人的感觉和反应. 在基于人工

智能的车辆跟驰模型中，驾驶员多被视为一个复杂的非线性系统，通过应用模糊理论和人工神经网络等方法，尝试更精确地模拟驾驶员的行为.

2.4.1 问题的分析与假设

车辆通过收费站的行驶过程可以概括为：车辆首先减速行驶进入收费站，然后在收费车道排队等待，交费后离开收费亭并加速行驶，通过合流点后重新回到高速公路. 因此，收费站的车辆跟驰模型不仅要考虑前导车对跟驰车行为的影响，还要考虑收费站的合流点、收费亭、分流点等对车辆行为的影响.

安全距离模型是一种简单实用的跟驰模型，这类模型大多数情况只需知道驾驶员将采用的最大制动减速度. 在建立跟驰模型前，对车辆行为和外在条件等做如下假设：

(1) 收费站前后的公路上没有障碍物，道路平直，不允许超车. 除了在分流点和合流点处，车辆不变换车道.

(2) 车辆进入收费站时，穿过公路车道和收费亭车道之间的过渡区在瞬时完成，且车辆选择进入最短的队列.

(3) 在前方车辆较远时，驾驶员采取自由驾驶，并尽可能达到所允许的最大行驶速度. 设高速公路最大限速为 v_{max}.

(4) 跟驰行驶时，驾驶员保持合理的安全车距 G 以避免碰撞.

(5) 影响车辆行驶方式变化的外部条件主要是车辆进入收费站的分流点和车辆驶离收费站时的合流点. 驾驶员能够正确地估计距分流点的距离.

(6) 所有驾驶员具有相同的预期反应时间 Δt 和非预期反应时间 γ.

(7) 不对车辆进行区分，即所有车辆具有相同的车长、加速度和减速度等. 一般情况下车辆以定常加速度 a_+ 和减速度 a_- 进行加速和减速行驶，紧急制动时减速度为 a_b.

(8) 车辆到达收费亭时是完全停止的，车辆在等待队列中加速或减速到某个位置所需要的时间取决于它之前的队列长度.

(9) 所有收费亭的服务时间服从均值为 $1/s$、方差为 σ 的正态分布. 设高速公路某时刻 t 的单车道的车流量为 $\rho(t)$.

车辆在收费站延误时间的长短描述了由于收费站的存在造成车辆在经过时的延误，它反映了收费站的交通条件的好坏. 收费亭的设置情况将影响车辆的延误时间，同时也会影响收费站的建设与运营成本. 综合考虑两方面的影响，设最优的收费亭配置为使得车辆的平均延迟时间最小时的情形，如两种不同的收费亭方案得到相似的延迟时间，则选择收费亭数目较少的方案.

2.4.2 名词解释与符号说明

为了便于描述，建立收费站的位置坐标并设定收费亭的标号以及车辆的描

述. 相关定义如下：

(1) 收费站（图 2-11）. 设置有 n 个收费亭的收费站用空间 $[-d, d]\times\{1, \cdots, n\}$ 表示，其中 d 为半个收费站的长度. 集合 $\{0\}\times\{1, \cdots, n\}$ 的元素为收费亭，记第 i 个收费亭 $\{0\}\times\{i\}$ 为 τ_i. 子空间 $[-d, 0)\times\{1, \cdots, n\}$ 为进入区域，$[0, d)\times\{1, \cdots, n\}$ 为离开区域. 收费站前的公路路段用 $(-\infty, -d)\times\{1, \cdots, m\}$ 表示，之后的公路路段用 $(d, \infty)\times\{1, \cdots, m\}$ 表示，其中 m 为高速公路的车道数目. 集合 $\{1, \cdots, m\}$ 和 $\{1, \cdots, n\}$ 分别表示公路车道和收费站车道. 根据实际情况，这里设 $m\leqslant n$.

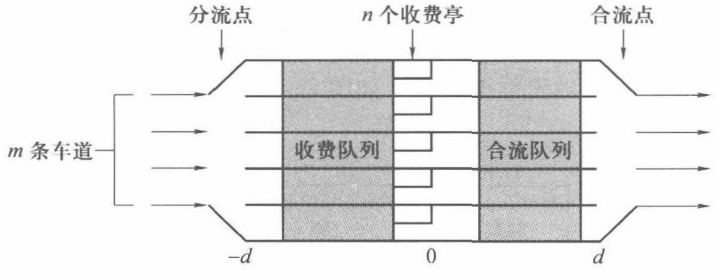


图 2-11 收费站

(2) 合流点与分流点. 车辆从高速公路驶入收费站的位置为分流点，其坐标为 $-d$；车辆从收费站回到高速公路的位置为合流点，其坐标为 d.

(3) 车辆 C 用 4 元组 (L, a_+, a_-, a_b) 和位置函数 $p(t)=(x(t), k(t))$ 表示. 其中，$x(t)$ 是 t 时刻车辆 C 的车头所在的位置，$x(t)$ 是时间 t 的连续函数. $k(t)$ 是 t 时刻车辆 C 所在车道（收费站或公路）的标号. L 是车辆 C 的长度（单位：m），a_+是舒适加速度，a_-是舒适减速度，a_b 是最大制动减速度，加速度的单位为 $\mathrm{m/s^2}$.

(4) 安全距离 G. 车辆在给定的速度下行驶时，通常会和距它最近的障碍物保持一定的距离，如果车辆以最大制动速度 a_b 制动，在到达障碍物之前停下的距离称作安全距离 G. 车辆前方 G 以外的地方为安全范围，车辆可自由行驶.

(5) 反应时间. 车辆的反应时间是发现需要调整加速度与实际调整时的时间差. 反应时间可分为预期反应时间和非预期反应时间. 预期反应时间 Δt，指从观察到预期的事件（红绿灯的变化、收费亭等）到做出反应所用的时间；非预期反应时间 γ，指从观察到非预期的事件（突然停车）到做出反应（刹车、加速、改变方向等）所用的时间.

(6) 仿真步长 Δt. 仿真过程中仿真时钟向前推进的时间间隔. 这里考虑到驾驶员的反应时间，仿真以等时间步长 Δt 往前推进.

2.4.3　基本的跟驰模型

1. 车辆跟驰行驶的一般表示

在直线行驶、无超车情况下，跟驰车辆驾驶员的反应过程可以归纳为感知、决策和控制三个阶段. 在感知阶段，驾驶员通过视觉、听觉和触觉等来感知道路交通环境因素和车辆性能、状态因素；在决策阶段，驾驶员对感知信息进行分析判断，决定采取相应的操控来保证汽车安全行驶；在控制阶段，驾驶员依据判断决策来操控汽车或者通过汽车信号传递给其他道路使用者. 跟驰车驾驶员通过加速、减速及制动来控制汽车的纵向运动. 从控制论的角度来看，单个车辆的跟驰过程是一个不可分解的持续调整反馈控制系统，跟驰理论的一般形式可用传统控制理论框图表示，如图 2-12 所示[20].

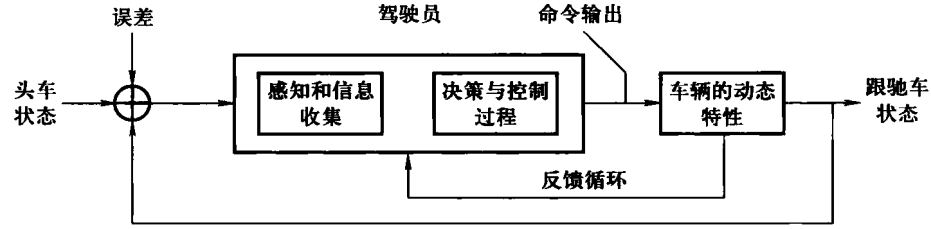


图 2-12　车辆跟驰过程框图

2. 基于安全距离的跟驰模型

在 20 世纪 70 年代以后，欧美发达国家的交通工作者对车辆跟驰模型进行了大量研究，其中最富有成果也最基本的应算是基于安全车间距的车辆跟驰模型[21]. 最初的安全距离跟驰模型是由 Kometani 和 Sasaki[22] 提出的，其基本假设是：驾驶员应在不能完全判断前车运动的情况下，保持合理的安全距离 G（即前车车尾到后车车头的距离，一般也被视为足够的刹车距离）以避免碰撞. 之后有学者提出了很多更为复杂的安全车距模型，主要采用两种方法来保证其安全间距[23]：一是紧急刹车时能够有效地防止碰撞；二是在仿真过程中，让跟随车与前车始终保持一定的间距. Pittsburgh 大学 Bullen 开发的 Pitt 跟驰模型的基本思想是：在仿真中让后车与前车始终保持一定的间距 G. 在 Pitt 模型的基础上，NETSIM 跟驰模型和在其基础上建立的 CARSIM 模型，将跟驰距离要求放松为：仅在紧急刹车时能够有效地防止碰撞.

NETSIM 模型是典型的防止碰撞模型，它的算法逻辑可描述为：当仿真过程向前推进一步，前车属性刷新并行驶到一个新的位置，跟驰车辆推进到一个位置，满足当前车以最大的减速度刹车时，后车也能及时刹车从而避免碰撞. 假设前车从 $t+T$ 时刻开始以最大减速度刹车，最后停止在 x_l' 处，其中 T 为时间步长. 后车从 t 到 $t+T$ 以加速度 a_f 加速，从 $t+T$ 到 $t+T+r$ 时刻维持速度不变，然后开始以最大减速度刹车，最后停止在 x_f' 处，其中 r 为驾驶员的反应时

间. 两车运动和避免碰撞的示意图见图 2-13. 图中 v_l 和 x_l 为 $t+T$ 时刻前车的速度和位置；v_f^i 和 x_f^i 为 t 时刻后车的速度和位置；v_f 和 x_f 为 $t+T$ 时刻后车的速度和位置；Δs_f 为从 t 时刻到 $t+T$ 时刻后车的行驶距离；Δr 为后车在反应时间范围内行驶的距离；s_l^T 和 s_f^T 为前后两车刹车后的运行距离；L 为前车的长度.

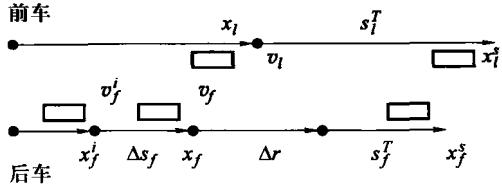


图 2-13 跟车运动过程示意图

从图 2-13 中可以看出，避免碰撞的必要条件为 $t+T+r$ 时刻和最终两辆车停止的时刻，两车均没有发生碰撞，则有

$$x_l+v_l r-\frac{d_l r^2}{2}-x_f-\Delta r\geqslant L,$$

$$x_l+s_l^T-x_f^i-\Delta s_f-\Delta r-s_f^T\geqslant L.$$

其中前车刹车运行距离为 $s_l^T=v_l^2/(2d_l)$，后车刹车运行距离为 $s_f^T=v_f^2/(2d_f)$，式中 d_l 和 d_f 分别为前后两车的减速度. Δs_f 和 Δr 分别依据仿真步长和驾驶员的反应时间进行计算，因此，基于约束条件可以确定后车的加速度 a_f.

车辆属性按以下两个动力方程进行计算，即有

$$v_f=v_f^i+a_f T,\quad x_f=x_f^i+v_f^i T+\frac{1}{2}a_f(T)^2.$$

2.4.4 单车道车辆跟驰模型

1. 模型的优化目标

车辆通过收费站的时间延迟，可以用没有收费站时车辆驶过相同长度的高速公路所用的时间，与有收费站时车辆通过收费站所用的时间之差来表示. 对于某车辆 C，用 x_o 和 t_o 分别表示从最大限速 $v_{\max}$ 开始减速以驶入收费站的位置和时刻，用 x_f 和 t_f 表示 C 驶离收费站回到公路上且车速重新达到 $v_{\max}$ 的位置和时刻. 设时间延迟 T 为车辆由于通过收费站所带来的多余时间消耗，则

$$T=t_f-t_o-\frac{x_f-x_o}{v_{\max}}.$$

因此，收费亭最优设置的目标是使得时间 T 较小.

2. 安全车间距的确定

车辆 C 的安全距离 G 取决于它自身的速度、前车 C' 的速度以及制动减速度等. 设 C 和 C' 的当前速度为 v_1 和 v_2，假设车辆 C' 尽最大可能刹车，即采用制动减速度 a_b 进行减速，则 C' 在 v_2/a_b 时间内停下，且在这段时间内的行驶距离为

$$v_2\frac{v_2}{a_b}-\frac{1}{2}a_b\left(\frac{v_2}{a_b}\right)^2=\frac{v_2^2}{2a_b}.$$

如果车辆 C 经过反应时间 γ 后开始紧急刹车，即它也采用最大制动减速度 a_b 进行减速，则它停下来需要的时间为 $\gamma+v_1/a_b$，行驶距离为

$$\gamma v_1+\frac{v_1^2}{2a_b}.$$

因此，在过去的时间内，车辆 C 和 C'两车之间的间距减少量为

$$\gamma v_1+\frac{v_1^2-v_2^2}{2a_b},$$

这个距离也是避免碰撞时两车的最小距离. 再考虑到车辆 C'的长度 L，所以 C 和 C'的最小距离，即安全车距为

$$G=L+\gamma v_1+\frac{v_1^2-v_2^2}{2a_b}.$$

如果最近的危险是障碍物，如车辆合流点等，此时车辆 C 会保持一个允许正常刹车的安全距离，以舒适的减速度接近，而不是紧急刹车. 即采用制动减速度 a_-，而不再是 a_b. 另外，此时车辆 C 是对预期事件的反应，反应时间为 Δt. 由于障碍物的长度和速度都是0，所以安全距离为

$$G=\Delta t v_1+\frac{v_1^2}{2a_-}.$$

3. 车辆的加速模型

车辆在收费站的行驶过程包括减速、排队、加速等多项活动，且车辆行为受前方车辆、收费亭、合流点等多个因素的影响. 下面对车辆在通过收费站的过程中处于不同情况下的加速度设置分别进行描述.

(1) 如果前车或障碍物与车辆 C 的距离大于安全距离，则此时车辆自由行驶. 在自由行驶状态，如果 C 的当前速度没有达到最大限速 $v_{\max}$，则它以加速度 a_+加速以达到最大限速，车辆加速到目标速度后匀速行驶.

(2) 如果车辆 C'与车辆 C 的距离小于安全距离，也就是说如果车辆 C'紧急刹车，车辆 C 以当前的速度行驶时间 γ 后开始紧急刹车，则两辆车将会发生碰撞. 因此，在这种情况下，车辆 C 以某个速度 a_C 减速，使得在下一个时刻 C'不在它的安全距离内.

a_C 的选择取决于车辆 C 和 C'的速度 v_1 和 v_2 与位置 x_1 和 x_2. 假设车辆 C'保持速度不变，则一个时间步长 Δt 后，新的位置和速度分别为

$$x_1'=x_1+v_1\Delta t-\frac{1}{2}a_C(\Delta t)^2,\quad x_2'=x_2+v_2\Delta t,$$

$$v_1'=v_1-a_C\Delta t,\quad v_2'=v_2,$$

此时新的安全距离为

$$G'=L+\gamma v_1'+\frac{v_1'^2-v_2'^2}{2a_b}.$$

为了使车辆 C'不在车辆 C 的安全距离内，下一时刻两车的距离需要满足：

$$x_2'-x_1'=G'.$$

将前面的式子代入，得到：

$$x_2+v_2\Delta t-\left(x_1+v_1\Delta t-\frac{1}{2}a_C(\Delta t)^2\right)=L+\gamma(v_1-a_C\Delta t)+\frac{(v_1-a_C\Delta t)^2-v_2^2}{2a_b}.$$

通过上式可以求得 a_C，即 C 尾随 C'时的加速度为

$$a_C=\frac{1}{\Delta t}\left(\frac{\Delta ta_b}{2}+v_1+\gamma a_b-\frac{1}{2}[(\Delta ta_b)^2+4\Delta ta_b^2\gamma+(2\gamma a_b)^2-4v_1\Delta ta_b$$

$$+8(x_2-x_1)a_b+8v_2\Delta ta_b-8La_b+4v_2^2]^{\frac{1}{2}}\right).$$

（3）通过合流区域的动作：车辆 C 一旦离开收费亭将会加速前进，直到有车辆或者合流点进入其安全距离之内；如果另一辆车 C'进入 C 的安全距离，C 减速且跟随 C'，直到 C'通过合流点，此时合流点在 C 的安全距离内；当合流点进入 C 的安全距离时，把合流点视为一个障碍物，依据与障碍物的安全距离确定制动减速度. 当 C 能够通过合流点时，它通过并加速前进.

（4）如果有障碍物在车辆的安全距离内，计算车辆与障碍物的安全距离，不同于跟随情况，减速度用 $a_$代替 a_b，反应时间用 Δt 代替 γ，且不考虑车长，其他参数同上. 此时，可得到制动减速度为

$$a_o=\frac{1}{\Delta t}\left(\frac{3\Delta ta_}{2}+v_1-\frac{1}{2}[5(\Delta ta_)^2-4v_1\Delta ta_+8(x_2-x_1)a_]^{\frac{1}{2}}\right).$$

综上所述，车辆通过收费站的加速度确定过程如图 2-14 所示.

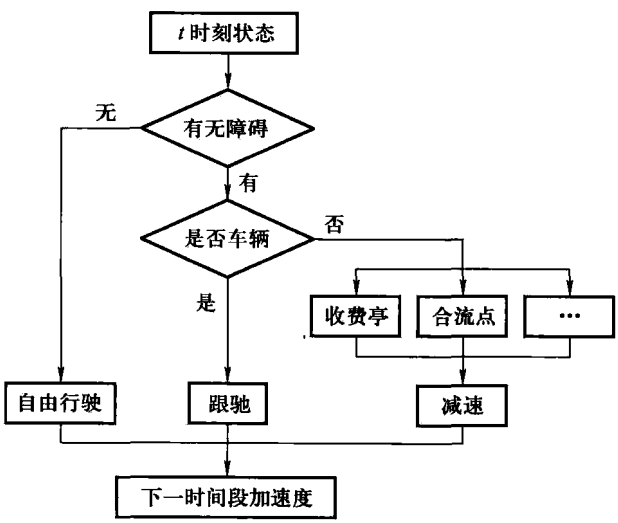


图 2-14 车辆通过收费站的加速过程

假设车辆在一个时间步长内加速度不变，则一旦确定了车辆 C 在某一时刻的加速度 a，就可以确定车辆在下一时刻的位置和速度. 设 x、v 和 x'、v'分别是车辆当前时刻的位置、速度和下一时刻的位置、速度，则

$$v'=v+a\Delta t,\quad x'=x+v\Delta t+\frac{1}{2}a(\Delta t)^2.$$

2.4.5　多车道车辆行为模型

1. 车流的生成方法

为了对收费站的交通流特性进行分析，需要对多辆车通过收费站的行为进行仿真，并在此基础上统计多辆车通过收费站的平均延迟时间. 为此，首先需要确定车辆到达的时间表，即单车到达时间的分布. 这里使用车流密度函数 $\rho(t)$生成到达时间表，基于 $\rho(t)$ 随机地生成某辆车到达坐标 0 点（即收费亭前）的时间. 假定在没有收费区时车辆正常到达 0 点的时间为 t，减去车辆通过收费亭前的队列所用的总时间，即为车辆正常到达队列的时刻 t'. 如果 t 时刻第 i 队的队列长为 $l_i(t)$，则可以确定收费队列 i 下一辆车的到达时间为

$$t'=t-l_i(t)/v_{\max}.$$

然而，由于车辆是从最大速度 $v_{\max}$开始减速，因此驶入收费站的位置 x_o 和时刻 t_o 分别为

$$x_o=-l_i(t)-v_{\max}^2/2a_-,\quad t_o=t'-v_{\max}/a_-.$$

2. 收费亭前队列的分配方法

当一辆车到达时，它被分配到长度最短的队列上去. 令 c_i 为在收费车道 i 上的车辆数，车辆在车道上具有相等的间隔距离 ε，收费站区域的半径为 d，队列的长度为 $l_i=c_i(L+\varepsilon)$，其中 L 为车长. 如果 $l_i>c_i(L+\varepsilon)$，那么队列延伸出分流区到了高速路上，而那里有 m 条车道. 假设队列的长度大致相同，增加的最小长度则是一辆车的车长，车的总数大约为 n，因此平均到 m 条车道上，则每条车道上的队列长度增加 n/m 辆车. 于是有

$$l_i=\begin{cases}c_i(L+\varepsilon), & c_i(L+\varepsilon)<d,\\ d+\dfrac{n[c_i(L+\varepsilon)-d]}{m}, & c_i(L+\varepsilon)\geqslant d.\end{cases}$$

队列的长度随着加入的车辆和离开的车辆而变化，即

$$\frac{\Delta c_i}{\Delta t}=\frac{m}{n}\rho\left(t-\frac{l_i(t)}{v_{\max}}\right)-s_i.$$

其中 s_i 为收费亭 i 的服务率. 用上式和 l_i 跟踪收费亭前队列长度的变化.

让车辆 C 加入到某个队列中去，如果队列长度 $l_i>0$，则 C 必须等到在它之前进入队列的所有车辆都完成交费后，C 才能完成交费进入合流区域.

3. 车辆的合流方法

当车辆交费后驶出收费站进入高速公路的车道为空时，合流点才不会对车辆形成障碍. 此时，车辆可以加速通过合流点进入高速公路. 否则，需要考虑多辆车合流时对合流速度和时间的影响. 当车辆驶入合流区域接近合流点时，它加入车辆合流队列（至多 n 辆车）. 当车辆通过合流点后，其加速行驶至最大限速. 设 t_m 为车辆 C 合流的时刻，v_{out}是此时的速度，则 C 驶往高速公路上且车速重新达到 $v_{\max}$的位置 x_f 和时刻 t_f 分别为

$$x_f = d + v_{\text{out}} \frac{v_{\max} - v_{\text{out}}}{a_+} + \frac{(v_{\max} - v_{\text{out}})^2}{2a_+}, \quad t_f = t_m + \frac{v_{\max} - v_{\text{out}}}{a_+}.$$

2.4.6 模型参数的确定

1. 车流密度

为了确定合理的密度函数，考虑采用实际的车流数据. 这里所用的大部分数据来源于 Orlando-orange 国家高速公路官方网站[24]，还有一些是关于车辆的报道. 首先对车流密度函数相关问题给出一些简单的假设：

（1）为了定义合理的平均延迟时间，选择有代表性的一天车流的高峰期和低谷期，限定车流密度函数 $\rho(t)$ 的时间区间为 $[0, 3\ 600\times24]$ s. 在大多数工作日里，车流基本具有相似的模式，因此，这种选择是合理的.

（2）车流密度函数 $\rho(t)$ 是分段常数函数，按小时变化. 由于车辆是离散的，实际上 $\rho(t)$ 是在较长时间内的平均值.

（3）车辆到达的时间间隔服从正态分布.

基于 Orlando-orange 国家高速公路官方网站上的相关报道，给出佛罗里达一些高速公路每小时的车流量，结合车辆到达时间间隔服从正态分布，可以得到车流密度函数 $\rho(t)$. 然后基于车流密度函数生成高速公路上的车辆到达时间表. 为了检验模型，我们考察了 6 条佛罗里达的高速公路，车道数从 2 车道到 4 车道变化，且具有不同的车流量. 从观测数据可以看到，6 条高速公路每小时的车流密度主要区别于数量大小，曲线的形状较为相似. 因此，基于两条这样的标准高速公路，包括 4 车道的 Holland West（高密度公路）和 3 车道的 Bee Line（低密度公路），建立车流密度模型，其分布如图 2-15 所示.

2. 收费亭服务率

假设每个收费亭的平均服务率为 s 辆/s，但是不同类型的收费亭的服务率有很大差异，如人工收费、电子收费等. 因此，综合考虑各类收费亭的服务效率确定一个综合的 s 值. 根据 Edie 的数据[25]，人工收费亭的平均服务时间是 12 s，依据 Orlando-orange 国家高速公路网站的数据，电子收费亭的平均服务时间是2 s. 类似的，根据城市 Houston 的报道[26]显示人工收费亭的平均服务时间是 10 s，机器收费时间为 7 s. 基于这些统计数据，这里设平均服务时间为 5 s，即

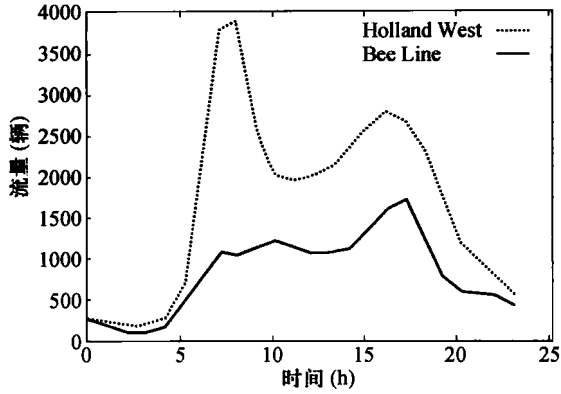


图 2-15　车流量密度分布图

车辆的平均等待时间是 5 s，亦即收费亭的平均服务率为 $s=0.2$ 辆/s.

3. 车辆基本参数

根据车辆动力学原理，车辆在加速和制动过程中，能够达到的最大加速度、减速度主要由道路的路面附着系数（静摩擦系数）μ_r 来决定. 在平直、干燥的柏油路面上（路面附着系数约为 0.8），车辆的最大制动减速度为 $a_b=\mu_r \cdot g=8\ \mathrm{m/s^2}$. 通常认为最大的“舒适”减速度为 $-3\ \mathrm{m/s^2}$. 美国各州公路及运输工作者协会提供的车辆在接近交叉路口时的减速度一般在 $-2 \sim -2.6\ \mathrm{m/s^2}$. 这里考虑到车辆接近收费亭时会有指示信号，驾驶员的舒适加/减速度的大小取值为 $\pm 2\ \mathrm{m/s^2}$.

综上所述，并结合前面的假设，对车辆的基本参数的取值进行设定，具体如表 2-6 所示.

表 2-6　车辆的基本参数

名称	符号	取值
舒适加速度	a_+	$2\ \mathrm{m/s^2}$
舒适减速度	a_-	$2\ \mathrm{m/s^2}$
最大制动减速度	a_b	$8\ \mathrm{m/s^2}$
车长	L	4 m
最大限速	v_{max}	30 m/s

4. 驾驶员的反应时间

勒纳（Lerner）等人从广泛的研究中总结出了制动反应时间（包括制动开始的时间），分析了两种类型的反应：

（1）驾驶员之前不知道什么时候应该开始制动，甚至不知道是否应该制动的情况下快速反应；

(2) 驾驶员被告知将要制动的信号，但不知道确切的时间.

对未知的情况下实现制动，有95%的人反应时间为2.45 s，这也接近美国各州公路工作者协会的建议：所有车速安全停车距离的反应时间是2.5 s. 这个2.5 s的估计值被广泛应用于高速公路上的所有情况，包括停车视距和其他各种视觉距离. 另有调查情况表明，驾驶员的平均反应时间是0.75 s，标准差为0.28 s. 这里假设所有车辆具有相同的预期反应时间和非预期反应时间，它们的取值分别为

$$\Delta t=1\ \text{s},\ \gamma=2\ \text{s}.$$

基于此，仿真步长取 Δt.

2.4.7 模型的仿真与结果分析

基于上面建立的车辆运动规律主要模型构建仿真程序，以实现收费系统中的交通流情况的仿真.

1. 仿真流程

跟驰模型属于典型的基于时序的仿真系统[27]，时间推进方式采用时间扫描法，即以等长的时段将仿真时钟向前推进，每一步长对系统更新一次. 在每一个时间步长循环更新每辆车的状态，并确定每辆车的运动行为. 在设定仿真时长之后，基于前面的模型可以确定车辆在系统中的任何时刻加速或减速，以及加速度的大小等. 基于跟驰模型的交通流仿真的基本流程如下：

(1) 初始化：收费亭前队列的初始化，模拟时钟 $t=0$.

(2) 更新时钟：$t=t+\Delta t$.

(3) 根据车流密度，确定 Δt 时间内到达的车辆，以及它们在 t 时刻的位置、速度、加速度，并更新队列长度.

(4) 更新系统中车辆、队列等的状态，并确定车辆在下一时间段 Δt 内的加速度.

如果车辆在收费亭前的队列中，根据该时段内是否有车辆离开收费亭更新车辆的位置；否则，根据加速度模型确定车辆的加速度.

从系统路段的最前端开始更新路段上车辆的状态，直到路段的末端结束. 车辆状态方程更新如下：

$$v'=v+a\Delta t,\ x'=x+v\Delta t+\frac{1}{2}a(\Delta t)^2.$$

(5) 判断是否满足仿真结束条件，若不满足，转到步骤 (2).

(6) 仿真结束.

基于以上仿真流程，编写仿真程序，便可进行收费站系统交通流的仿真.

2. 仿真结果分析

通过仿真，统计不同收费亭设置时的平均延迟时间. 通过比较该统计量的

大小，确定最佳的收费亭数目 n.

使用 Orlando-orange 国家高速公路网站提供的数据，可以建立 14 种测试场景：交通流包括高密度和低密度两类交通流量；高速公路车道包括从 1 车道到 7 车道. 针对每一种情形，将收费亭的数目 n 从 m 到 $2m+2$ 变化，通过仿真统计出每种情况下的平均延迟时间. 对于最优收费亭方案的确定，按照下面的步骤：

首先，找出最小平均延迟时间 $t_{\min}$(s)，以及对应的收费亭数目 n_0.

然后，查找平均延迟时间在 $t_{\min}\pm1$ 范围内的收费亭方案. 如果没有满足条件的方案，则 n_0 为最优的收费亭数目；否则，找出满足条件的方案中收费亭数目最小的方案，将方案所对应的收费亭数目与 n_0 比较，如果小于 n_0，则该方案为最优.

不同情形下得到的最优收费亭数目如表 2-7 所示. 可以看出，高密度交通流情形下，最优收费亭的数目趋向于高速公路车道数的 2 倍；而低密度交通流情形下，最优收费亭的数目等于高速公路的车道数. 另外，从平均延迟时间来看，高密度交通流情形下，如果仅有与车道数相同的收费亭数目，平均延迟时间大概是 500 s，几乎是最优收费亭设置延迟时间 25 s 的 20 倍. 因此，如果不论在一天的任何时段具有较高的车流密度，建议不要仅仅建设与车道数相同的收费亭数目. 然而，对于低密度交通流情形下，平均延迟时间为 22 s，与车道数相同的收费亭数目是最优的.

表 2-7 不同情形下的最佳收费亭数目

	高密度	低密度
高速公路车道数	1 2 3 4 5 6 7	1 2 3 4 5 6 7
最佳收费亭数目	3 4 5 6 8 9 11	1 2 3 4 5 6 7

3. 模型检验

为了检验模型的合理性，将仿真结果与实际的收费亭设置进行对比分析. 针对六个实际的收费站，通过仿真确定不同条件下的最佳的收费亭设置数目，其结果如表 2-8 所示.

表 2-8 不同情形下的最佳收费亭数目

收费站	最优收费亭数目	实际收费亭数目	比较结果
Hiawassee	4	4	相同
John Young Parkway	4	4	相同
Dean Road	5	4	差别显著
Bee Line	3	5	接近
Holland West	7	6	接近
Holland East	7	7	相同

从表中可以看出，有3个收费站的仿真预测数目与实际数目相同，其他3个稍有差异. 进一步做仿真计算，还可以得到Dean Road收费站仅设置4个收费亭，其平均延迟时间为70 s，与最佳设置时的平均延迟时间25 s相比，差别较为显著. 而对于Bee Line和Holland West收费站，平均时间差异至多为1 s. 这些结果表明所建立的模型与实际情况基本一致.

2.5 基于元胞自动机的交通流仿真模型

元胞自动机（Cellular Automaton，CA）技术研究在空间中离散个体间的相互作用而形成整体上的复杂行为. 在元胞自动机中，用一定的规则网格将空间分割为许多单元，这些规则网格中的每一个单元都称为元胞（cell），并且元胞的状态只能在有限的离散状态集中取值，所有的元胞遵循同样的演化规则，依据确定的局部规则做同步更新，通过元胞个体的状态变化来实现整体上的运动. 不同于一般的系统动力学模型，元胞自动机模型不是由严格定义的物理方程或函数确定，而是由一系列的演化规则构成.

交通流元胞自动机模型是20世纪90年代兴起的一种新型交通流模型. 元胞自动机模型的时间、空间、状态变量均为离散的，并且各个元胞在空间上的相互作用和在时间上的因果关系都是局部的，具有良好的并行性. 该模型具有计算效率高并且调节灵活的优点，便于考虑真实交通条件的各种效应. 因此，该模型被广泛应用于各种交通流现象的研究，如车流从自由运动到局部阻塞的变化、路障与“瓶颈”效应等. 这些复杂系统的非线性行为都可用元胞自动机模型来进行有效地仿真. 目前元胞自动机模型已成为研究交通流问题的重要模型之一.

2.5.1 问题的分析与假设

为了建立收费站交通流的元胞自动机模型，需要确定收费站的元胞空间描述以及车辆在收费站行驶的规则. 在交通系统的CA模型中，时间及车辆的位置、速度和加速度都被认为是离散的变量，速度定义为在一个时间步长内车辆向前移动的格点数. 用一维网格来表示一条道路，每一格点代表一元胞，这一元胞的状态或是空的，或是被一辆车占据.

下面基于传统的NaSch模型，考虑收费站交通流的特点，确定相应的规则. 在NaSch模型中，减速规则是基于车辆前面的空间，而收费亭强迫整个交通车流慢下来，考虑使用安全距离来实现车速的更新. 另外，在NaSch模型中，通过随机减速以使车辆达到稳态. 而在收费站问题中，车流速度在逐渐减慢，因此，不再采用随机减速，而是引入车辆随机到达和优先变道.

为了便于模拟，对系统做以下假设：

（1）车辆都是理想的，车身长度相同. 道路也是理想的平坦道路，在行驶过程中车辆不存在任何的侧倾、横摆等运动. 车辆的行驶方向始终向前保持不变. 在仿真路段内没有驶出或进入的出入口.

（2）车辆在收费通道内排列紧密. 驾驶员不希望其他车道的车辆插入队列，跟车距离将会比较近.

（3）在车辆进入分流区域时，是否变道取决于变道能否使得它们距期望的收费亭更接近一些. 车辆在合流区域会变换车道以便驶出死胡同车道. 忽略车辆在变道过程中的加减速行为，若车辆满足变道条件，则直接置换到相邻车道的前一个格点. 另外，不允许车辆在收费亭元胞前后的五个元胞内变换车道.

（4）在收费站内，收费亭前后的车道长度均为300英尺（1英尺=0.3048米）. 设置路障可以提高车道扩展和合并的效率，假设可以通过在收费亭前的区域内设置障碍，禁止变换车道. 车辆分流区域和车辆合流区域内设置的变道路障和默认车道是对称的.

（5）收费亭包括人工、自动和电子三种付费方式，有信号告知驾驶员每个收费亭接受的付费方式. 如果驾驶员倾向于某种方式，将会到相应类型的收费亭去.

（6）收费亭的运行费用忽略不计. 与收费站的建设和高速公路的维护相比，收费亭的运行费用是微不足道的，尤其是自动和电子收费亭比人工收费需要的维护费更少. 因此，只从顾客的角度来评价收费亭的设置是否最佳.

（7）车辆通过收费站的延迟时间是指车辆通过整个仿真区域所用的时间.

为了分析收费站的布局对车辆通过收费站延迟时间的影响，需要考虑强行变道、急刹车、没有电子通行证的车辆卡在电子通道等多种情形. 为了验证所建仿真模型的有效性，需要对仿真结果进行分析，比较仿真的输出是否与实际的交通行为相一致. 另外，进一步讨论收费站如何使用变道路障引导车道的扩展，或者配置适当数量的电子收费亭，以提高收费站的效率.

2.5.2　名词解释与符号说明

为了能够描述多种不同类型的收费亭设置，下面建立收费站车道设置的一般模型. 设收费站前后的高速公路都有 m 条车道，收费站内有 n 个收费亭. 高速公路车道分隔线的标号从1到 $m+1$，收费亭车道分隔线的标号从1到 $n+1$. 变道路障用数组 (x, y) 表示，x 表示公路车道分隔线的标号，y 表示收费亭车道分隔线的标号. 图2-16给出了 $m=4$，$n=6$ 时的情况，变道路障在(1, 1)，(2, 3)，(3, 4) 和 (5, 7)，点线表示变换车道可以通过. 假设变道路障的有序集合为 $\{(x_1, y_1), \cdots, (x_k, y_k)\}$，其中对于 $i=1, \cdots, k-1$ 有 $x_{i+1}>x_i$，则变道路障的集合必须满足以下规则：

（1）车辆不能驶出道路：（1，1）和（m+1，n+1）必须是变道路障.

（2）变道路障相互不能交叉，即对于 i=1，…，k−1，有 $y_{i+1}>y_i$.

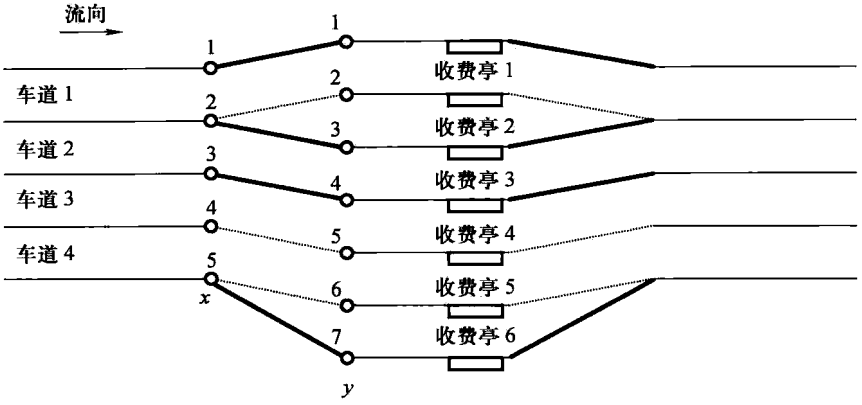


图 2-16 收费广场扩展车道示意图

默认车道：在收费亭前的合流区域，为每个高速公路车道安排一个默认的收费亭车道，使得车辆不用变道就能进入收费区. 图 2-16 中高速公路车道从 1 到 4 分别对应收费亭车道 1、3、4 和 5，并且相应的默认收费亭车道在收费亭之后又回到相应的高速公路车道，合流区域的其他的车道则是死胡同. 如果附近的车道没有变道路障，车辆可以进入.

扩展车道：通过使用变道路障引导车道的扩展，如车道 1 中的车辆可以驶入收费亭车道 1 或车道 2.

随机发车：依据概率分布生成到达的车辆流. 仿真从车辆在收费亭前的 1 mile处开始，直到车辆到达仿真路段的末端，即收费亭后 1 mile 结束.

车辆间距：两相邻车辆间的距离.

2.5.3 交通流元胞自动机模型

Cremer 和 Luding 最先基于 Wolfram 提出的 184 号基础元胞自动机规则，研究了如何利用确定性元胞自动机建立交通流模型. 20 世纪 90 年代，Nagel 和 Schreckenberg 提出了著名的 NaSch 模型[28]. 之后，元胞自动机交通流模型受到了更多研究者的广泛关注，提出了许多改进模型，以模拟更真实复杂的交通状况.

NaSch 模型是一种较简单的元胞自动机模型，该模型用一维网格代表一条单车道，即将所研究的单车道分成 n 个长度为 L 的小路段（元胞），每个网格代表一个元胞. 每个元胞的状态或者空，或者被车辆占据. 定义元胞长度 L 为道路阻塞时的平均车头间距，车辆速度的取值范围为 $\{0, 1, 2, \cdots, v_{\max}\}$，其中 $v_{\max}$ 为允许的最大车速. 在 NaSch 模型中，车辆按照以下的规则进行更新：

（1）加速规则. 假设驾驶员总是期望以最大的速度行驶，如果当前车速小于车辆的最大速度 $v_{\max}$，速度加1；如果车速等于最大车速，车速不变，即

$$v_n \rightarrow \min(v_n+1,\ v_{\max}).$$

（2）减速规则. 由于在CA模型中速度定义为一个时间步长内车辆向前移动的格点数，因此，如果与前车的间隔点数（即与前车的距离）小于当前车速，为了避免与前车碰撞，则减速以满足与前车的安全距离，即

$$v_n \rightarrow \min(v_n,\ d_n).$$

这样即使前车在当前时刻停下，在下一个时刻前后车也不会发生碰撞.

（3）随机慢化. 由于不确定因素，如过度刹车、道路条件变化、心理因素等，驾驶员驾驶行为具有不确定性，以随机慢化概率 p 减速，即

$$v_n \rightarrow \max(v_n-1,\ 0).$$

（4）位置更新. 车辆依据由前三步确定的车速在道路上向前移动，即

$$x_n \rightarrow x_n+v_n.$$

其中，v_n 和 x_n 分别表示第 n 辆车的位置和速度；$d_n=x_{n+1}-x_n-1$ 表示其与前车之间空置的元胞数.

NaSch模型是可以重现道路交通流基本特征的一个最小化模型，四个规则缺一不可. 虽然NaSch模型形式简单，但却可以在一定程度上描述出实际的交通流状况. 然而，如果要捕捉更为复杂的交通流现象或考察更为复杂的交通条件，就需要添加一些新的规则. 针对NaSch模型的单车道、不允许超车等局限性，很多学者相继对其进行改进和扩展，建立了多车道交通流模型，包括定义变道规则等. Rickert等和Chowdhury等分别通过引入一套变道规则，将单车道的NaSch模型扩展到双车道系统中去.

Chowdhury等人[29]构造了一个对称的双车道元胞自动机模型（Symmetric Two-lane CA），在其规则中，变道动机为车辆在原车道上不能按期望的速度行驶且目标车道行驶条件比原车道好，变道的安全条件为确保不会与后面的车发生碰撞. 变道规则可以用下面的公式表示：

$$C_n=\begin{cases}1, & d_n<\min\{v_n+1,\ v_{\max}\},\ d_{n,other}>d_n,\ d_{n,back}>d_{safe},\\ 0, & \text{其他}.\end{cases}$$

其中，C_n 表示第 n 辆车是否具备换道条件，d_n 为第 n 辆车与前方车辆的间距，$d_{n,other}$ 为第 n 辆车与相邻车道前方车辆的间距，$d_{n,back}$ 为第 n 辆车与相邻车道后方车辆的间距，d_{safe} 为模型中限定的安全变道间距. 换道条件中，$d_n<\min\{v_n+1,\ v_{\max}\}$ 表示车辆在原车道受到阻挡，$d_{n,other}>d_n$ 表示在另一车道上可以达到更快的速度，$d_{n,back}>d_{safe}$ 表示安全换道间距符合条件，即在另一条车道上，后方的车辆与其有一定的距离.

2.5.4 收费站的元胞自动机模型

1. 最优收费亭设置的标准

车辆通过收费区时，需要经过减速、停止、付费等过程，比通过相同距离的高速公路路段花费的时间长. 通常顾客对于等待的不满意度是等待时间的非减凸函数，因此等待时间最好有较均匀的分布，而不是具有较高的标准差. Klodzinski 和 Al-Deek 建议延迟的 85% 是一个较好的标准[30]. 这里，综合考虑能够较快通过的车辆，避免结果被异常值影响，采用落在 50% 和 85% 之间数据的均值作为各类车辆的等待时间标准. 为了分析不同类型车辆的等待情况，把车辆分为轿车、卡车和有电子通行证的车辆，计算不同车辆等待时间的平均值，然后再根据不同车辆的数量比例进行加权平均.

2. 基本参数定义

(1) 元胞长度. 定义一个一维点阵来代表一条单车道，即将所研究的单车道分成 n 个长度为 L 的小路段，每个点阵或空闲或有一辆车. 这里设每英里有 250 个元胞，即每个元胞表征的实际长度是 21 英尺，元胞表征的实际长度略大于平均车长（约 15 英尺）.

(2) 最大速度 $v_{\max}$. 设车辆的最大速度为每个时间步长前进 5 个元胞，如果取时间步长为 1 s，则对应的最大速度是 72 mph，大致等于高速公路上的期望车速.

(3) 更新时间间隔. 更新时间间隔取 1 s.

(4) 边界条件. CA 模型采用的边界条件是封闭的道路系统，即在收费站附近没有匝道使得车辆进出，这样可简单地将所研究路段上的车辆数控制在一定的数量上，以研究此车流密度下的交通状况.

3. 跟车模型

假设车辆尽可能快地通过收费站，车辆的速度受限于与它前面车辆的距离，两车之间的距离足以使得紧急刹车时能够避免碰撞. 对于任意前后跟随的两辆车，即使前车以每个时间步长 1 个单位（即 21 英尺/s^2）的减速度进行减速，后车以相同的减速度进行减速，两车能够避免碰撞. 假定在某个时间段，前车和后车的速度分别是 v_c 和 v_{fc}，如果以 1 个单位的减速度减速直到它们都停下来时，它们行驶的距离为

$$v_c+(v_c-1)+\cdots+1=\frac{1}{2}v_c(v_c+1),$$

$$v_{fc}+(v_{fc}-1)+\cdots+1=\frac{1}{2}v_{fc}(v_{fc}+1).$$

为了避免撞车，停下来时应该保证后车仍在后面，因此要求：

$$gap+\frac{1}{2}v_c(v_c+1)-\frac{1}{2}v_{fc}(v_{fc}+1)>0,$$

即

$$gap>\frac{1}{2}(v_{fc}-v_c)(v_c+v_{fc}+1).$$

每次速度更新时，应该检查后车加速 1 个单位后是否还满足这个不等式，否则必须减速以保持该不等式成立，即保证不发生碰撞.

然而，根据这个模型，如果两辆车以相同速度行驶，理论上它们会发生碰撞，实际情况是后车会留一定的空间，因此，我们考虑增加$\left\lfloor\frac{v_c}{2}\right\rfloor$的距离. 对于高速行驶的情况，这可能不是一种安全距离，但在收费站附近车辆出现拥挤时，车辆的速度不会太快. 因此，最终的标准是车辆观察它前面的空间并调整速度，使得满足下面的不等式：

$$gap>\left\lfloor\frac{v_c}{2}\right\rfloor+\frac{1}{2}(v_c-v_{fc})(v_c+v_{fc}+1).$$

4. 车辆变道模型

在车辆行驶过程中，诱发驾驶员产生变道意图的原因有多种，但总体上可以分为三类：强制性变道、期望性变道和随意性变道. 强制性变道是指车辆由于转弯、前方发生交通事故或车道使用的限制必须变道；期望性变道是指车辆在将要达到强制变道区域之前，提前进行变道；随意性变道则是因为车辆的行驶速度达不到或超出了驾驶员的心理期望值而进行变道. 随着变道的意图不同，驾驶员往往会采用不同的变道行为.

车辆通过收费站的变道主要包括：进入收费站后收费亭车道的选择、离开收费亭后在合流区域车道的选择等. 车辆进入收费站时根据自身的付费类型选择相应的期望车道，如果直接进入的车道符合付费类型，则采用随意性变道，以更快地到达收费亭. 在车辆驶向合流点时，如果未能在合流点进入它的安全距离之前变换车道完成期望性变道，则进行强制性变道. 在强制性变道过程中，车辆必须在道路某一关键点之前完成其变道行为. 当产生强制性变道意图后，驾驶员首先减速并选择目标车道，再判断向目标车道变道的可行性. 若变道可行则实施变道，否则停下来等待，直到有机会完成变道.

在车道变换时，驾驶员必须首先查看目标车道上前、后车的位置和速度，判断它们之间是否存在足够的间隙来实现变道. 假如车辆并入后，形成的两个新的间隙能满足安全要求，则这个间距是可接受的，就可以实施变道，这种变道模型称为间隙接受模型（gap acceptance model）. 基于此模型，只有在路段上车辆很少、存在很多较大车间距的时候变道行为才可能发生，而密度较大或是拥挤的情况下，变道行为将不会发生. 因此，车辆进入收费亭前的队列之后，将不再有变道行为.

综合考虑安全变道、期望收费亭类型、强制变道等多种因素，收费站内车

辆的变道决策过程概括如下：

（1）车辆驶入收费站时，选择符合车辆付费类型的最短队列作为车辆的目标车道.

（2）在驶向收费亭前的队列的过程中，根据目标车道的位置，判断与目标车辆同一侧的近邻车道是否满足变道的安全条件. 这里变换车道是为了能够达到目标车道，所以变换的条件为是否能够安全进入近邻车道，即

$$d_{n,other}>d_{safe},\ d_{n,back}>d_{safe}.$$

如果满足安全条件，则变换车道，否则留在原车道.

（3）在车辆离开收费亭后，车辆尽量离开不能直接通向高速公路的车道. 设定车辆的目标车道为其近邻的能直接通往高速公路的车道.

（4）在车辆驶向合流点的过程中，根据目标车道的位置，判断与目标车道同侧的近邻车道是否满足变道的安全条件，如果满足则变换车道，否则留在原车道. 当到达目标车道后，采用随意性变道. 如果到达目标车道之前合流点进入安全距离内，则采用强制性变道.

（5）随意性变道. 如果驾驶员左、右侧车道条件相同，则在两车道之间随机选择.

（6）强制性变道. 如果近邻可用车道满足变道条件，则变道，否则等待变道条件成立时，完成变道.

5. 车辆速度的更新

对于每条车道，从前向后更新车辆的位置和速度，车道的顺序随机选择. 然而，这样的结果是信息往后传递，比如，都以速度 1 行驶的整个车辆链会全部停下来. 对于这个无限波速问题可以通过引入随机减速来解决.

考虑到收费站内车辆低速行驶，车辆连续缓行更合适，因此，这里不考虑随机减速，且使用序列更新. 序列更新可以避免两个不同车道的车辆同时进入它们中间的车道问题. 对于每个时间步长，随机生成车道更新的优先顺序，减少车道变换的不对称性. 一条车道上车辆速度更新的流程如图 2-17 所示.

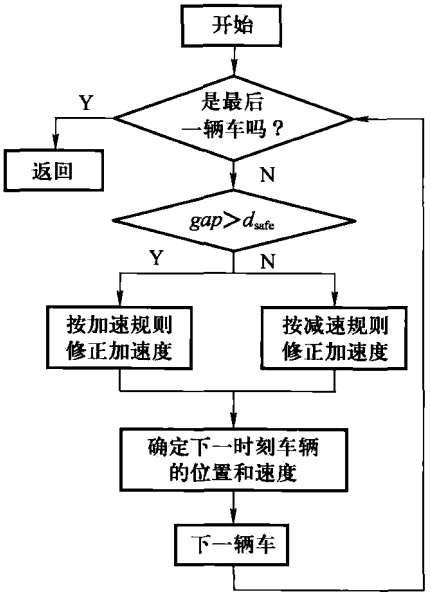


图 2-17 速度更新程序框图

2.5.5　模型的仿真与结果分析

1. 仿真参数的确定

（1）车流到达. 仿真一个收费站 70 min 的车流情况，假设车流密度从低到高逐渐增长，40 min 后开始从高到低变化. 图 2-18 给出了低密度、中密度及高密度交通车流的生成比率，基准的总车流量为 70 分钟 2 500 辆车. 对于四车道的高速公路上的收费站，70 min 内三种交通状况的车流量分别大约为 2 200辆、3 300 辆和 3 600 辆.

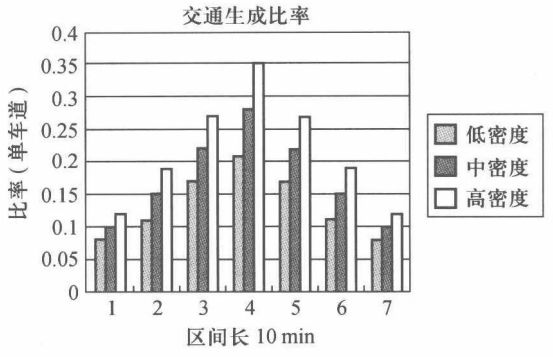


图 2-18　交通流生成速率

（2）收费的延迟时间. 使用电子通行证的车辆通过电子收费亭的速度为每个时间增量通过两个元胞（≈30 mph），它也可以通过人工或自动收费亭，但需要等待 3 ~ 7 s. 其他车辆通过自动收费亭的时间是 8 ~ 12 s，通过人工收费亭的时间是 13 ~ 17 s. 使用随机的均匀分布来实现服务时间延迟，以保证车辆不会同时离开收费亭.

（3）收费亭设置. 考察两种类型的收费亭：第一类是不同类型的收费亭集中在一起；第二类是不同类型的收费亭存在交叉. 对于 4 车道的高速公路，使用 2 个电子收费亭、4 个自动收费亭和 2 个人工收费亭. 首先将相同类型的收费亭按顺序排在一起，然后再改变类型. 另外，考察设置变道路障和不设置变道路障的两种情况，假设每条高速公路车道的默认收费车道为编号为奇数的车道，且每条高速公路车道有一条默认的收费车道.

（4）半数车辆具有电子通行证，10% 的车辆是卡车且没有电子通行证.

2. 仿真结果与分析

图 2-19 给出了 10 次仿真运行的平均结果，x 轴表示不同的收费亭方案，“e”代表电子收费亭，“m”代表人工收费亭，“a”代表自动收费亭，y 轴表示通行时间落入 50% ~ 85% 之间的所有数据的平均值. 如果没有收费亭，车辆以全速行驶这段距离需要的时间是 100 s.

从图 2-19 可以看出，有变道路障的情形比不设置变道路障的情形略差.

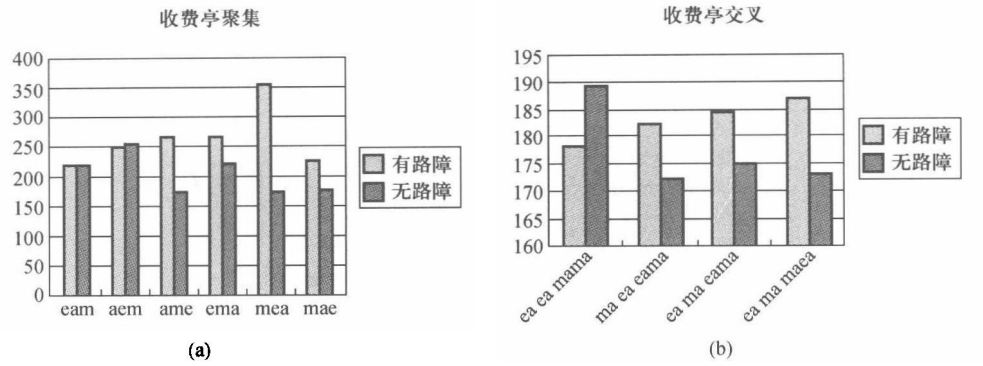


图 2-19 收费广场不同设置情况下的仿真结果

这可能是由于如果对车道分流区域不做限制，可能会使车道变换更容易些. 从图（a）和图（b）的对比可以看出，收费亭的交替配置似乎能减少些延迟时间，这在高速公路不同车道上不同类型车辆分布相同的情况下是成立的. 但在不同类型收费亭交叉设置时，也会出现更多的危险车道变更的警告和车辆卡在收费亭通道内等情况.

在正常的交通流密度情况下，研究不同类型收费亭比例的确定. 以 4 车道高速公路上 8 个收费亭的收费站为例，且在车道扩展区域未设置变道障碍，把电子收费亭放在左侧，人工收费亭放在右侧. 对于轿车、卡车和电子通行车辆的不同分布，改变每类收费亭的数量，并且在正常的交通密度状况下进行了多次仿真. 仿真结果表明：除非卡车的比例非常低，否则仅设置 1 个人工收费亭给卡车，将会出现大量的卡车卡在它们不能使用的通道内；电子收费亭的数量应为 2 或 3，依赖于有电子通行证的车辆是否多于没有电子通行证的车辆. 由于 4 通道公路收费亭的合理配置是 8～12 个，因此，这里建议设 1/4 的电子收费亭，1/2 的自动收费亭，1/4 的人工收费亭.

在收费亭类型设置不变的情况下，通过改变收费亭的总体数量以确定最优的收费亭数目. 依据上面的比例，分别使用低密度、中密度、高密度的交通车流数据，可以得到不同收费亭数目情况下的延迟时间，如图 2-20 所示.

最后，考虑没有卡车且没有电子通道的情形，即所有收费亭都是自动的，分别使用低密度、中密度、高密度的交通车流数据，得到不同收费亭数目情况下的延迟时间，如图 2-21 所示. 从仿真结果可以看出，由于没有电子收费亭，车辆将经受更长的时间延迟，因为每辆车都必须在收费亭前停下. 若每车道仅有一个收费亭，正常交通流量（70 分钟 3 000 辆车）情况下，许多车辆的等待时间超过 45 分钟. 在这种情况下，要想达到最小延迟时间大约需要 12 个收费亭，而不是使用自动和电子通道时的 9 个收费亭.

基于以上仿真分析，可以得出以下结论：

（1）收费站应该基于所提倡的行为模式排列收费亭，把较快的收费亭放

在左侧，较慢的收费亭放在右侧. 这主要是由于高速公路上的快车道通常在公路的左侧.

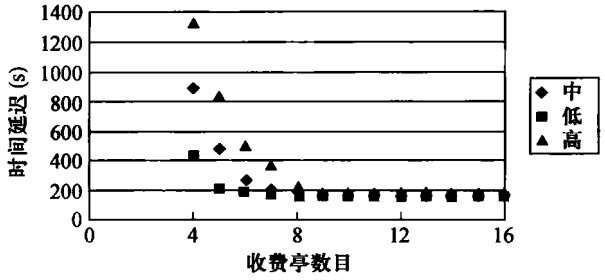


图 2-20　不同收费亭数目对应的延迟时间

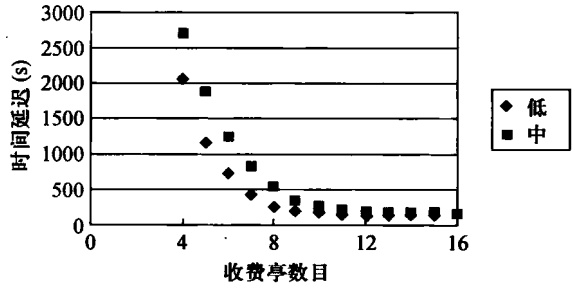


图 2-21　自动收费亭数目对应的延迟时间

（2）一般不设置变道障碍. 变道障碍将妨碍驾驶员进入他们所需要使用的车道.

（3）收费亭的数目由不同类型车辆的分布来决定. 因为车辆类型的分布对时间延迟具有更大的影响，而交通车流密度对收费亭数量的影响不大.

（4）收费亭的一个有效的比例是 1 个电子收费亭、2 个自动收费亭、1 个手动收费亭.

2.5.6　模型的检验与分析

1. 模型的合理性检验

为了检验所建模型的合理性，对车辆通过收费亭的行为模式进行分析，判断仿真结果是否符合实际情况. 为此，在正常交通流量情况下，对 4 车道高速公路上 8 个收费亭（从左至右：2 个电子收费亭、4 个自动收费亭、2 个人工收费亭）且没有任何路障的收费站进行仿真，记录车辆的到达时间和通过收费站的行驶时间. 图 2-22 给出车辆行驶时间与到达时间的关系.

从图 2-22 可以看出，车流通过收费站有两个主要特点：

（1）底部的点代表有电子通行证的车辆. 即使在交通繁忙时，它们也不会耽搁太长时间，因为它们可以不用停下就可以通过收费亭.

（2）顶部的点代表没有电子通行证的车辆. 它们等待时间的分布与车流

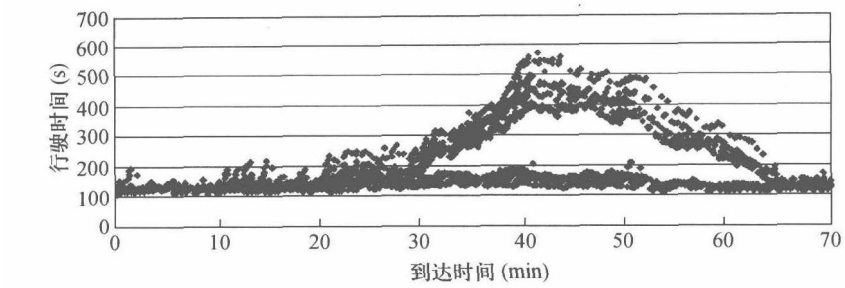


图 2-22 车辆行驶时间与到达时间的关系

分布后移约 10 分钟. 甚至可以看到图形被分割成数个“通道”，这显示了较慢的货车通道（人工收费亭）和普通汽车（自动收费亭）的差异.

2. 参数的灵敏度分析

变化车道扩展和收缩区的长度，既不会对延迟时间产生影响，也不会对不良行为记录产生显著的影响. 在适当范围内改变具有电子通行证的汽车和卡车的比例，也不会产生显著影响. 对于公路的通道数 n，在收费站使用变道障碍时，$2n$ 或 $2n+1$ 个收费亭后增加一个新的收费亭的边际收益较小，除非交通车流量增大.

2.6 问题的综合分析与进一步研究

2.6.1 问题的综合分析

在交通流理论中，大多数交通流模型可分为两类：微观模型和宏观模型. 微观模型指的是交通流离散模型，研究单个车辆和驾驶员的行为，使用车辆的间距和速度来刻画整个交通流，这类模型包括车辆跟驰模型、元胞自动机模型等. 宏观交通模型将车流看做是流体的流动，利用流体力学的理论进行研究. 宏观模型评价“平均”的行为，通常用的变量包括稳态速度、车流流量和交通流密度等. 微观模型评价“统计”的行为，用统计量的均值、方差等. 还有一些模型建立起两类模型之间的桥梁，如气体动力模型，在该模型中车辆被看做是相互作用的粒子，通过积分关于相空间密度分布函数的 Boltzmann 方程，然后引入近似关系来封闭所得到的方程组，就可以导出宏观交通流模型[31].

下面对收费站交通流的几类模型进行综合分析，比较不同模型在解决此问题时的优缺点.

1. 宏观交通流模型

宏观仿真模型中的交通流被视为可压缩的媒体或流体，交通流的运动按照流体机制来处理. 宏观仿真通过流量与密度的关系来控制交通流的运行，在模型中不追踪单个车辆的移动. 在高密度情形下，离散模型和连续模型非常吻

合. 通过对基本交通流变量的了解，可以了解系统运行的状态，它是衡量系统运营效果的有效办法.

与微观模型相比，宏观模型只需要求解几个由参量构成的偏微分方程或方程组，其计算量相对较小. 另一方面，还可以采用适当的数学工具进行解析分析. 连续交通流模型的缺点是：交通流不一定是连续的，所以在模型中只能近似，不能完全反映现实情况. 实际交通流的状况是在不断变化的，而宏观模型只能给出稳态的结果，虽然它对系统的宏观规划和控制有一定的参考价值，但难以体现真实交通行为的复杂性. 另外，在连续交通流理论中，速度与密度关系只适用于平衡状态，满意的速度与密度关系很难找到，通常是通过假设或推断得出.

由于简单连续交通流未考虑加速度和惯性的影响，因而未能如实反映出非平衡状态交通流的动力特性. 于是学者们引进了高阶连续介质模型，认为由密度和速度所确定的速度不是当时就能达到的，而是在一段时间之后，并考虑了各车相邻车距对考察车辆车速的影响. 因此，某一地方的车流速度不再由当地密度值决定，而是依赖于它前方某处的车流密度值.

2. 排队模型

排队理论是交通流理论中的重要组成部分，它以概率论为基础，在研究车辆延误、通行能力、信号灯配时以及停车场、加油站等交通设施的设计与管理方面得到广泛的应用. 车辆排队系统模型可以方便地计算系统稳态情况下的各项运行指标，它的优点是可以支持参数的很多变化，给定合理的参数取值，能够直接计算相应的指标值，算法易于使用和检验. 排队模型的不足之处是：为了避免无限队列，假定系统的顾客（车辆）到达率低于服务率，这个假设在长期运行中是成立的，但实际情况是在某些时间段，到达率可能会超过服务率. 因此，尽管最优解对应的平均等待时间较小，但对个别车辆的最大等待时间可能会相当大.

车辆排队系统用于收费站交通流建模时的一个难点是车流合流部分服务率的估计. 一些方法基于一定的假设，比如车辆均匀加速通过合流区等，近似计算车辆通过合流区的时间. 目前，所见的研究成果多是对收费亭前的排队现象进行研究，而之后的排队问题研究成果很少，对车辆通过合流区的具体排队过程还缺乏必要的认识. 因此，车辆合流部分的排队问题，还需要进一步采集数据、统计分析，以检验假设的合理性和所建模型的有效性.

另外，在排队模型中对车辆到达时间间隔和收费亭的服务时间进行了服从一定分布的假设，这种假设是否合理、假设在什么情况下成立等是需要考虑的问题. 因此，排队模型的检验也是模型应用的必要组成部分.

3. 微观仿真模型

微观交通仿真模型的研究对象是车辆个体运动，运算量大，要求运算速度

快. 近 10 年来，随着计算机技术的飞速发展，微观交通仿真模型的复杂性有了巨大的增长. 同时，在交通工程与交通规划领域，微观交通仿真模型的使用也有了巨大的增长. 目前，全球已有一百多个微观交通仿真模型，并有几十种商业化的微观交通仿真系统.

微观仿真模型利用车辆跟驰、车道变换以及可接受间隙等理论，刻画交通网络中运行的单车行为，能够反映出车辆间的相互作用，描述车流运行的随机行为、车流中不同车型构成时的交通特性以及不同驾驶员的道路选择偏好，还可以预测单个车辆的出行时间延误、停车次数、排放等性能指标及其分布. 微观交通仿真模型对现有的运输系统的运行状况与对其潜在的改进所导致的运行状况可以提供非常有价值的信息[32].

跟驰理论在描述微观的跟车行为方面具有以下特点：

(1) 跟驰模型的描述与实际的跟车行为有着很好的对应.

(2) 考虑了车辆个体的不同特性.

(3) 跟驰模型的使用往往只需获得各种车型的加、减速特性，模型参数的标定相对容易得多.

跟驰模型的不足之处是：跟驰理论的假设条件是车辆处于密度较高的非自由流状态，因而仅适用于单车道没有超车行为发生的交通流情形，而在实际交通流中大量存在超车、变道等现象，因此，跟驰理论具有一定的应用局限性.

CA 模型从简单的假设实现复杂的行为，其主要优点有：

(1) 模型简单，特别易于在计算机上模拟实现.

(2) 能够再现各种复杂的交通现象，反映交通流特性. 在仿真过程中人们通过考察元胞状态的变化，不仅可以得到每一辆车在任意时刻的速度、位移以及车头时距等参数描述交通流的微观特性，还可以得到平均速度、密度、流量等参数，呈现交通流的宏观特性.

(3) 能够再现单车道、多车道以及路网的交通流建模，机动车和非机动车交通流的建模等.

(4) 可处理收费站各种可能的配置，容易改变收费亭类型或设置变道障碍.

使用 CA 模型的难点是：需要获得真实世界的参数. 虽然现有高速公路的 CA 模型较好地描述了交通流的微观特性，但其规则比较简略. 全面考虑制约车辆运动的各种因素，开发出既符合动力学原理，又便于计算机仿真的 CA 模型，是需要今后进一步研究的内容之一.

2.6.2 进一步研究的问题

为了通过建立数学模型研究高速公路收费站的合理设置问题，在建模的过程中对实际问题进行了简化. 然而，现实中收费站的规划问题需要考虑一些实

际因素的影响，下面进行简要描述.

1. 交通流的特性分析

(1) 交通流量的时间统计特性. 交通流量是随时间变化的随机变量，其时间统计特性按照时间的长短，可分为年变、月变、周变和时变等. 道路功能不同，变化规律也会不同. 如何在设置收费站时对交通流量的时变特性予以考虑，是数学模型是否适用的条件之一. 高峰小时交通流量占该天全天交通流量之比称为高峰小时流量比（以百分比表示），通常用它反映高峰小时交通流量的集中程度. 一天中的交通高峰是交通分析的重点.

(2) 交通流量的车道分布. 对于多车道的公路，交通流量在各个车道上的分布存在一定的差异. 交通流量的车道分布取决于交通规则、交通组成、车速和交通总流量、驾驶员的习惯等因素. 高速公路上的交通流量主要取决于快速道的设置、超车习惯等，在收费站附近还受不同类型收费亭设置位置的影响.

(3) 驾驶员的行为特性. 驾驶员的行为特性是交通流模型的重要影响因素之一，尤其是对于微观交通流仿真. 道路上来往车辆、交通标志、路面状况等外界信息，通过驾驶员的视觉、听觉和触觉等感觉器官传入驾驶员的大脑，然后驾驶员依据其驾驶经验予以加工处理，做出相应的判断和决策，最终发出调整方向和速度的指令，从而达到改变汽车运动状态的目的. 交通流中的车辆相互之间通过驾驶员对信息的处理和控制形成一定的关系. 其他车辆的速度和加速度变化会被相邻车辆驾驶人感觉到，从而采取相应的措施，其中前导车辆和侧向车辆的影响比较明显. 驾驶员的行为特性会影响到交通流中的追赶、超车、合流、停车等行为.

另外，高速公路上运行的车辆类型多种多样，常见的有大货车、中小货车、小客车、大中客车等类型，车辆类型会影响到驾驶员的驾驶特性，车辆类型的分布将会对收费亭类型的分布产生影响.

2. 收费站的车道分布

在探讨高速公路收费站的车道设置时，有必要对高速公路适合采用的 ETC 车道形式和相应的 ETC 车道布局进行研究. ETC 车道的分类如图 2-23 所示. 从车道的专用性可分为两类：一类为专用车道，一类为混合车道. 专用车道只服务 ETC 车辆，而混合车道既服务 ETC 车辆，也服务其他付费类型的车辆. ETC 专用车道包括高速 ETC 专用车道和低速 ETC 专用车道. 高速 ETC 车道包括单车道自由流和多车道自由流，其中多车道自由流采用门架式 ETC 系统，允许车辆高速通过，且可变车道行驶，服务水平高，车辆通过时几乎不产生延误.

目前组合式收费系统中车道设置有两种不同的形式：

(1) ETC 车道内置式. 在半自动或人工收费系统中，收费站等待收费服

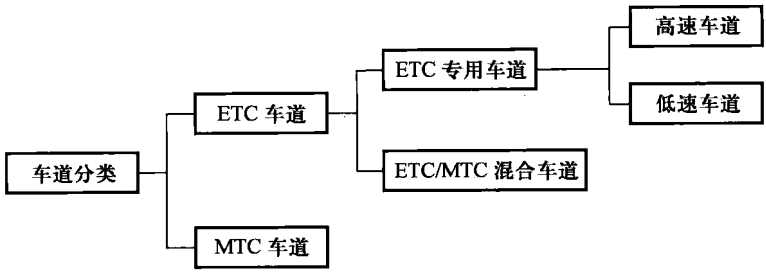


图 2-23 高速公路车道类型

务的分布具有中心车道倾斜性，即在收费服务机率相等时，驾驶员习惯于或倾向于驶入中心收费车道收费. 另外，由于内侧车道为快速车道，当 ETC 车道设置在内侧时能保持与公路相同的车道顺序，符合驾驶员的驾驶习惯. 这种设置方式有利于 ETC 车辆的通行，对收费站的改动较小，但误入 ETC 车道的车辆较多.

(2) ETC 车道外置式. 这种设计方式专用性强，在营运初期误入 ETC 车道的无效车辆少，但不符合公路车道的顺序原则，在营运中后期将会对 ETC 车辆产生较大的不便.

车辆合流区也是收费站车流运行过程中容易出现瓶颈的地段，因此收费亭后加速车道设置的合理与否对收费站的运行状况会产生很大影响. 加速车道较长，驾驶员会有更多的时间来寻找可插入间隙，在汇入车流前，驾驶员改变速度的幅度较小，舒适度和安全性提高，降低合流车辆的延误. 加速车道过短会影响车流的通过效率，过长则会造成建设资金的浪费. 因此，加速车道设置的长度及其通行能力也是收费站设置的重要组成部分.

3. 优化目标函数

收费亭数目的增加可以提高服务效率，减少顾客的等待时间以及等待所带来的其他不良影响，如燃料消耗、环境污染等. 由于收费站的服务效率低所带来的这些负面影响，统称为顾客等待费用. 收费亭数目的增加会增加收费站系统的建设成本和运营成本，这些成本统称为收费站的服务成本（费用）. 如何综合考虑这两方面的费用，从而使两类费用之和为最小，是收费站设置应考虑的问题. 收费站系统的费用合成如图 2-24 所示.

收费站服务机构的成本一般可以计算或估计出来，但车辆等待费用的估算则相对较难. 大多论文作者用车辆通过收费站的时间来度量，而时间与顾客的满意度之间的关系还需要进一步刻画. 另外，交通延迟所产生的费用不仅体现在时间上，还体现在能源消耗、空气质量等方面. 因此，需要建立这些因素与收费站设置方案之间的关系模型，在此基础上综合考虑时间、能源、环境等多方面的影响，建立相应的综合评价模型.

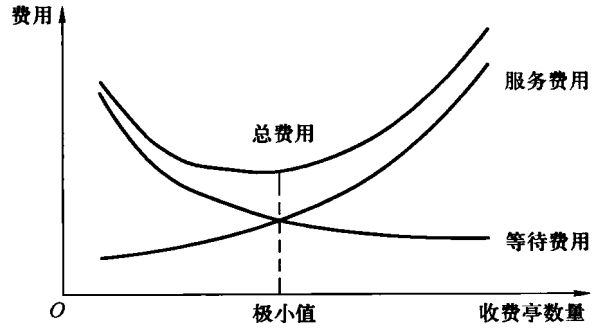


图 2-24　收费站的总费用合成示意图

4. 交通影响模型

目前，中国汽车消耗的汽油占我国汽油总产量的 90%，汽油消耗和废气排放已成为有效评估交通管理策略的重要组成部分. 自 1970 年起，研究人员在燃料消耗方面已经进行了大量的研究，得出了许多燃料消耗模型. 通过在不同驾驶条件下收集来的大量燃料消耗数据可以建立相应的数学模型，进一步研究相关的问题.

假设给定车辆、驾驶员的特性和环境状况，就可以建立交通流相关因素对燃料消耗的影响模型. 城市交通和公路交通中的燃料消耗具有一定的差异. 许多研究都表明，城市交通中单位距离的燃料消耗与平均速度的倒数近似成线性关系，燃料效率随平均速度提高而增加，当速度大于 55 km/h 时，空气阻力对燃料消耗的影响也将逐渐增加. 由于公路条件下车辆平均速度较高，当平均速度超过 55 km/h 时，必须把阻力作为重点来考虑. 对于车辆通过收费站时的燃料消耗，可以通过重新标定相关模型的参数来确定.

影响汽车排放物的主要因素是行驶状况及车速. 污染物质排放量随着车速的减小而增大. 行车速度很低时，发动机转速较低，燃料不能充分燃烧，所以 CO 和 HC 排放量明显增大. 一般情况下，在怠速或低速运行下的车辆，其 CO 和 HC 排放量是正常运转下的 2 ~ 3 倍.

有许多成熟的方法来评价空气污染水平，包括简单的线性源高斯模型和数字化模型（TRB 1981）. 大部分模型中运用的数据主要包括气象数据和交通数据. 气象数据包括风速和风向的变化以及温度的变化. 交通数据包括交通流、车速、车长和车型、地型等. UMTA 模型是空气质量模型中最简单的，也是与车速和排放水平相关的模型. 虽然该模型不太精确，但是计算速度快. CAL3HQC 模型能够说明车辆在道路交叉口附近的排放情况，怠速排放是整个交叉口排放的主要原因. 由于收费站附近车辆大都处于怠速状态，对空气质量问题的研究可以借鉴这些已有模型.

5. 仿真模型的校验

在国内外仿真系统开发过程中都会面临这样一个问题，如何保证开发出的

仿真系统的仿真结果是可信的，这就需要通过微观仿真系统标定、验证与确认来保证，可以简写为 V&V.

标定与验证的概念出现于 20 世纪 60 年代. 模型的标定是指通过各种方法对模型中关键变量、参数及随机变量的确认. 模型的确认具有特定的含义，即通过比较在相同输入条件和运行环境下模型与实际系统输出之间的一致性，评价模型的可信度或适用性. 而模型验证则是判断模型的计算机实现是否正确. 尽管确认和验证在各个文献中的定义不尽相同，但对于二者之间的区别，专家的看法却是基本一致的. 简单地说，模型确认强调仿真的理论模型与实际系统之间的一致性，模型验证则强调理论模型与计算机程序之间的一致性. 微观仿真系统标定与验证是一个复杂的过程，贯穿于微观仿真系统开发的始终.

参考文献

[1] Paul J Campbell. UMAP Journal [J] . COMAP, Inc, 2005, 26 (3): 190-191.

[2] ENO Foundation for Transportation. Transportation in America [R]. Transportation policy Assoc. contained in: National Transportation Statistics, D O T. Report DOT-VNFEC RSPA 92 1, 1992.

[3] Rillings J H. Automated highways [J]. Scientific American, 1997, 365: 60-63.

[4] 唐孝威，张训生，陆坤权. 交通流与颗粒流 [M]. 杭州：浙江大学出版社，2004.

[5] 李增光. 高速公路封闭式收费系统计费模型与拆账方法研究 [D]. 长沙：长沙理工大学，2004.

[6] 高速公路之美 [EB/OL]. [2011-10-01]. http://210.65.0.170/photogallery/? photoid =22.

[7] 张智勇. 基于 M/G/K 排队模型的北京地区高速公路收费站通行能力研究 [J]. 公路，2001 (7).

[8] 赵智明. 高速公路跨区域电子联网收费系统的研究与实现 [D]. 北京：华北电力大学，2006.

[9] 陈亚妮. 组合式电子收费系统的研究及车道优化设计 [D]. 长沙：长沙理工大学，2008.

[10] Chowdlmry D, Sauteu L, Schadschneider A. Statistical physics of vehicular traffic and some related systems [J]. Phys. Rep, 2000, 329: 199-329.

[11] TRANSIMS. Transportation Analysis and Simulation System [EB/OL]. [2005-01-01]. Los Alamos National Laboratory, Los Alamos, http://transims. tsasa. lanl. gov/.

[12] Helbing D, et al. Master macroscopic traffic simulation based on a gas-kinetic, non-local traffic model [J] . Transportation Research Part B, 2001, 35: 180-211.

[13] Lighthill M J, Whitham G B. On kinematics waves: I. Flood movement in long river [C]. Proceedings of the Royal Society, Series A, London, 1995.

[14] Lighthill M J, Whitham G B. On kinematics waves: II. A theory of traffic flow on long

crowded roads [C]. Proceedings of the Royal Society, Series A, London, 1995.

[15] Greenshields B D. A study in traffic capacity [C]. Proceedings of Highway Research Board, 1935, 14, 447-448.

[16] Greenberg H. An analysis of traffic flow [J]. Operations Research, 1959, 7 (1): 79-85.

[17] J Medhi. Stochastic Models in Queuing Theory [M]. New York: Academic Press, Elsevier Science, 2003.

[18] P Bocharov, Apice C D, Pechenkin A, et al. Modern Probability and Statistics: Queuing Theory [M]. Utrecht: Brill Academic Publishers, 2004.

[19] David R MaDonald Jr and Robert E, et al. Contribution to the development of guidelines for toll plaza design [J]. Journal of Transportation Engineering, 2001, 127 (3): 215-222.

[20] 邵春福. 交通流理论与方法 [M]. 北京: 中国铁道出版社, 2010.

[21] Chen B, Jin W D, Gao L. Numerical simulation and analysis of car following model on particular progress [J]. China Journal of Highway and Transport, 2005, 18 (1): 90-94.

[22] Kometani, Sasaki T. Dynamic behavior of traffic with a nonlinear spacing speed relationship [C]. General Motors In Proceedings of the Symposium on Theory of Traffic Flow. New York: Elsevier, 1959.

[23] Aycin M F, Benekohal R F. Comparison of car-following models for simulation [C]. The 78th Annual Meetings of TRB, 1999.

[24] Orlando-Orange Country Expressway Authority [EB/OL]. E-Pass. http://www.expresswayauthority.com/trafficstatistics/epass.html,2004.

[25] Edie A C. Traffic delays at toll booths [J]. Journal of Operations Research Society of America, 1954, 2: 107-138.

[26] Texas Transportation Institute. Houston's travel rate improvement program [R]. http://mobility.tamu.edu/ums/trip/toolbox/increase system efficiency.pdf,2001.

[27] 刘运通, 石建军, 熊辉. 交通系统仿真技术 [M]. 北京: 人民交通出版社, 2002.

[28] Nagel K, Schreckenberg M. A cellular automaton model for freeway traffic [J]. Phys. I (France), 1992, 2: 2 221-2 229.

[29] Chowdhury D, Wolf D E. Particle hopping models for two-lane traffic with two kinds of vehicles: effects of lane-changing rules [J]. Physics A, 1997, 235 (3): 417-439.

[30] Klodzinski Jack, Haitham A D. New methodology for defining level of service at toll plazas [J]. Journal of Transportation Engineering, 2002, 128 (2): 173-181.

[31] Prigogine I, Herman R. Kinetic theory of vehicular traffic [M]. New York: Elsevier, 1971.

[32] Garber, Nicholas J, Lester A H. Traffic and Highway Engineering [M]. CA: Brady/Cole Publishing Company, 1999.

第3章　不可再生资源的管理问题

不可再生资源的管理问题是2005年ICM竞赛题，研究的是资源保护领域如何更好地保护不可再生资源问题．由于不可再生资源总量是有限的，更是人类赖以生存的三大类资源之一，所以不可再生资源的合理开发与利用是实际研究中一个非常重要的课题．从数学建模的角度分析研究不可再生资源的优化管理问题也是很有实际应用价值的工作．

3.1　问题的综述

3.1.1　问题的提出

选择一种对人类生存和发展至关重要的不可再生或者即将耗尽的资源（例如水资源、矿产资源、能源、食物等），找到有关这种资源在世界范围内的预计储量、探明储量、年度消耗量和价格等数据．需要完成的任务如下：

（1）使用预计储量、探明储量和年度消耗量数据，运用资源建模方法，对世界范围内的资源耗尽问题或者资源储量减少问题进行建模分析．

（2）考虑未来经济、人口、政策和环境等因素的影响，调整你的数学模型，要求模型有足够的细节，提供可视化的结论，并讨论模型的局限性．

（3）提出一个公平、可操作的开采/管理策略，例如经济激励和惩罚措施，从而确保资源的长期可用性，避免资源极度混乱的消耗、退化或迅速枯竭．

（4）提出一种安全开采策略以保护资源，防止偷盗、滥用、破坏、不必要的退化或耗尽．此外，讨论与这些策略相应的政策和安全管理方面的可行方案．

（5）资源开采过程可能带来短期或者长期的环境危害，请提出一种控制这种危害的有效策略．制定策略时需要考虑：环境污染问题、对自然灾害的敏感度、资源废料的处理和存放以及其他你们认为重要的因素．

（6）将这种资源和用途相近的其他资源进行比较，为减少这种资源的消

耗，应该发展什么样的新科学或技术？提出一种能推动这些新领域的研发策略.

不可再生资源管理问题的原文如下：

The Exhaustible Resource Problem

Select a vital nonrenewable or exhaustible resource (water, mineral, energy, food, etc.) for which your team can find appropriate world-wide historic data on its endowment, discovery, annual consumption, and price.

The modeling tasks are:

(1) Using the endowment, discoveries, and consumption data, model the depletion or degradation of the commodity over a long horizon using resource modeling principles.

(2) Adjust the model to account for future economic, demographic, political and environmental factors. Be sure to reveal the details of your model, provide visualizations of the model's output, and explain limitations of the model.

(3) Create a fair, practical "harvesting/management" policy that may include economic incentives or disincentives, which sustain the usage over a long period of time while avoiding severe disruption of consumption, degradation or rapid exhaustion of the resource.

(4) Develop a "security" policy that protects the resource against theft, misuse, disruption, and unnecessary degradation or destruction of the resource. Other issues that may need to be addressed are political and security management alternatives associated with these policies.

(5) Develop policies to control any short-or long-term "environmental effects" of the harvesting. Be sure to include issues such as pollutants, increased susceptibility to natural disasters, waste handling and storage, and other factors you deem appropriate.

(6) Compare this resource with any other alternatives for its purpose. What new science or technologies could be developed to mitigate the use and potential exhaustion of this resource? Develop a research policy to advance these new areas.

3.1.2 竞赛的基本情况

2005年ICM（第7届ICM）共有来自4个国家88所大学的164个队参赛，本届竞赛从美国东部时间2月3日20：00（北京时间2月4日9：00）开始，到2月7日20：00（北京时间2月8日9：00）结束. 该问题的命题人是美国比洛特（Beloit）学院数学与计算机科学专业教授、UMAP杂志编辑 Paul

Campbell，以及该校地球科学与土木工程师 Ted Hromadaka 和他的同事们.

在竞赛后的评阅过程中，文章的摘要、模型叙述和全文的组织结构是判断一篇论文是否优秀的主要标准. 对于评阅中出现的意见分歧，评阅人将充分交流讨论，以求实现意见统一，如果无法统一，则再增加一个评阅人进行评阅.

2005 年 ICM 评阅出 3 份特等奖（Outstanding）论文，分别是：

（1）中国华东理工大学

论文题目：The Coming Oil Crisis

论文作者：Wei Deling，Chen Jie，Xu Hui

指导老师：Ni Zhongxin

（2）美国欧林工程学院（INFORMS 奖获得者）

论文题目：Preventing the Hydrocalypse：A Model for Predicting and Managing Worldwide Water Resources

论文作者：Steven Krumholz，Frances Haugen，Daniel Lindquist

指导老师：Burt Tilley

（3）美国 Maggie Walker Governor's School

论文题目：The Petroleum Armageddon

论文作者：Jonathan Giuffrida，Palmer Mebane，Daniel Lackek

指导老师：John Barnes

其他奖项的获奖数量及比例如表 3-1 所示.

表 3-1 2005 年 ICM 获奖情况统计表

	Outstanding	Meritorious	Honorable Mention	Successful Participation	总计
参赛队个数	3	26	89	46	164
百分比	1.8%	15.9%	54.3%	28%	100%

不可再生资源管理问题，要求 3 名本科生在 4 天时间内完成具有很大的挑战性. 要求参赛队选择一种不可再生或即将耗尽的资源，建立这种资源的损耗模型. 在问题的研究过程中，除了需要对赛题中的具体问题给出科学和定量的分析外，还需要考虑有关经济、人口、政策、环境、安全和相关技术细节等定性的问题.

相对 MCM 以及往届 ICM，这次 ICM 赛题的开放性更强，不同的参赛队可能选择具有不同特性的不可再生资源展开研究，从而导致不同的模型之间往往不具备可比性.

3.2　问题的背景资料

3.2.1　不可再生资源概念

不可再生资源是指不能按照人们消耗的速度进行制造、种植、繁殖或者使用的自然资源. 不可再生资源一旦消耗殆尽，日后将无法再次获得. 从广义上讲，消耗速度远远大于自然生产速度的自然资源也被称作不可再生资源. 例如化石燃料（煤炭、石油和天然气），各种核燃料（铀）以及地下水资源等都是不可再生资源. 但是木材（特指能够保持生态平衡的砍伐和生产），或者金属（能够回收冶炼）不属于不可再生资源[1].

3.2.2　Hubbert 峰值理论

Hubbert 峰值理论（Hubbert peak theory）[2] 是美国地理学家 M. King Hubbert 提出的，该理论指出：针对一个特定的地理区域，不论是一个独立的原油产区还是整个世界，石油开采速度与时间的关系往往遵循钟形曲线，这是研究石油峰值产量的主要理论之一.

Hubbert 钟形曲线的函数表达式为

$$x=\frac{e^{-t}}{(1+e^{-t})^2}.$$

该函数是 logistic 分布函数的密度函数，与正态分布函数的形态略有不同，如图 3-1 所示.

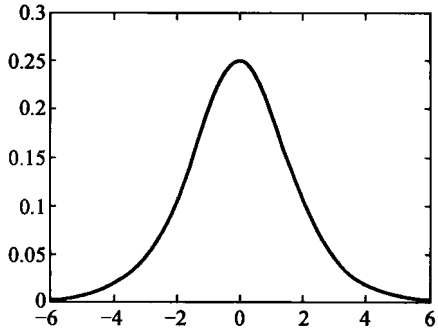


图 3-1　Hubbert 曲线

在 Hubbert 峰值理论中，基于石油勘探速率、开采速率以及累积开采量，确定了一条能够确定石油最大产量的曲线. 曲线的前半部分（峰值前）反映的是开采速率会因为勘探速率的提高以及石油开采的基础设施建设的发展而逐步提高；曲线的后面部分（峰值后）反映的是开采速率会因为资源的枯竭而

逐步降低.

Hubbert 峰值理论是基于任何地区的可探明地下石油资源是有限的这一假设，所以在初期，可勘探的资源量比较大，石油的勘探速度增长较快，当达到最大值时，未勘探石油资源逐步变小，此时勘探速度开始减小. 在美国，石油的开采时间落后于石油勘探曲线 32～35 年.

Hubbert 峰值是指已发现油气资源地区的年产量峰值. Hubbert 在 1956 年发表的一篇论文中预测，美国大陆石油的 Hubbert 峰值将出现在 1965—1970 年间（如图 3-2 所示），并且预测全球石油的 Hubbert 峰值将出现在半个世纪后（2020 年左右），那时的年石油产量将达到 1.2×10^{10} 桶（如图 3-3 所示）.

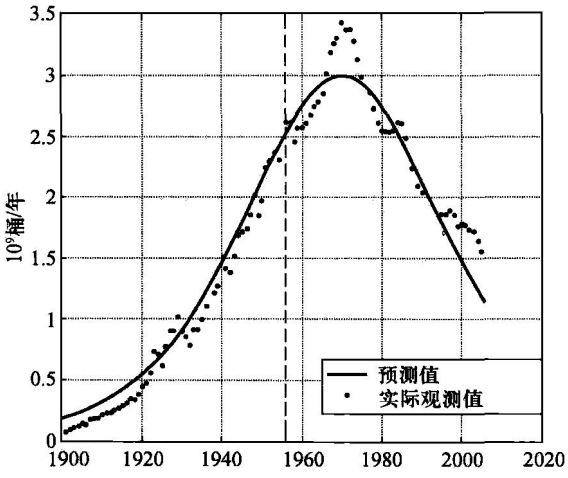


图 3-2 Hubbert 于 1956 年预测的美国石油年度开采量

从预测结果和实际观察值的对比来看，美国在 20 世纪 70 年代早期达到陆地石油资源的 Hubbert 峰值. 当时每天的石油产量是 1.02×10^8 桶（约合 162 万立方米）. 自那以后，美国陆地石油产量开始逐渐减少. 从这个数据上看，Hubbert 比较准确地预测了美国的石油年度产量.

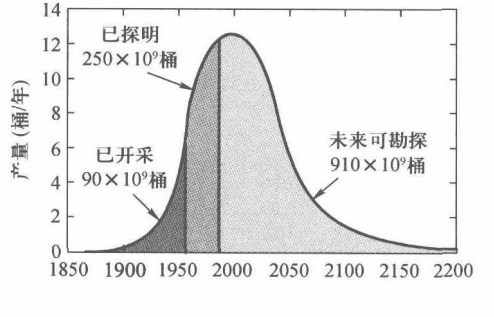


图 3-3 Hubbert 于 1956 年预测的全球石油产量

在1976年的一次电视采访中，Hubbert指出，OPEC的各种努力可以使Hubbert曲线更平缓一些，峰值后推一些，但最多也只能将峰值推后10年左右.

虽然Hubbert峰值理论主要应用于石油资源产量的峰值预测，但也有一些使用Hubbert峰值理论对其他自然资源产量峰值预测的研究. 如Doug Reynolds于2005年预测北美天然气产量的峰值将出现在2007年；Bentley预测全球天然气产量的峰值将出现在2020年.

Hubbert在研究美国无烟煤的年产量规律时，发现峰值发生在1920年，美国宾夕法尼亚的煤炭年产量也十分符合Hubbert曲线. Hubbert预测全球的煤炭资源可以维持到2150年，这并不是说这些地区的煤炭资源枯竭了，而是因为石油的使用场合远远大于煤炭的使用场合，现阶段更多的人更愿意使用石油，对煤炭的需求有所减少，所以煤炭的Hubbert峰值提早出现了. 如果将来石油资源消耗殆尽，又没有合适的替代能源，可以预见，煤炭资源的年度开采量将再次形成一个Hubbert峰值.

Hubbert峰值理论最初主要用于研究不可再生资源的消耗问题，无法应用于可再生资源消耗的研究，但是一些过度开采的可再生资源也往往呈现出Hubbert曲线的特征，修正后的Hubbert曲线可以应用于任何开采速度远远大于再生速度的资源消耗预测.

例如地下水资源，世界上大多数国家的地下水资源的开采速度都大于再生速度，这使得可用淡水资源（地下水、湖泊和河流）变为有限资源，其消耗规律类似于石油资源. 目前针对地下水资源消耗的研究主要集中在农业灌溉用水、生活用水等，针对诸如核电、煤炭开采、沥青砂开采等水资源密集型的工业耗水的研究并不多.

3.2.3 Michael Olinick理论

固定消耗储备量：指某种资源在总供应量S固定的前提下，每年的消耗量U也固定，该资源可用的年数可表示为S/U.

指数消耗储备量：指某种资源在总供应量S固定的前提下，初始年度消耗量为U，每年使用量的增长率为r时（此时该资源的年度消耗量呈指数增长），该资源的可用年数可表示为

$$\frac{\ln\left(1+\frac{S}{U}\right)}{\ln(1+r)}.$$

按照上述定义，2005年美国的煤炭资源按照当时的消耗速度将可以维持250年，而如果按照年度消耗增长率2.25%，美国的煤炭资源只能维持85年.

美国地质勘探局指出，目前世界石油资源和天然气资源的年度消耗量的增

长率分别为1.9%和2.2%，美国石油资源的年度消耗量从1973年起一直保持2.5%的增长，并且据预测，这种增长率将一直维持到2025年左右.

3.2.4 需求价格弹性理论

需求价格弹性（Price Elasticity of Demand）理论，在经济学中一般用来衡量需求的数量随商品价格的变动而变动的情况[3]. 假设 Q 为某个商品的需求量，P 为该商品的价格，则需求的价格弹性值 E_d 为

$$E_d=\frac{\Delta Q/Q}{\Delta P/P}=\frac{\Delta Q}{\Delta P}\cdot\frac{P}{Q},$$

其连续形式为

$$E_d=\frac{\partial Q}{\partial P}\cdot\frac{P}{Q}.$$

某种商品的需求价格弹性曲线如图3-4所示. 可以看出，需求随价格的降低而增加.

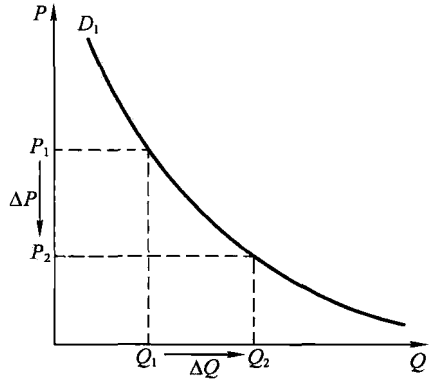


图3-4 某种商品的需求价格曲线

通过求解上述微分方程，可以得到

$$Q=Ce^{E_d\log P},$$

其中 C 是由微分方程初值决定的一个常量.

通常，因为商品价格的下跌会导致需求数量的增加，所以一般情况下需求价格弹性系数为负数. 需求价格弹性越大，表明需求对于价格越敏感. $E_d=0$ 时表示完全无弹性，$E_d=\infty$ 时表示完全弹性.

影响需求价格弹性的因素包括替代品的品种、该商品消费占支出的比例、商品的耐用性和是否是必需品. 表3-2给出了常见商品的需求价格弹性指数.

从表3-2中可以看出，美国的石油需求价格弹性指数远高于世界平均水平，与烟草、咖啡的需求价格弹性指数接近. 该结果一方面说明美国对石油需求的依赖性大于世界平均水平，另一方面说明美国对石油价格的敏感性不高，石油的支出占所有可用支出的比例小于世界的平均水平，价格的波动并不会在

很大程度上影响美国人的石油消费. 上述结果只是针对现阶段的数据统计得到的. 一旦石油资源供应紧张，价格升高到美国民众无法承受的程度，美国的石油需求价格弹性指数将逐步改变到-0.45 左右.

表 3-2　常见商品需求价格弹性指数[3]

商品名称	烟草	大米			石油		汽车燃油		钢铁
区域或类别	美国	中国	日本	美国	美国	全球	短途	长途	全球
E_d	-0.06	-0.8	-0.25	-0.55	-0.06	-0.4	-0.25	-0.64	-0.25

商品名称	酒类			饮料			看电影	鸡肉
区域或类别	啤酒	葡萄酒	烈酒	可口可乐	威士忌	咖啡	美国	美国
E_d	-0.7	-1.0	-1.5	-3.8	-4.4	-0.07	-0.87	-0.5

通过需求价格弹性指数可以建立价格与需求量之间的关系，也为通过价格机制调整需求，实现节约使用不可再生资源的调控政策的科学性提供了理论支持.

3.2.5　OPEC 对石油供求关系的影响

石油输出国组织（Organization of Petroleum Exporting Countries）成立于1960 年，简称 OPEC. 其宗旨是协调和统一成员国的石油政策，维护各自的和共同的利益，现有 12 个成员国. 2003 年该组织成员国石油的总储量占世界石油储量的 69%，产量占全世界的 39%.

OPEC 组织垄断了全球大部分原油资源，在一定程度上操纵着国际原油价格，也极大地影响了石油的需求量. OPEC 通过控制其成员国的石油产量和价格竞争机制调控原油价格，从而实现其成员国利益的最大化.

但是，尽管原油价格已经从 2002 年的 20 美元/桶升高到 2005 年的 50 美元/桶，但是 OPEC 的收入也远比 1980 年（按照 2004 年美元购买力计算，相当于 66 美元/桶）的少了很多. 例如沙特阿拉伯的人均总收入从 1980 年到 1999 年下降了 70%，达到了历史最低点.

此外，原油的价格通常以美元结算，从 2002 年开始美元持续贬值，从而导致了石油价格需要上涨 30% 才能抵消美元通胀. 对 OPEC 而言，最关心的问题是如何从石油生产中获得最大的利益. OPEC 会综合考虑货币通胀、石油资源的剩余量等安排成员国的生产，一方面将当前收益保持在一个可接受的水平，另一方面极力维护自身的石油资源开采量，从而实现可持续开采.

如果 OPEC 原油开采量过大，将导致石油价格下降，一方面可能无法保证总收入的最大化，也可能导致石油资源的提前枯竭. 如果 OPEC 开采量过小，可能抬高价格，但也将导致需求量降低，从而影响总收益. OPEC 需要根据通胀率、

需求价格弹性指数、短期经济发展和长远发展等因素确定最佳的开采量.

3.3 问题的分析与假设

不可再生资源的优化管理问题要求使用数学建模的方法研究下面的6个问题:

(1) 预测不可再生资源的枯竭日期;

(2) 多因素时不可再生资源耗尽时间的预测;

(3) 资源的公平开采策略;

(4) 资源的可持续开采策略;

(5) 考虑环境因素的资源开采策略;

(6) 存在替代资源的开采策略.

3.3.1 问题 (1): 预测不可再生资源的枯竭日期

这个问题是要求解决的第一个任务, 即要求根据预计储量、探明储量和年度消耗量的数据, 对世界范围内的某种不可再生资源的耗尽日期进行预测分析.

不可再生资源的耗尽时间可以根据预计总储量和预测的年度消耗量计算得到, 将问题转化为如何对总储量和年度消耗量的预测问题. 随着科学技术的发展, 地球上各类资源的探明储量基本接近预计的总储量, 即预计储量是一个相对稳定的值, 此时影响资源耗尽时间的主要因素是不可再生资源的年度消耗量.

不可再生资源的年度消耗量与诸多因素有关, 例如经济发展对该资源的依赖性, 是否存在替代资源, 其当前储有量、年度开采量、开采成本和价格等. 为了得到该资源的年度消耗量, 可以采用如下两种方法:

(1) 直接利用历史数据进行拟合预测;

(2) 建立影响资源年度消耗量与各个因素之间的微分方程模型, 同步预测各个因素发展状态.

1. 数据拟合预测法

Hubbert 在 1956 年使用 Hubbert 曲线对美国以及世界石油年度消耗量进行了预测, 事实证明, Hubbert 得到了相当精确的预测结果.

使用 Hubbert 曲线对资源年度消耗量进行预测的好处是绕过了各个复杂细节对资源消耗的影响, 直接从结果上寻找规律. 同时这也是拟合预测方法的缺陷, 该模型无法准确地解释资源消耗的本质规律, 无法根据现实中各种条件的改变而对资源年度消耗量进行调整. 例如它无法解释 OPEC 对石油年度消费量的影响, 也无法预测战争、自然灾害等因素对资源消耗的影响.

2. 供求关系微分方程法

资源的年度消耗量与诸多因素有关，3.3.2 节将研究其与经济、人口、政策和环境因素的关系，因此在这里暂不研究这些因素的影响. 影响资源年度消耗量的最根本因素是价格，价格的上涨必将导致资源消耗量的减少. 不同资源的价格制定方法不同.

对于石油、煤炭资源来说，价格受供求关系的影响. 供大于求导致价格下降，供不应求导致价格上涨. 这里的“需求”是指对资源的潜在需求，由经济发展或人们生活对该资源的依赖性决定，它是一条相对光滑的曲线，不会因为价格和供应量的改变而波动.

当供应量与需求量发生偏移时，就可能产生价格波动. 当供应量低于需求量时，为了通过价格竞争获得资源，导致资源价格上涨. 价格的上涨引起更多的资金投入资源开采，从而增加供应量. 由于从投产到收益有一个时间周期，因此往往会导致供大于求，最终导致价格下降；如果资金撤离资源开采，又将会导致供不应求，如此反复，如图 3-5 所示.

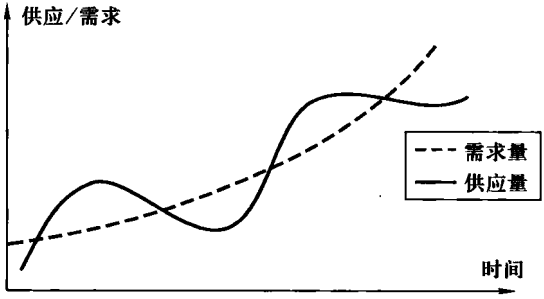


图 3-5　供应-需求曲线示意图

一般来说，不可再生资源的存储成本很高，可以认为每年的供应量等于实际消耗量. 因此图 3-5 中的供应量就是其真实消耗量，对消耗量的预测就转化为根据供求关系对供应量的预测. 对于剩余储量较大、市场化程度较高的资源，例如煤炭、木材等资源，供应量基本围绕需求量上下波动.

该模型中，供应量与需求量的差值决定了价格，而价格决定了供应量. 在一些基本假设下，对资源的需求量进行预测（例如指数增长），建立上述关系的微分方程模型（例如价格的改变量与供求差额成线性关系，供应量与价格成线性关系），即可得到资源的年度供应量的预测模型.

3. “剩余资源-价格”微分方程法

上述模型只适合短期预测，随着资源开采的深入，当资源的剩余量接近枯竭时，价格往往不仅仅受供求关系的制约，还受到剩余资源的影响.

例如当石油资源即将消耗殆尽时，价格必将攀升. 假设石油价格与石油剩余储量成反比关系，即剩余储量越小，价格越高. 根据该假设，对历史数据进

行回归，即可估计出反比例系数，此时根据预测的剩余储量即可推算出各年的石油价格. 针对稀缺资源而言，高价格并不会吸引大量的投资导致产业的产量增加，因为此时已经没有足够的石油资源可供开采，价格将影响石油贸易的成交量，即实际消费量.

根据需求的价格弹性原理，价格越高，需求越低. 这里的需求不是指潜在需求量，而是指石油贸易的实际年度成交量（即年度开采量），人们对石油的潜在需求只受经济发展和替代能源的影响. 文献［3］中指出，全球的石油需求价格弹性指数是-0.4，据此可以根据价格建立年度开采量的关系模型. 根据年度开采量，可以进一步推算各年的石油资源剩余储量，从而计算下一年度的石油价格，反复上述过程即可推算各年的石油开采量.

对于水资源来说，水资源定价受国家调控的影响，国家根据水资源的充沛程度以及人们的收入制定一个合理的价格，该价格能够兼顾人们生活质量和水资源的保护. 水资源定价时，需要针对不同的用途分别定价，以保证人们的基本生活用水.

水资源的定价取决于剩余水资源量，这一点与石油资源类似；不同的是，政府定价可以调控水价对剩余资源的敏感程度，提早防范水资源的枯竭. 不同的用水途径，其需求价格弹性指数是不一样的，一般认为，农业用水的需求价格弹性高于工业和生活用水，发达国家一般在-0.5～-1.4之间，甚至有研究显示该弹性高达-3.0[4]. 但对于发展中国家，由于水价偏低，因此价格弹性表现较小，说明农民对低水价不敏感，但适当提高水价可导致需水量大量减少[5].

政府根据水资源的危急程度（剩余水资源量）对不同的用水途径和不同的需求价格弹性指数制定不同的水价，根据需求价格弹性原理，可以计算出不同用水途径的用水量，从而获得各年的水资源消耗量.

3.3.2 问题（2）：多因素时不可再生资源耗尽时间的预测

这个问题需要综合考虑人口、经济、政策和环境等因素修正上述模型. 在上一节中，若采用反映供求关系的微分方程法，则可以在其基础上添加一些因素，从而得到多因素的预测模型.

1. 多因素时供求关系微分方程模型

在供求关系微分方程模型中，有一个关键量——潜在的需求量变化曲线. 人口、经济、政策和环境等因素将通过改变潜在需求量的变化曲线影响年度的消费量.

人口的增长、经济的发展将导致潜在需求变化曲线上升. 对于水资源等民生资源，潜在需求量与人口的关系更密切，对于石油、煤炭等工业资源的潜在需求量更受GDP的影响. 于是可以假设潜在需求量 Q 满足关系：

$$Q=k\cdot P(t)^{k_1}\cdot[G(t)\cdot R(t)]^{k_2},$$

其中 k 是比例系数，$P(t)$ 是 t 时刻的人口数，$G(t)$ 是 t 时刻的 GDP，$R(t)$ 是单位 GDP 所消耗的资源量，该值反映了国家的资源利用率. k_1 和 k_2 是权重系数，分别反映了人口和 GDP 对潜在需求量的影响的大小，对于水资源可取 $k_1>k_2$，对于石油和煤炭资源可取 $k_1<k_2$.

根据历史数据，拟合得到比例系数 k 和权重系数 k_1 和 k_2，再对人口和 GDP 进行预测，从而可以计算得到各年的潜在需求量，代入 3.3.1 节中的供求关系模型，即可得到考虑人口和经济因素的供求关系微分方程模型.

考虑资源消耗产生的排放物对环境的污染以及资源的不可再生性，政府部门可以通过制定消费税的方式控制资源的消费. 增加消费税后，资源价格上涨，但这些收入并没有流向资源开采商手中，因此无法吸收更多的投资，也不会导致供应量的增加.

当考虑政策和环境因素后，成交价格分为两部分，一部分是市场价格部分，另一部分是消费税部分. 市场价格部分仍然由供求关系决定，消费税的比例由政府调控决定. 供应量的增加量由市场价格决定，而不是成交价格.

按照上述分析，通过人口和 GDP 的预测修改潜在需求量变化曲线，通过消费税修正微分方程模型，即可得到考虑多因素时的供求关系微分方程模型. 适当调整各个参数，计算模型结果，即可得到不同人口、GDP 增长速度下，不同消费税征收比例下的资源消耗的预测结果.

2. 多因素时“剩余资源-价格”微分方程模型

在“剩余资源-价格”模型中，价格与剩余资源成反比例关系. 考虑到人口和 GDP 对资源的依赖性，可以认为人口和 GDP 将影响反比例系数. 人口越多，GDP 越高，单位 GDP 耗费资源就越多，则价格对剩余资源的敏感性也就越大. 实际上可以将人口和经济因素看成是对资源的潜在竞争需求，人口越多，经济发展就越依赖于资源，则对该资源的竞争也就越激烈，从而导致价格对剩余资源量越敏感.

同理，政策和环境因素也可以体现在反比例系数中.

针对石油资源，其“剩余资源-价格”关系具有特殊性. 石油资源将在近几十年内消耗殆尽，并存在分布不平衡的现象. 资源的稀缺性和分布的不平衡性都导致了石油资源成为垄断商品，垄断商品的供求关系不再满足自然市场的供求规律. 石油资源的最大垄断商是 OPEC，OPEC 并不会因为石油供不应求就增大石油供应量，而是从长远的角度出发规划石油资源的开发，实现其成员国利益的最大化. 实际上，世界的石油价格基本被 OPEC 操控.

根据当前的人口、GDP 以及单位 GDP 对石油的消耗需求，可以得到全世界对石油资源的潜在需求量. 这个潜在需求量和剩余石油资源量影响了人们对石油资源的需求价格弹性指数，需求量越大，需求价格弹性指数越小.

根据当前人口和经济条件，OPEC 可以计算出当前石油的“需求价格弹性指数”，并根据该指数估计不同定价措施下全球的石油年度成交量. 不同的定价，可能导致不同的成交量，价格、开采成本和成交量都将影响 OPEC 成员国的收入. 安排的产量过低，可能导致价格过高，交易量减少，从而使得 OPEC 的年度收益减少；安排的产量过高，可能导致价格下降，从而导致年度总收益的下降. OPEC 需要根据当前的人口经济形势，制定最佳的价格和安排石油开采，实现自身收益的最大化.

3.3.3 问题 (3)：资源的公平开采策略

在研究资源的公平开采策略时，要综合考虑资源分布的不均衡性、资源利用率的不均衡性以及资源开采的利益分配问题. 自然资源是国家的私有资源，涉及国家利益. 一般来说，实现国家间自然资源的无条件配置是不现实的. 资源的公平开采策略问题，主要有两方面值得研究，一方面是如何通过国际合作实现资源的有序开采，并且合作各国共同分享合作的利益；另一方面是国内不同地区的资源开采配置.

1. 通过国际合作实现资源有序开采

OPEC 是国际合作实现资源有序开采的一个典型案例. 在 OPEC 组建之前，中东等石油输出国形成恶性竞争，盲目扩大产量，打压价格，以获得更多的眼前利益，从而导致石油资源极度混乱的消耗、退化甚至枯竭. OPEC 的形成在一定程度上制止了这种石油资源的极度混乱的消耗. OPEC 成员国合作形成一个经济体，与石油进口国开展价格博弈，并将博弈的利润合理地分配给各个成员国. 该合作形成的关键要素是合作利润的分配（即生产配额的规划）. 如果某个成员国不遵从 OPEC 规划，单独生产能够比合作生产获得更多的收益，则这种合作是不牢固的.

此外针对水资源的国际合作规划也有一些成功案例.

2. 国内不同地区资源的开采配置

对于一个国家不同地区的资源开采配置的自由度就要大得多. 不同地区的资源储量不同、交通条件不同、开采成本不同、开采利用率也不同，政府部门需要综合考虑开采成本和对环境的影响等因素，规划国内的资源开采策略，并制定相关政策对资源输出地进行经济补偿，用于修复资源开采对当地环境的破坏.

国内资源的开采配置模型与 OPEC 合作博弈模型类似，各个资源开采企业通过合作，制定最佳的资源开采策略，实现资源有序、平衡和可持续开采.

3.3.4 问题 (4)：资源的可持续开采策略

从理论上讲，越到开采后期，资源的剩余量越少，价格就越高，开采单位

资源所获得的利润也越高. 因此，从资源开采企业的总收益上看，将资源留到足够远的将来再开采可以获得更多的利润. 但是，对于大多数资源输出国而言，资源出售是这些国家的主要经济来源，尤其是中东地区的石油输出国. 为了维持当前经济发展的需要，这些国家必须保持每年一定的开采量.

假设一个资源输出国在保持当前生活水平不下降的基础上，综合考虑货币通胀和资源价格变化，可以计算出每年最少的资源开采量. 按照这个开采量规划生产，可以实现在保持生活水平的前提下，最大程度地延长资源开采时间，从而实现资源输出国利益的最大化.

对于资源自产自销型的国家，保持当前经济发展需要一定量的资源. 根据3.3.2节可以计算得到各个国家保持一定经济增速时的资源消耗量. 资源消耗量可通过两种途径获得，一种是国内资源的开采，另一种是资源进口. 国家需要综合考虑长远发展以及当前进口成本，规划本国的资源开采量以及进口量的最佳比例.

3.3.5 问题（5）：考虑环境因素的资源开采策略

对于不同的地区，资源开采对自然环境的破坏程度是不同的. 例如露天煤矿对环境的破坏程度要比地下煤矿严重得多，浅层地下煤矿对环境的破坏程度要比深层地下煤矿大. 在干旱地区取水对环境的影响要比在水资源充沛地区取水的影响严重.

正规的大型开采企业相比小型开采企业，其生产效率更高，开采成本更低，因此能够投入更多的资金保护开采地周边的环境，开采单位资源对自然环境的破坏就要更小. 使用新技术开采，一方面能够提高资源产出率，另一方面能够降低对自然环境的影响.

在考虑环境因素制定资源开采优化策略时，可以根据历史数据统计得到各类开采企业的开采产出率、开采单位资源用于环境保护的投入、企业的产能，以及各种资源的储备类型，建立规划模型，确定最佳的资源开采策略.

3.3.6 问题（6）：存在替代资源的开采策略

在存在替代资源情况下，各国对原始资源和替代资源总的消耗量的规律，可以通过3.3.1节和3.3.2节中的相关模型估计得到. 下面讨论在已知资源总消耗量的前提下，如何利用替代资源逐步替代原始资源.

在“剩余资源-价格”微分方程模型中，实际资源的成交量与价格相关，其关系式中有一个关键量——需求价格弹性指数. 当原始资源没有替代资源时，需求价格弹性指数将处于一个比较低的水平，价格波动对资源使用量的影响不大. 当存在替代资源时，原始资源涨价后，就会有消费者转向替代资源. 此时消费者对原始资源的需求量开始价格敏感，即需求价格弹性指数升高. 例

如在出租车行业中，在早期只有汽油燃料，不论汽油价格是否上涨，汽油消费基本保持不变. 当出现天然气燃料后，汽油价格上涨时，越来越多的出租车司机考虑使用天然气燃料.

随着资源日益枯竭，原始资源的价格日益升高；随着技术的进步，替代资源的成本日益降低. 根据需求价格弹性指数以及两类资源当前的价格，可以计算出人们对两类资源的需求量. 替代资源的需求量将逐步增加，最终实现完全替代原始资源.

3.4 石油资源的预测模型

本节以石油资源为研究对象，首先建立供求关系模型，然后叠加经济发展的指数规律得到石油消耗量与时间的关系式，并用该关系式进行外推，预测未来一段时间内的石油消耗量. 最后得到结论：若不采取有效措施，世界石油资源将在2032年耗尽. 在拟合预测的基础上，考虑经济、人口、政策和环境因素得到更为精确的预测模型.

为了满足当前人们的需求，同时不威胁子孙后代的能源需求，模型制定了一套石油资源在各代人之间配置的标准，并且用一个例子说明在该准则下如何实现最优配置. 模型还提供了一种延缓近期石油危机的开采策略. 最后，基于边际效用替换原理，给出了石油资源和替代能源的平衡关系. 由于该模型是基于供求关系理论和不可再生资源的本质规律，所以可以推广到其他不可再生资源的研究.

3.4.1 模型1.1：石油资源耗尽时间的预测模型

1. 问题的假设与符号说明

先给出模型的基本假设：

(1) 石油精炼能力能够满足实际需求.

(2) 在全球石油资源枯竭之前，只要有需求，就能供应，所有的石油都可以被开采.

这里 $U(t)$ 表示 t 年度尚未被探明的石油储量；$R(t)$ 表示 t 年度已探明储量中尚未被开采的石油；$D(t)$ 表示 t 年度石油的年需求量；$S(t)$ 表示 t 年度全世界石油供应量；$P(t)$ 表示 t 年度的石油价格；P_0 表示石油的平均价格.

2. 模型的建立与求解

由上述假设，t 年度全球尚存的石油资源总量可以表示为 $U(t)+R(t)=C$. 1997年预计未探明的全球石油储量为1 800亿桶[2]，已探明的石油储量是10 185亿桶[3]，即 $C=11\ 985$ 亿桶.

假设石油资源将在第 n 年消耗殆尽，则第 $t\sim n$ 年度的石油需求总量为 $\sum_{i=t}^{n} D(i)$，其中 n 满足：

$$\sum_{i=t}^{n} D(i) \leqslant U(t)+R(t) \leqslant \sum_{i=t}^{n+1} D(i)$$

要研究石油将在何时耗尽，需要先预测将来一段时间的石油需求量.

石油的需求量与价格密切相关，每次价格波动都会带来消耗量的剧烈波动. 下面建立基于价格机制的“供-需平衡”常微分方程模型.

设 $\widetilde{P}$ 是实际售价与石油价值 P_0 的差，即 $\widetilde{P}(t)=P(t)-P_0$，当 $\widetilde{P}(t)>0$ 时，说明石油价格高于平均价格，此时石油供应国将提高产量以获取更多的利润. 假设增产的速率与 $\widetilde{P}(t)$ 成正比，即

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t}=a\,\widetilde{P}(t).$$

石油价格受到供求关系的影响，当供应大于需求时，石油价格会出现下降，假设石油差价 $\widetilde{P}(t)$ 下降的速率与供求差 $S(t)-D(t)$ 成正比，即

$$\frac{\mathrm{d}\widetilde{P}(t)}{\mathrm{d}t}=-b[S(t)-D(t)],$$

石油的需求受到价格的影响，价格上涨将导致石油需求量的下降. 假设石油需求量的改变率与价格差 $\widetilde{P}(t)$ 成正比，即

$$\frac{\mathrm{d}D(t)}{\mathrm{d}t}=-c\,\widetilde{P}(t).$$

求解上述微分方程组，可得到通解：

$$\widetilde{P}(t)=k\sin(\omega t+\varphi), \tag{3.1}$$

$$S(t)=S_0-\frac{ak}{\omega}\cos(\omega t+\varphi), \tag{3.2}$$

$$D(t)=D_0+\frac{ak}{\omega}\cos(\omega t+\varphi), \tag{3.3}$$

其中 $k=\sqrt{\tilde{c}_1^2+\tilde{c}_2^2}$，$\varphi=\arctan\left(\dfrac{\tilde{c}_1}{\tilde{c}_2}\right)$，$\omega=\sqrt{b(a+c)}$，$S_0=D_0$，$\tilde{c}_1$，$\tilde{c}_2$ 是待定参数.

从式 (3.3) 可以看出，石油的需求量总是一个余弦周期函数，即石油的需求量总是在一个基准附近上下波动. 显然上述结论是不太合理的，随着工业化的发展，在没有替代能源的前提下，石油的需求量必将逐年递增. 因此该模型的作者认为：由于人口在一段时间内满足指数增长规律，所以石油的需求量

也应满足指数增长规律，即石油的实际需求量应该是在式（3.3）周期性波动的基础上叠加指数增长趋势，即将式（3.3）改造为

$$D(t)=a_1+a_2\cos(a_3t+a_4)+a_5\exp(a_6t). \tag{3.4}$$

基于上述分析，通过历史数据对式（3.4）进行拟合，求解参数 a_1，a_2，a_3，a_4，a_5，a_6 即可对将来的石油需求量进行预测.

1970—2004 年的石油需求量如表 3-3 所示（数据来源：美国能源信息部 2004）.

表 3-3 1970—2004 年石油消耗量 千桶/天

年度	消耗量	年度	消耗量	年度	消耗量	年度	消耗量
1970	46 808	1980	63 108	1990	66 443	2000	76 954
1971	49 416	1981	60 944	1991	67 061	2001	78 105
1972	53 094	1982	59 543	1992	67 273	2002	78 439
1973	57 237	1983	58 779	1993	67 372	2003	79 892
1974	56 677	1984	59 882	1994	68 679	2004	82 631
1975	56 198	1985	60 087	1995	69 995		
1976	59 673	1986	61 825	1996	71 522		
1977	61 826	1987	63 104	1997	73 292		
1978	64 158	1988	64 963	1998	73 932		
1979	65 220	1989	66 902	1999	75 826		

将表 3-3 的数据代入式（3.4），运用 MATLAB 的拟合工具箱进行拟合，得到拟合表达式为

$$D(t)=31950+1557\cos(0.605t-3159.659)+1.183\times10^{-16}\exp(0.02369t).$$

石油的需求量预测曲线如图 3-6 所示.

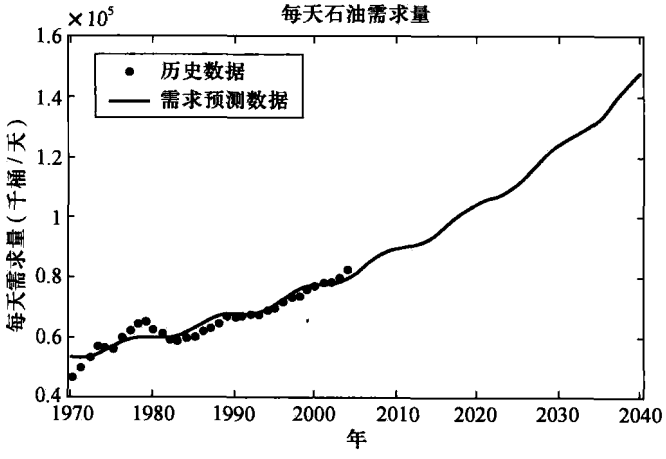


图 3-6 石油日需求量预测值与实际值比较

由图3-6中可以看出，石油的需求量围绕一条指数上升的曲线波动，比较合理地解释了实际中石油需求量上下波动、总体上升的规律.

从图中可以看出，随着时间增长，需求量逐步增长，价格导致的需求波动相对于社会发展导致的需求增长越来越弱. 为了简化模型，略去式（3.4）的价格波动影响项，只保留常数项和指数增长项，即

$$D(t)=a_1+a_5\exp(a_6t),$$

为了比较不同拟合方式下结果的差异，分别给出指数函数拟合、直线拟合和常量需求预测下石油资源的耗尽时间预测函数：

指数拟合函数：$D(t)=29\ 820+2.265\times10^{-15}\exp(0.022\ 23t)$，

直线拟合函数：$D(t)=771.2t-1.467\times10^{6}$，

常量需求函数：$D(t)=D(2\ 004)=82\ 631$.

如果采用指数模型或线性模型，全球的石油资源将在2032年到2033年之间耗尽，既使未来的石油需求稳定在2004年的水平，全球的石油资源也将在2037年耗尽.

3. 模型的分析与评价

1979年未探明的石油储量$U(1979)$是一个难以得到的数据，但是，即使将该数据进行上下10%的改变，对于石油耗尽时间预测的误差也不超过一年.

该模型根据供求关系建立了供求微分方程模型，得到了需求的周期性规律，并在此基础上叠加由于经济发展而导致的石油需求量的指数增长的总体趋势，在历史数据的基础上对模型参数进行估计，最终得到未来50年每年的石油消耗量，进而预测石油的耗尽时间.

该模型比较好地解释了石油需求量周期性波动并持续上升的特点，短期预测可以得到较为精确的结果. 但模型过于简单，没有考虑随着技术的发展，各种替代能源可能逐步取代石油，石油的年消耗量可能出现下降的趋势，因此该模型不适宜用于长期预测，该模型预测石油耗尽时间也过于消极.

此外该模型直接在供求模型的基础上叠加指数模型缺乏依据，如果能够通过修改供求微分方程模型，引入经济发展对石油资源需求量的增长导致需求量的增长，将会得到更好的模型.

3.4.2　模型1.2：考虑多因素的石油耗尽时间预测模型

假设年度石油需求量反映了年度消耗量；不考虑各种因素之间相互影响；忽略影响石油消耗的其他不显著因素.

1. 经济因素影响模型

一般认为石油的消耗量与GDP是密切相关的，表3-4给出了1995—2003年的石油消耗量与GDP的相关数据.

表 3-4 1995—2003 年世界 GDP（亿美元）与世界石油消耗量（千桶/天）

年度	1995	1996	1997	1998	1999	2000	2001	2002	2003
GDP	27.134	28.247	29.433	30.257	31.377	32.850	33.640	34.648	36.000
石油消耗	69 955	71 522	73 292	73 932	75 826	76 954	78 105	78 439	79 813

经统计计算，石油日消耗量 OCD 与 GDP 的线性相关系数达到了 0.992 4，因此，可以认为石油消耗量 OCD 与 GDP 近似满足线性关系. 于是经线性回归拟合，得到石油消耗量与 GDP 的函数关系为

$$OCD = 1108GDP + 40390.$$

以 2001 年为起点，这一年的总储量（包括未探明储量和已探明储量）为 $U(2001)+R(2001)=11.178\times10^{11}$ 桶，计算 GDP 年增长率分别为 10%、5%、3% 和 1% 时，石油资源耗尽的时间预测如图 3-7 所示.

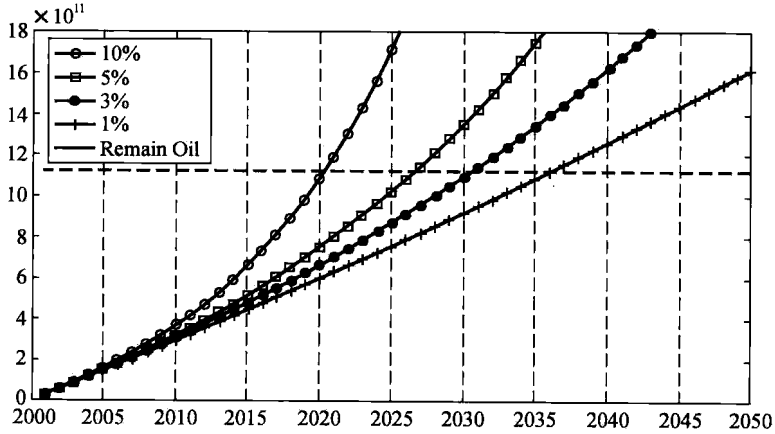


图 3-7 不同 GDP 增长速率下石油资源耗尽时间预测（单位：桶）

可以看出，当 GDP 年增长率为 10% 时，石油资源将在 2020 年耗尽；当 GDP 年增长率为 5% 时，石油资源将在 2026 年耗尽；当 GDP 年增长率为 3% 时，石油资源将在 2030 年耗尽；当 GDP 年增长率为 1% 时，石油资源将在 2036 年耗尽.

2. 人口因素的影响模型

利用 Logistic 模型预测世界人口的发展规律，即人口总量与时间 t 的关系为

$$x(t)=\frac{k}{1+\left(\frac{k}{x_0}-1\right)e^{-(t-t_0)r}},$$

其中起始时间 $t_0=1980$；$x(t)$ 是 t 时刻的人口，$x_0=x(1980)=4.458\,5\times10^9$；$k$ 是地球容许最大人口数，设 $k=10\times10^9$；r 是人口自然增长率，通过历史数据

拟合得到.

将1980年、1990年和2000年的人口数代入模型，计算得到人口的增长率为 $r=0.0327$，以此预测未来的人口数如表3-5所示.

表3-5　1980—2060年世界人口预测（单位：10^9人）

年度	1980	1990	2000	2010	2020	2030	2040	2050	2060
人口	4.4585	5.2736	6.0744	6.8212	7.4849	8.0495	8.5126	8.8811	9.1560

对比分析人口POP增长与石油日消耗量OCD的增长数据，可以发现二者有很大的相关性，其线性相关系数达到0.9877. 利用相关数据作线性因归拟合，得到人口数与石油年消耗量之间的线性关系为

$$OCD=1443\times POP-11170.$$

基于上述Logistic模型，全球石油资源将在2033年耗尽.

3. 政治因素的影响模型

相关的政治因素有很多，在这里讨论最主要的因素——战争因素对石油消耗的影响.

根据过去的数据资料，1970—2004年的石油日消耗量的增长率如图3-8所示.

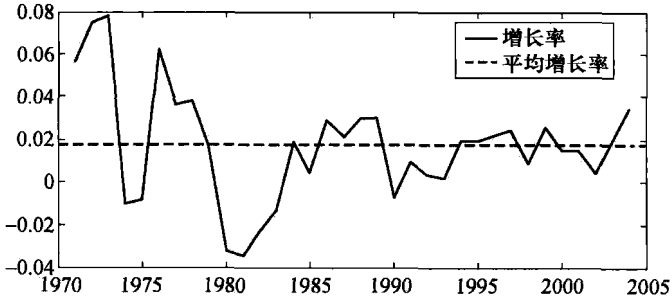


图3-8　1970—2004年世界石油日消耗量增长率拟合图

从世界石油日消耗量增长曲线可以看出，全球的石油日消耗量增长率有4个低谷，分别是1974年、1980年、1990年和2002年. 我们注意到，1973年发生了第四次中东战争，1980年发生了两伊战争，1990年发生了海湾战争，2002年发生了伊拉克战争. 这些战争都发生在世界的产油中心——中东地区，战争导致原油产量下降，价格上升，从而抑制了石油的消费，这也充分反映了战争对石油消耗的影响.

4. 环境因素的影响模型

石油的使用不可避免地会对环境产生一定的污染，特别是 CO_2 的排放. 为此，寻找新能源替代石油产品的一个重要目的就是减少 CO_2 的排放. 表3-6给出了1993—2001年全球范围内使用石油而导致的 CO_2 排放量指标.

表 3-6 1993—2001 使用石油导致的 CO_2 排放量（单位：吨）

年度	1993	1994	1995	1996	1997	1998	1999	2000	2001
CO_2 排放量	9 220	9 284	9 388	9 586	9 691	9 766	9 939	10 138	10 292

经统计发现，石油的日消耗量 OCD 与 CO_2 的排放量呈现较强的线性相关性，相关系数为 0.993 7. 线性回归方程为

$$OCD = 10.09 \times CO_2 - 25320.$$

上式给出了石油消耗与 CO_2 排放之间的关系. 为了保护环境，政府往往会做出碳排放的规划，图 3-9 给出了不同碳排放规划（碳排放年增长率分别为 3% 和 1%）下的石油日消耗量（单位：桶）.

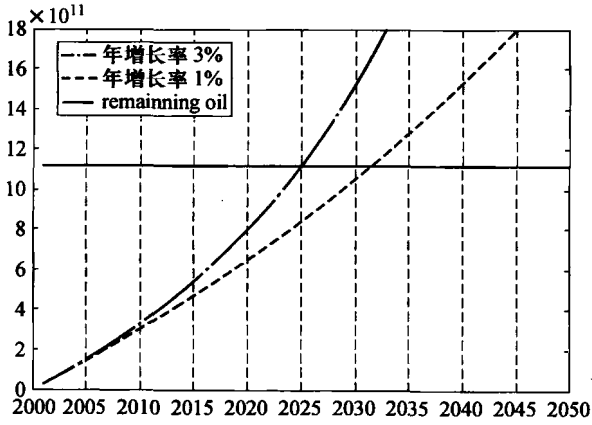


图 3-9 不同碳排放规划下的石油耗尽时间图

事实上，由计算可知，当 CO_2 排放的年增长率为 3% 时，石油资源将在 2025 年耗尽；当 CO_2 排放的年增长率为 1% 时，石油资源将在 2032 年耗尽.

该模型从多个角度研究了不同因素对石油消耗的影响. 模型并不直接通过历史数据预测未来的石油消耗量，而是通过研究历史数据中石油消耗与各个影响因素之间的关系，再预测这些因素的发展规律，最后通过各个因素指标发展的预测值反推石油的消耗. 这样做的好处是可以讨论不同的因素发展状况对石油消耗的影响.

实际中石油的消耗往往同时受到多种因素的影响，而且这些因素之间又有错综复杂的关系，各因素之间的相互关系将在后面进一步研究.

3.4.3 模型 1.3：石油资源的可持续开发模型

为了避免石油资源的无序开采和消耗，需要理性规划石油资源的开发与利用. 这里建立相应的模型进行分析讨论.

设一代人跨越的年限为 n 年. 如果在第 t 年石油资源总量为 $U(t)+R(t)$，

其中 m_1 的资源供这一代人使用，剩下的 m_2 留给下一代人使用，则有

$$U(t)+R(t)=m_1+m_2.$$

定义 η 为 m_1 与 m_2 之比，即

$$\eta=\frac{m_1}{m_2}\times 100\% \quad (0<\eta<+\infty)$$

如果 η 值增大，表示留给下一代人使用的石油资源量减少，反之则会增多.

此外，还需要提高石油的利用率. 随着技术的发展，石油精炼技术不断提高. 为了进一步提高石油的利用率，研究 n 年范围内的石油资源优化配置问题，建立如下线性规划模型：

$$\max \sum_{i=1}^{n} c_i d_i$$

$$\text{s. t.} \begin{cases} \sum\limits_{i=1}^{n} d_i \leqslant \eta' r, \\ c_i d_i \geqslant \alpha c_{i-1} d_{i-1}, \ i=1, 2, \cdots, n, \\ d_i \geqslant 0, \ i=1, 2, \cdots, n. \end{cases}$$

其中，c_i 是第 i 年原油的利用率，体现了第 i 年原油精炼的技术水平；η' 是为这一代人配置的石油资源量百分比的阈值；d_i 是第 i 年的石油消耗量；r 是初始石油储量的总和；α 是一个给定阈值，表示下一年的成品油消耗量不能小于 α 倍的上一年成品油消耗量.

模型的目标函数是实现 n 年内精炼油产量的最大化. 第一项约束条件是这一代人（共 n 年）的石油消耗总量不允许超过配置量.

由于新能源不可能在短时间里替代石油能源，为了保持社会稳定，石油的使用量不可能出现急剧下降，只能是逐步减少. 因此增加第二项约束，第 i 年的成品油消耗量不能低于第 $i-1$ 年消耗量的 α 倍.

假设石油精炼技术是不断发展的，即 c_i 是随时间 i 单调递增的. 设 c_i 满足如下关系：

$$c_i=1-a_1 e^{a_2(i-2004)}$$

根据历年的原油利用率数据做拟合，得到 $a_1=0.4$，$a_2=-0.0347$. 于是当 η'，c_i，r，α 给定后，就能够得到 n 年内最佳的石油配置计划. 为了说明模型的可行性，令 $n=20$，$\eta'=0.51$，$r=1.0\times 10^{12}$，$\alpha=1$. 基于上述数据可得：

（1）这一代人 20 年内允许消耗的石油资源总量为 0.51×10^{12} 桶；

（2）每年的成品油消耗量都在增长；

（3）2005 年的成品油消耗量不少于 2004 年的消耗量.

由于原油的利用率一直在提高，越是后期，原油利用率就越高，因此应该尽量将石油留在后期提炼，从而得到更多的成品油. 由于需要确保成品油供应

量不减少，为了提高原油利用率，可在前 19 年中保持成品油供应量与 2004 年相同，即 $c_i d_i = c_{2003} d_{2003}$，$i = 2005, 2006, \cdots, 2023$. 到 2024 年，剩余原油量为

$$d_{2024} = \eta' r - c_{2004} d_{2004} \sum_{i=2005}^{2023} \frac{1}{c_i},$$

此时将所有剩余的原油精炼为成品油，便可以做到这 20 年石油利用率最高.

实际上，如果 2024 年并不需要这么多的成品油，那么可以考虑将这些石油留给下一代精炼，因为越到后期精炼技术越高，石油的利用率也越高.

如果在上述过程中出现 $d_{2024} < 0$，则说明配置给这一代人的石油资源无法实现每年成品油的供给量不减少，即不能满足实际的需求量，此时需要增加石油配给量，或者降低石油供给量. 图 3-10 给出了最优配置计划的每年石油消耗量（单位：桶/年）.

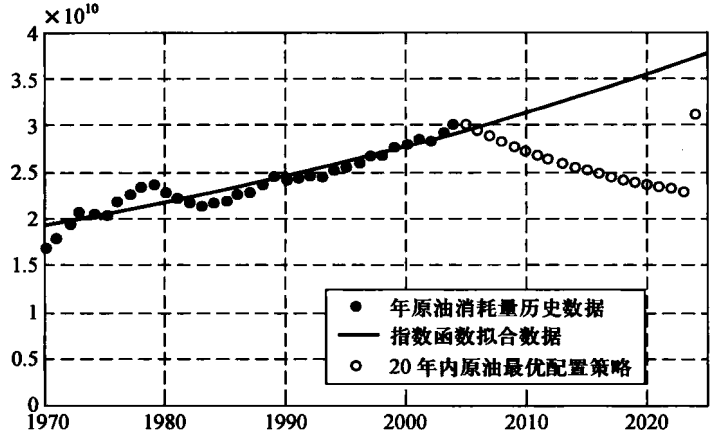


图 3-10 20 年内原油最优配置策略

在图 3-10 中，2005 年以后的数据为优化配置方案值，实线为指数函数拟合的预测需求曲线. 从图中可以看出，最佳的石油配置方案中所给出的石油配置远远小于指数函数拟合预测出的石油份额. 最优配置曲线前段逐年减少，这是因为随着炼油技术的提高，为了达到相同的成品油供应量，所需的原油量在逐步减少. 最后一年突然增高，这是因为将 20 年的配置额度安排在最后一年内使用完，所以导致这一年的原油消耗量增高. 实际中应该根据最后一年的剩余油量调整 20 年间的原油配置量，将多余的原油留给下一代人.

3.4.4 模型 1.4：石油资源的安全开发模型

可以注意到，不同国家的石油利用率不同，如果石油资源在利用率较低的国家消耗太多，而在利用率较高的国家实际需求得不到满足，则将会导致石油资源的极大浪费. 这里依据各个国家石油资源的利用率，建立不同国家间最佳

石油资源的配置模型，实现石油资源利用率的最大化.

假设每年的全球石油消耗量取自模型1.3中所给的结论；不考虑国与国之间的贸易壁垒，国与国之间的石油交易没有障碍. 同时，假设全球有 m 个主要的石油消耗国，第 t 年的最佳石油配置模型为

$$\max \sum_{i=1}^{m} l_i(t)x_i(t)$$

$$\text{s. t.} \begin{cases} \sum_{i=1}^{m} x_i(t) = d(t), \\ x_i(t) \geqslant \alpha_i(t)x_i(t-1), \ i = 1, 2, \cdots, m, \\ x_i(t) \geqslant 0, \ i = 1, 2, \cdots, m. \end{cases}$$

式中，$l_i(t)$ 是国家 i 在第 t 年的石油资源利用率；$x_i(t)$ 是国家 i 在第 t 年的石油消耗量；$d(t)$ 是第 t 年全球的石油消耗总量，该值由模型1.3确定；$\alpha_i(t)$ 是国家 i 相邻两年石油配置量比例的下限阈值，即第 t 年的石油配置量不能少于该国第 $t-1$ 年石油配置量的 $\alpha_i(t)$ 倍.

模型的目标函数是全球第 t 年的精炼油产量最大化. 第一个约束表示该年度各国石油配置之和等于全球配置总量，第二个约束是避免某些国家石油资源配置量骤减.

事实上，各国为了自己的利益，可能设置各种贸易壁垒，从而很难做到最优的石油资源配置.

3.4.5　模型1.5：石油开发对自然环境的危害模型

开发油田需要占用大量土地，破坏邻近区域的植被，改变土壤成分，破坏周围的环境和动物的生存空间. 此外油田开发还将影响地表水位，导致土地荒漠化. 同时，一旦有石油溢出还将严重污染附近的水资源.

为了简化，这里主要讨论油田开发所导致自然危害的短期影响. 假设所有开发生产出来的石油都被消耗；石油的生产能够满足经济发展需要.

设全球油田总数为 k，第 i 个油田在第 t 年的产量为 $x_i(t)$. 为了保证经济的发展，油田的产油总量应与世界石油消耗总量 $d(t)$ （模型1.3确定）平衡. 即

$$\sum_{i=1}^{k} x_i(t) = d(t)$$

开发油田所导致的自然危害与该油田的产量满足一定的函数关系，石油产量越大，所造成的危害就越大，不妨假设两者之间满足线性关系.

此外，规模相同的油田，新油田对环境的危害小于老油田对环境的危害. 可以使用罚函数 $e^{-\alpha[1-\lambda_i(t)]}$ 对不同开采程度的油田进行加权，$\lambda_i(t)$ 表示第 i 个油田从开发以来到第 t 年累计开采量占该油田总储量的比例. 开采程度越高

(即油田开采时间越长)，$\lambda_i(t)$ 的值越趋近于 1，即罚函数越接近于 1；开采程度越低，$\lambda_i(t)$ 越趋近于 0，即罚函数 $e^{-a}<1$，a 由新油田对环境的危害程度评估获得.

综合上述分析，则可给出油田开采对环境危害的模型：

$$p_i(t)=k_i x_i(t)e^{-a[1-\lambda_i(t)]},$$

式中，$p_i(t)$ 表示第 i 个油田开采第 t 年对环境所造成危害的程度；k_i 为环境危害系数，它取决于油田的开采时间和技术，其值越小，表示油田越新，开采技术越先进. 根据上述分析，建立石油开采对环境危害的最小化模型：

$$\min \sum_{i=1}^{k} p_i(t)$$

$$\text{s. t.} \begin{cases} \sum_{i=1}^{k} x_i(t)=d(t), \\ x_i(t)\geqslant 0,\ i=1,\ 2,\ \cdots,\ k. \end{cases}$$

求解该模型，我们发现，具有较小环境危害系数 k_i 的油田将获得更多的生产配额. 此外，增加油田的数量也能够降低对环境的危害，这是因为罚函数抑制了早期油田的继续开发，促使开发更多的新油田以降低对环境危害的总量.

根据上述模型，为了降低油田对自然环境的危害，应该增加采用了新技术的油田产量，降低技术落后油田的产量. 此外还要尽可能增加新的油田，减产甚至关闭老的油田，以降低油田开采对自然环境的危害.

3.4.6 模型 1.6：新能源的替代石油资源的模型

即使通过各种手段控制石油的使用，也只能使原本 30 年即将耗尽石油资源延长 4～5 年，节约使用石油并不能从本质上解决人类的能源危机，我们迫切需要能够替代石油的新能源. 问题是如何发展新能源，保证过渡期经济的稳定增长.

1. 问题的分析与假设

(1) 只考虑一种新能源.

(2) 石油和新能源可以相互替换，即原本使用石油的场所都能够改为新能源.

(3) 以石油和新能源对 GDP 的贡献作为两种能源的度量.

(4) 石油的能量转化率是稳定的.

假设产生单位能量需要消耗 C_1 的石油，或者消耗 C_2 的新能源. 由于新能源技术尚未成熟，在现阶段，往往 $C_1<C_2$，即现阶段使用石油的成本低于使用新能源. 但由于石油资源的总量是不断减少的，油价逐步上升，使用石油的成本也将逐步上升；相反，新技术的发展将使新能源的成本不断降低. 石油与新

能源的价格关系如图 3-11 所示.

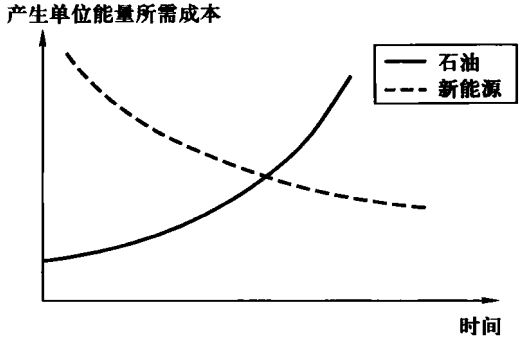


图 3-11　石油价格与新能源价格规律

随着石油价格的上涨和新能源成本的降低，两条曲线的交点即为石油和新能源具有相同的成本. 到那时，除了环保因素，基于经济因素也会有越来越多的人开始选择新能源，直到石油完全被新能源代替.

2. 模型的建立与求解

由各国石油资源开发的最优配置模型可知，在未来的几年里，石油消耗量的增速将会减缓. 不妨设某国 GDP 为 $G=G_1+G_2$，其中 G_1 是受石油消耗影响的 GDP，G_2 是受新能源消耗影响的 GDP，石油消耗量为 x，新能源消耗量为 y，时间为 t.

一般说来，GDP 是石油消耗量和替代能源消耗量的函数，取 $G=G(x,y)$. 此时有

$$\frac{\mathrm{d}G}{\mathrm{d}t}=\frac{\partial G_1}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t}+\frac{\partial G_2}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t},$$

在实际中，如果石油消耗量呈现指数衰减的规律，即

$$x(t)=x(t_0)\mathrm{e}^{-b(t-t_0)},\quad t>t_0,\quad b>0$$

对上式求导，得

$$\frac{\mathrm{d}x}{\mathrm{d}t}=-x(t_0)b\mathrm{e}^{-b(t-t_0)},$$

式中，t_0 为石油需求量开始减少的时刻. 将上式代入 GDP 微分方程，得

$$\frac{\mathrm{d}y}{\mathrm{d}t}=\frac{\dfrac{\mathrm{d}G}{\mathrm{d}t}-x(t_0)b\mathrm{e}^{-b(t-t_0)}\dfrac{\partial G_1}{\partial x}}{\dfrac{\partial G_2}{\partial y}},$$

式中，$\frac{\mathrm{d}y}{\mathrm{d}t}$为新能源替代石油资源的速率；$\frac{\partial G_1}{\partial x}$为石油对 GDP 的贡献率；$\frac{\partial G_2}{\partial y}$为替代能源对 GDP 的贡献率.

基于上述模型，在确保经济稳定增长的前提下，可以得到新能源替代石油

资源的速率. 假设 t_0 为 2010 年，这一年会出现可以推广使用的新能源. 基于上述模型仿真计算，可以得到新能源替代石油资源的过程如图 3-12 所示.

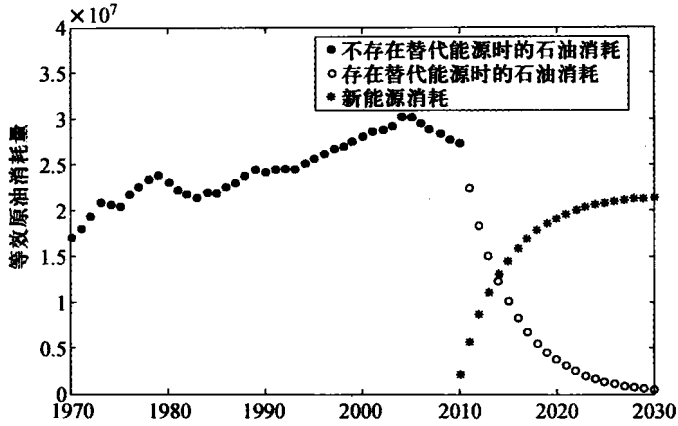


图 3-12 新能源替代石油资源的过程（单位：千桶）

从图 3-12 中可以看出，从 2010 年新能源可以产业化开始，新能源快速占有市场，石油消耗量迅速萎缩，最终新能源消耗稳定在一个常量上，石油的消耗趋于 0. 这里新能源的使用量已等价转化为原油桶数. 从图中可以看出，新能源的使用总量小于石油的减少量，这是因为新能源产业对 GDP 的贡献率大于石油产业，为了保持相同的 GDP 增长速度，新能源的消耗量应少于石油的消耗量.

目前可以看到，潜在的替代新能源包括太阳能、风能、地热能、水力发电以及潮汐发电. 石油的替代品包括天然气、生物燃料（植物或者动物油脂以及酒精）以及可燃气体水化物等.

生物燃料来自农作物，农作物通过吸收大气中的二氧化碳，形成可提供热量的碳水化合物. 燃烧生物燃料产生的二氧化碳将在下一年重新被农作物通过光合作用吸收，从而产生新的生物燃料. 此外，生物燃料不含硫化物、芳香烃以及重金属，其污染程度比化石燃料少得多. 不含硫化物意味着可以减少酸雨，不含致癌的芳香烃类物质（例如苯、甲苯以及二甲苯）和重金属意味着减少对人体健康的伤害.

可燃气体水化物是一种冰状的晶体，它的基本成分是水分子包裹着可燃气体分子. 可燃气体水化物主要存在于两极海底海床（浅水区）以及大陆架斜坡海床（深水区）. 虽然这些替代能源能够减缓石油危机，但是它们能否完全解决能源危机仍然是个未知数.

3.5 石油资源的消耗与控制模型

假设石油资源的生产和消耗总是平衡的，不考虑石油储备、运输和销售所导致的时延. 同时，根据相关的研究结果和统计数据，假设石油的新增探明储量和石油总产量满足 Logistic 曲线.

在这里研究 4 个关键因素：总产量、总探明储量、剩余石油总量和总需求量随时间变化的规律. 通过分析这 4 个因素之间的相互关系，建立微分方程模型. 模型中的相关参数可以通过历史数据拟合确定. 事实上，需求量与生产量密切相关，同时也与探明储量和剩余石油量有关，它们之间满足经济学中的供求关系原理.

通过分析改变相关参数，研究能源的技术革新、经济抑制或刺激、自然或人为灾害以及需求量增长或减少等情况下，石油消耗量的预测问题，为未来的石油生产规划提供建议. 例如，指导政府部门如何控制石油产量，以确保石油资源能够平稳地使用到可能存在合适替代能源的 22 世纪，以及提高油料的供应安全、降低对环境的污染以及开发替代能源等.

3.5.1 模型 2.1：石油资源耗尽时间预测模型

1. 模型的分析与假设

假设石油开采和消耗是同步，也就是说每年开采的石油都在当年被消耗，开采量与消耗量之间不存在时延；需求函数满足经济学中的供需规律，即供应量与价格呈反比关系，需求量和价格呈正比关系；模型的起始年份，设 2000 年为 $t=0$；短期内无法人工合成石油.

图 3-13 给出了美国历年的石油累积开采量，从图中可以看出，初期缓慢增长，中期快速增长，后期又趋于低速增长. 鉴于这一特征，我们采用 Logistic 模型对美国历年石油开采量进行最小二乘拟合，并向外推得到未来各年的石油开采量. 根据 Logistic 模型，容量上限是 18 000 亿桶，这是由当前已探明石油储量（以当前的技术，可盈利开采的石油资源总量）所决定的. 这一数值也是该领域饱受争议的问题之一，18 000 亿桶是这些有争议数据的平均值[2].

由图 3-13 所给出的 Logistic 模型，对其进行微分计算就可以得到每年开采量的拟合模型. Logistic 函数的微分是一个正态分布的密度函数，使用正态分布密度函数对每年的石油开采量进行最小二乘拟合，可得到每年石油开采量曲线，如图 3-14 所示.

图 3-14 中，20 世纪 70 年代出现开采量上升的原因是 OPEC 在此期间调高了油价，而 1979 年下降的原因是伊朗革命导致的原油供应不足. 从图中可

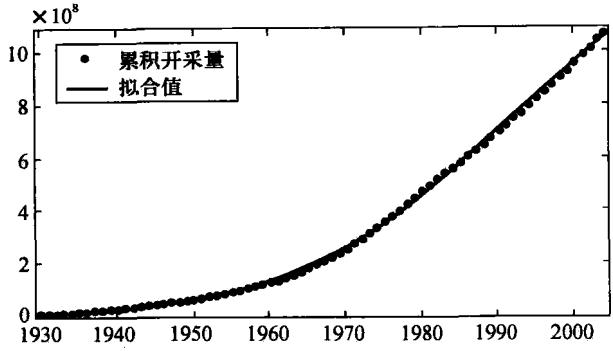


图 3-13 美国历年石油累积开采量（单位：千桶）

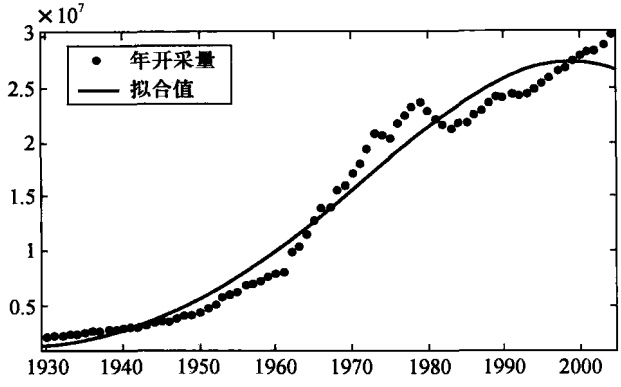


图 3-14 美国石油年开采量（单位：千桶）

以清晰地看出二次世界大战之后的经济快速发展与 20 世纪 80 年代初期经济不景气导致的原油产量起伏.

2. 模型的建立与求解

首先定义如下函数：$S(t)$ 表示 t 时刻的石油累积探明量；$H(t)$ 表示 t 时刻的石油累积开采量；$D(t)$ 表示 t 时刻的石油需求量；$M(t)$ 表示 t 时刻未开采的石油总量.

设全球石油资源的上限为 M_0，随着开采的深入，越到后期，累积探明量的年增长速度（即年探明量）越缓慢，当累积探明量达到上界 M_0（$M_0=1.8\times10^{12}$桶）之后，其年度探明量将变为 0. 此外累积探明量的增长速度还受到需求量的刺激，需求越旺盛，石油勘探会投入更多的勘探技术和设备，使得年探明量增加. 综上所述，建立容量为 M_0，增长因子受需求量作用的 Logistic 模型：

$$\frac{\mathrm{d}S}{\mathrm{d}t}=kDS\left(1-\frac{S}{M_0}\right).$$

式中 k 是一个为了平衡量纲和数值的常数，可以通过历史数据拟合确定.

与前面分析类似，Logistic 模型同样适用于石油累积开采量 $H(t)$ 的估计. 由于石油的开采量始终小于探明量，因此开采量 $H(t)$ 的 Logistic 模型的容量

为探明量 $S(t)$，则有

$$\frac{\mathrm{d}H}{\mathrm{d}t}=bDH\left(1-\frac{H}{S}\right).$$

式中 b 是一个为了平衡量纲和数值的常数，可以通过历史数据拟合确定. 全球的剩余石油储量满足：

$$M(t)=M_0-H(t).$$

对两边求导数，得

$$\frac{\mathrm{d}M}{\mathrm{d}t}=-\frac{\mathrm{d}H}{\mathrm{d}t}=-bDH\left(1-\frac{H}{S}\right).$$

按照供求关系的经济规律，价格随供应量的增加而降低，需求随价格的降低而增加. 为了简化问题，假设供应和需求成正比关系，即 $D(t)\Big/\dfrac{\mathrm{d}H}{\mathrm{d}t}=C$，$C$ 是比例系数. 综合上述4个微分方程，可以得到如下微分方程组：

$$\begin{cases}\dfrac{\mathrm{d}S}{\mathrm{d}t}=kDS\left(1-\dfrac{S}{M_0}\right),\\[2ex]\dfrac{\mathrm{d}H}{\mathrm{d}t}=bDH\left(1-\dfrac{H}{S}\right),\\[2ex]\dfrac{\mathrm{d}M}{\mathrm{d}t}=-bDH\left(1-\dfrac{H}{S}\right),\\[2ex]D(t)\Big/\dfrac{\mathrm{d}H}{\mathrm{d}t}=C.\end{cases}$$

如果历史数据中包含过去几年的石油累积探明量 $s(t)$、累积开采量 $H(t)$、每年的需求量 $D(t)$ 和未开采总量 $M(t)$，其中 $t\in[t_0,\ t_n]$ 表示历史数据的采集时刻，则采用 Euler 迭代方法表示微分项有：

$$S(t_{i+1})-S(t_i)+\varepsilon_i=\frac{\mathrm{d}S(t_i)}{\mathrm{d}t}=kD(t_i)S(t_i)\left(1-\frac{S(t_i)}{M_0}\right)$$

式中，$i=0,\ 1,\ \cdots,\ n-1$；ε_i 表示实际数据中随机性导致的误差. 此时参数 k 的估计方法为求解下面的规划问题：

$$\min\ z=\sum_{i=0}^{n-1}\varepsilon_i^2.$$

类似地，b，C 的求解方法与 k 相同. 得到参数 k，b，C，M_0 后，对上述微分方程组进行数值求解，即可得到在给定初值和参数的情况下4个量随时间变化的曲线，如图3-15所示.

图3-15给出了石油的消耗量、累积探明量、累积开采量和累积需求量的曲线. 这些曲线的导函数具有直观的含义：在时刻 t，H'是单位时间的开采量，D'是单位时间的需求量，S'是单位时间的勘探量. 图3-15中各函数的导函数曲线见图3-16.

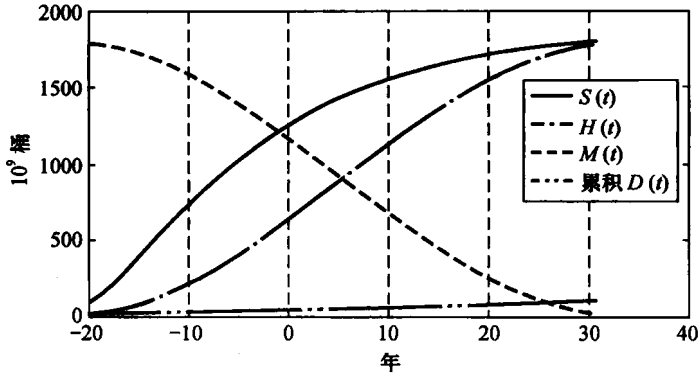


图 3-15 给定初值和参数条件下相关因素的函数曲线

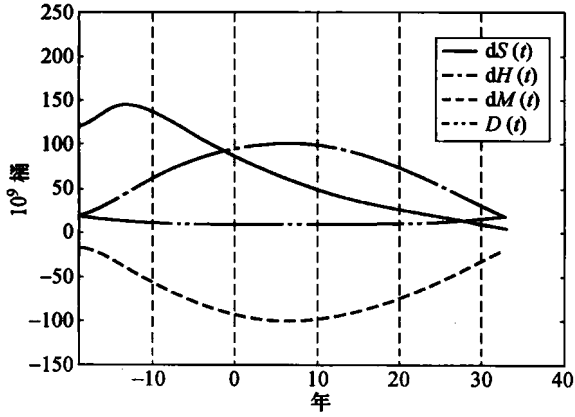


图 3-16 给定初值和参数条件下相关因素导函数曲线

从图 3-16 可以看出，石油的年度开采量 $H'(t)$ 的高峰期落后于石油的年度探明量 $S'(t)$ 的高峰期 25 年左右. 由于剩余石油量始终在减少，因此$M'(t)$是负值.

虽然模型能够给出有关石油资源消耗 4 个主要因素的变化规律，但是模型过于理想化，忽略了一些次要因素的影响.

3.5.2 模型 2.2：多因素影响石油消耗模型

为了更加准确地研究石油消耗的规律，在模型 2.1 中增加以下因素：

(1) 在模型 2.1 中，假设石油的年度开采量 $H'(t)$ 仅受年度需求量 $D(t)$ 和累积探明量 $S(t)$ 的约束，这样在需求量大、且探明量足够多的情况下，可能会导致无限制地开采. 实际中则不容许无限制地开采，政府部门、技术、环境、资金等限制将导致实际的年度开采量被限制在一定范围内. 因此，为 t 时刻的年度石油开采量 $H'(t)$ 需增加一个上限函数 $L(t)$，不妨设 $L(t)$ 为线性形式：$L(t)=mt+r$，其中 m 是 $H'(t)$ 的上限，r 是一个常量. 当年度开采量 $H'(t)$大于 $L(t)$ 时，使用 $L(t)$ 代替 $H'(t)$.

(2) 将 $H'(t)$ 的微分方程中的常量 b 分解为 2 个因子：基础开采因子 b_1 和高新技术开采因子 b_2，其中高新技术开采因子可以在特定的时间加入，用以仿真将来可能出现的技术革新而导致开采能力的迅速提升.

1. 控制石油开采量的响应模型

假设从 2010 年开始，将石油开采量控制在 120 亿桶/年，此时得到如下微分方程组模型：

$$\begin{cases} \dfrac{\mathrm{d}S}{\mathrm{d}t}=kDS\left(1-\dfrac{S}{M_0}\right), \\ \begin{cases} \dfrac{\mathrm{d}H}{\mathrm{d}t}=bDH\left(1-\dfrac{H}{S}\right), & t<2010, \\ \dfrac{\mathrm{d}H}{\mathrm{d}t}=-12\times10^9, & \text{其他}, \end{cases} \\ \dfrac{\mathrm{d}M}{\mathrm{d}t}=-bDH\left(1-\dfrac{H}{S}\right), \\ D(t)\Big/\dfrac{\mathrm{d}H}{\mathrm{d}t}=C. \end{cases}$$

采用 Euler 迭代方法求解上述微分方程组，可得 4 个主要因素的变化规律，如图 3-17 所示.

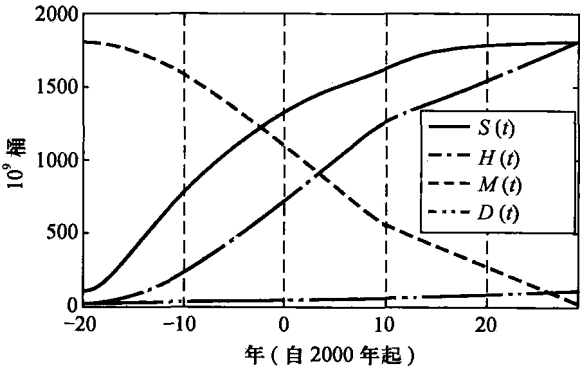


图 3-17　控制开采量时，$S(t)$，$H(t)$，$D(t)$，$M(t)$ 的曲线图

然而，世界范围内的限制产量是不现实的. 不论如何限制，石油资源总会枯竭，限制产量只不过是延缓枯竭的时间而已.

2. 控制石油需求量的响应模型

上述微分方程中的参数 b 包含两个含义，第一是表示石油开采的科技水平，b 的数值越大，表示开采技术水平越高，对于需求的响应就越迅速，从而导致开采速度增大；第二是表示对需求的政策响应，b 的数值越小，表示政策调控力度越大. 通过政策引导，采用各种节能措施，调控石油价格，从而减缓石油供给对价格的敏感性，减缓供给对需求的依赖性，以减少石油的开采量.

假设通过开发替代能源，降低对石油资源的依赖性，政策调控石油价格，降低石油生产对石油需求的敏感性，即将 b 的数值降低到 2010 年的 60%，并持续一段时间. 相应的数学模型如下：

$$\begin{cases} \dfrac{dS}{dt}=kDS\left(1-\dfrac{S}{M_0}\right), \\ \dfrac{dH}{dt}=b(t)DH\left(1-\dfrac{H}{S}\right), \\ \dfrac{dM}{dt}=-b(t)DH\left(1-\dfrac{H}{S}\right), \\ D(t)\Big/\dfrac{dH}{dt}=C, \\ \begin{cases} b(t)=b_0, & t<10, \\ b(t)=0.6b_0, & \text{其他}. \end{cases} \end{cases}$$

对于这个模型，同样采用 Euler 迭代方法来求解，得到 4 个主要因素随时间变化的曲线见图 3-18. 从图 3-18 中可以看出，加入控制开采措施后，石油开采量出现显著的下降. 然而这样做可能影响经济发展，导致经济的缓慢衰退. 相反地，如果提高参数 b 的数值，则可以在短期内促进经济的发展，但加快了石油资源的消耗.

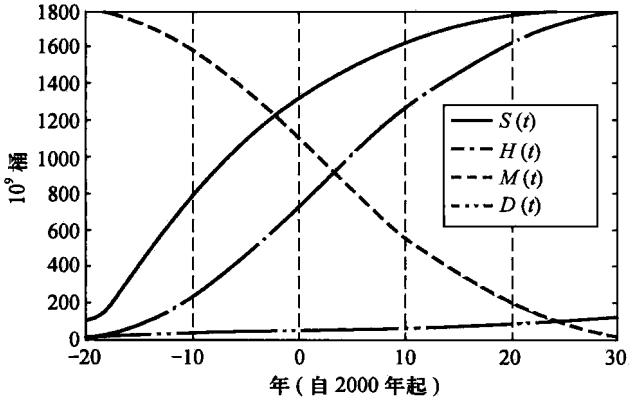


图 3-18 加入控制需求响应策略的 $S(t)$，$H(t)$，$D(t)$，$M(t)$ 函数曲线

3. 自然灾害对石油资源的影响模型

某些自然灾害，如地震、火山、海洋石油泄漏等原因可能导致部分已探明的石油资源无法继续开采.

假设 2010 年的一次自然灾害导致 0.1×10^{12} 桶已探明原油流失，即 $S(t)$ 在 $t=10$ 时，突然减少 0.1×10^{12} 桶. 相应的数学模型如下：

$$\begin{cases}\dfrac{dS}{dt}=kDS\left(1-\dfrac{S}{M_0(t)}\right),\\ \begin{cases}M_0(t)=1.8\times10^{12}, & t<10\\ M_0(t)=1.7\times10^{12}, & t\geqslant10\\ S(10)=S(10)-0.1\times10^{12},\end{cases}\\ \dfrac{dH}{dt}=bDH\left(1-\dfrac{H}{S}\right),\\ D(t)\Big/\dfrac{dH}{dt}=C,\\ \dfrac{dM}{dt}=-bDH\left(1-\dfrac{H}{S}\right).\end{cases}$$

在这里同样采用数值迭代方法求解，得到4个主要因素随时间变化的曲线如图3-19所示.

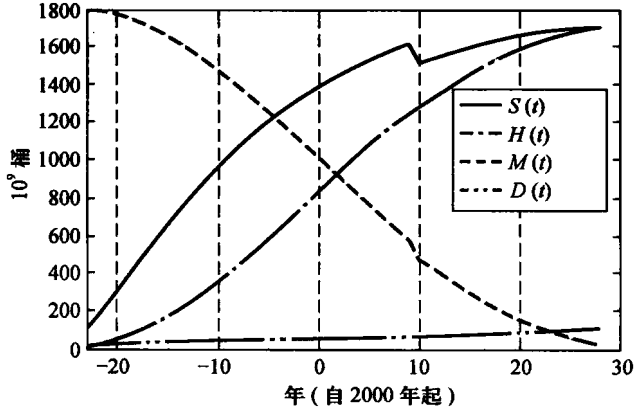


图3-19　2010年发生自然灾害后的 $S(t)$，$H(t)$，$D(t)$，$M(t)$ 函数曲线

从图3-19可以看出，2010年发生了自然灾害，则导致已探明的石油资源损失 0.1×10^{12} 桶，即 $S(t)$ 在 $t=10$ 时突然降低了 0.1×10^{12} 桶. 此外，剩余未开采原油资源 $M(t)$ 同样减少 0.1×10^{12} 桶. 受到容量上限 M_0 的限制，即可导致开采速度的降低，最终累积开采量 $H(t)$ 稳定在 1.7×10^{12} 桶的低值.

此外，关于成品油遭受恐怖袭击、原油泄漏等因素的影响可与前面模型类似地研究. 不同的是这些石油资源的损耗是对已开采石油的损耗，而前面模型是对未开采石油资源的损耗. 因此，在这类模型中需要将已开采量 $H(t)$ 在给定时刻减去一个损耗值，此时剩余石油资源量 $M(t)$ 和 M_0 没有改变.

4. 增加需求对石油资源的影响模型

如果在2010年因为某些新技术的应用，需要大量的石油，不妨设从这一年开始石油需求量增加一倍，即在上述模型中的需求量 $D(t)$ 增加了，则相应的模型修改如下：

$$\begin{cases}\dfrac{dS}{dt}=kDS\left(1-\dfrac{S}{M_0(t)}\right)\\ \dfrac{dH}{dt}=bDH\left(1-\dfrac{H}{S}\right),\\ D(t)\Big/\dfrac{dH}{dt}=C,\ t<10,\\ D(t)\Big/\dfrac{dH}{dt}=2C,\ t\geqslant 10,\\ \dfrac{dM}{dt}=-bDH\left(1-\dfrac{H}{S}\right).\end{cases}$$

类似地，采用数据迭代方法进行求解，得到4个主要因素随时间变化的曲线，如图3-20所示.

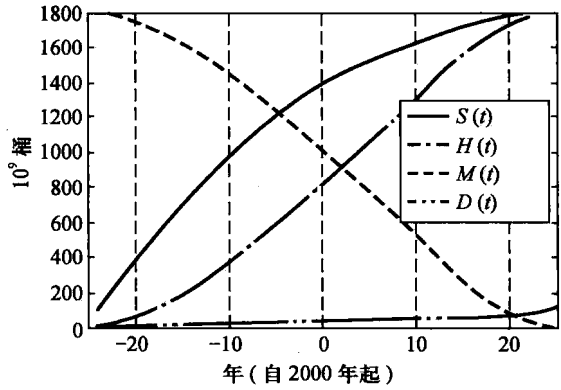


图3-20 石油需求量增加后的 $S(t)$，$H(t)$，$D(t)$，$M(t)$ 函数曲线

从图3-20可以看出，2010年石油需求量的增加，将会导致石油资源开采量的迅速增加，将加速石油资源的耗尽时间.

3.5.3 模型2.3：新能源对石油资源的影响模型

为了社会的进步和经济的发展，必须开发使用新能源来替代石油资源. 替代能源可以是可再生能源，也可以是不可再生能源. 例如可供美国使用几百年的煤炭资源，应用潜力巨大的核能、风能、太阳能、水电等. 最终将采用可再生的能源完全替代石油资源.

在这里，我们提出下一个世纪的石油开采规划. 从2010年起，政府部门带头并鼓励民众使用替代能源，政府用车、出租车、公交车等都要使用替代能源. 同时，预计在本世纪末和下世纪初，基本实现新能源全面替代石油资源，新能源将成为人类的主要能源. 我们提出的规划目标是：到2100年底，全球至少还有10%的石油资源（$M_0=1.8\times10^{12}$桶）得以保留，即累积石油开采量 $H(100)\leqslant 1.62\times10^{12}$桶.

对未来90年（2010—2100年）制定石油限产规划，以实现在2100年底保留1.8×10^{11}桶原油的目标，有两种策略：

- 对每年的产量进行限制，例如降低每年产量的20%，并用新能源填补能源空缺.
- 降低开采比例系数b值，降低开采量对需求量的响应敏感性，逐步降低产量.

在这两种策略下，总的能源空缺是相同的，但每年的空缺配置不一样.

1. 控制石油年产量的影响模型

若从2010年起，将每年的石油开采量控制在原计划开采量的40%，且一直持续到2050年. 这样，在替代过渡期完成之后，新能源逐步被民众接受，相关的产业和市场基本成熟，民众对石油的需求量将逐步减少，此时可以放开控制政策，不再限制石油的开采，并降低石油的需求量$D(t)$. 图3-21给出了控制石油年产量的开采规划曲线.

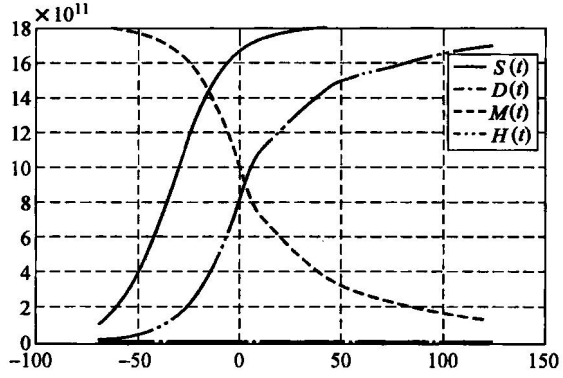


图3-21　控制石油年产量的开采规划曲线

由图3-21可以看出，通过从2010年开始控制石油的年开采量，并在2050年开始降低石油的需求量$D(t)$，最终实现2100年累积开采量为$H(100)=1.62\times10^{12}$桶. 图3-22给出了控制年度开采量后每年的实际开采量.

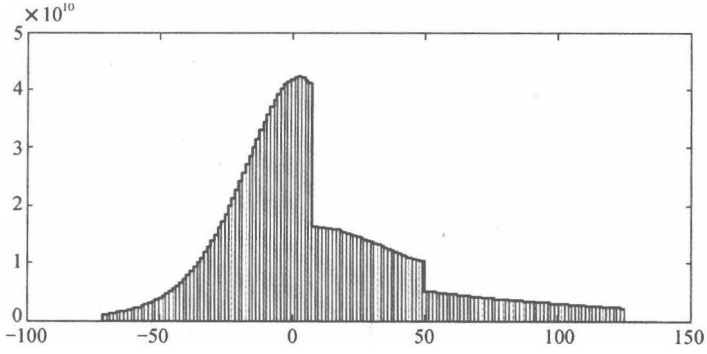


图3-22　控制年度开采量后的实际开采量

从图 3-22 可以看出，在采取控制比例系数的当年，将开采量降低到上一年开采量的 40%，则到 2050 年对石油的需求量就会减少. 于是，进一步降低石油的年开采量到上一年度的 50%，最终将不再开采石油.

2. 降低石油开采比例系数的影响模型

在上述模型中，石油开采比例系数 b 的取值影响很大，直接影响到石油资源的开采量和资源存储量的变化. 根据上面的模型，当将 b 值在 2010 年调整到之前数值的 70.5% 时，则可以实现 2100 年的累积开采量为 $H(100)=1.62\times10^{12}$ 桶.

当新能源替代石油的过渡期完成后，新能源逐步被民众接受，相关产业和市场基本成熟. 从而使得民众对石油的需求量逐步减少，此时可以放开政策，恢复开采比例系数 b 到正常值. 图 3-23 给出了 2050 年放开政策后的石油开采规划曲线图.

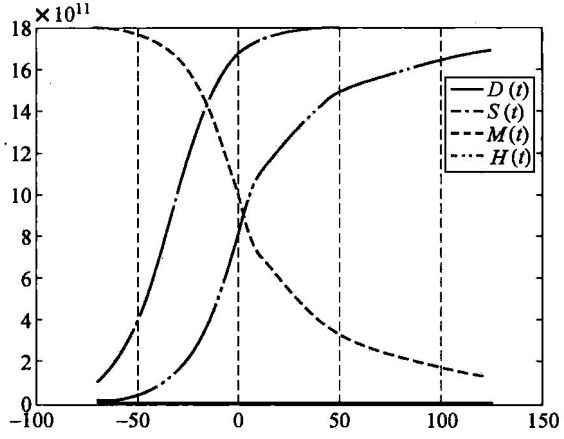


图 3-23 降低开采比例系数 b 的 $S(t)$，$H(t)$，$D(t)$，$M(t)$ 函数曲线

降低年度开采比例系数 b 后的年度实际开采量如图 3-24 所示. 从图中可以看出，控制比例系数 b 与控制年开采量的效果是类似的. 从 2010 年开始，

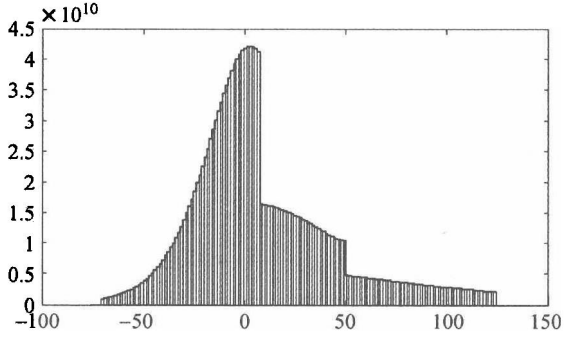


图 3-24 降低开采比例系数 b 的年度实际开采量

年产量降低可以在很大程度上延缓石油的消耗，但也将会或多或少地影响正常的生产和生活. 在这种情况下，需要大量的库存石油以填补正常的需求空缺，这将增加石油的使用成本，并将持续相当长一段时间. 这可能会导致世界经济危机或大量的企业破产，因此对这两种策略在实际中要慎重使用.

3. 石油资源的平稳控制过渡模型

在上面的两个模型中，都是通过控制策略使石油的年开采量 $H'(t)$ 突然降低，将出现供需不平衡的矛盾. 为了实现平稳过渡，应该使开采系数 b 值平稳地减小，而不是阶跃式地减小. 为此，提出如下石油资源的平稳控制过渡模型：

$$b(t)=\begin{cases} b_0, & t\in[0,\ 10], \\ b_0 e^{-k(t-10)^2}, & t\in[10,\ +\infty), \end{cases}$$

其中 k 是降速系数，其值越大，表示 $b(t)$ 的降速越快. $b(t)$ 衰减过程如图3-25 所示.

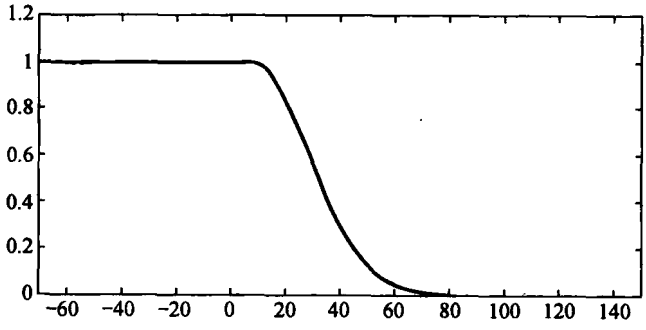


图 3-25　石油开采比例系数 $b(t)$ 的衰减过程示意图

由图 3-25 可以看出，开采比例系数 $b(t)$ 是平稳衰减的. 将 $b(t)$ 代入上面的微分方程模型中，并用 Euler 迭代方法求解计算，可以得到 4 个主要因素随时间变化的曲线，如图 3-26 所示.

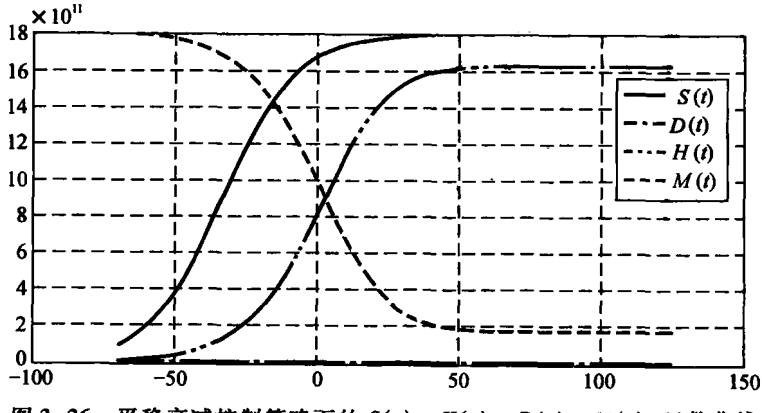


图 3-26　平稳衰减控制策略下的 $S(t)$，$H(t)$，$D(t)$，$M(t)$ 函数曲线

从图 3-26 可以看出，通过对石油开采比例系数进行平稳衰减控制，累积开采量 $H(t)$ 实现了平稳增长，并在 2100 年实现了累积开采量为 $H(100)=1.62\times10^{12}$ 的目标. 在平稳控制过渡策略下，年度石油开采量 $H'(t)$ 的规划图如图 3-27 所示.

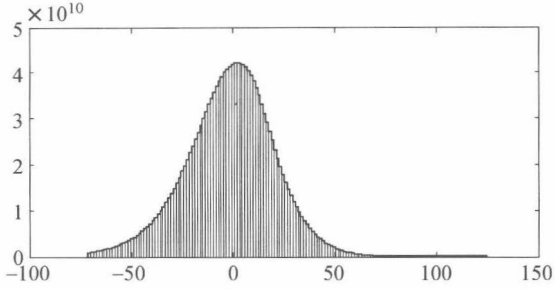


图 3-27 平稳衰减控制策略下石油年开采量 $H'(t)$ 的规划图

从图 3-27 可以看出，通过平稳衰减控制策略，石油的年开采量未出现前面模型中阶跃式的衰减，而是平稳降低，最终实现利用新能源替代石油资源.

3.6 世界水资源的预测与管理模型

该模型围绕世界水资源的消耗问题进行建模分析，提出了一些节约用水的对策. 模型研究了 3 类水资源的消耗，包括农业用水、工业用水和生活用水. 通过建立微分方程模型计算不同类型耗水量的变化速率，并给出研究对象区域内总用水量的变化规律. 此外，当一个国家的用水量接近这个国家的用水警戒水平时，需要综合考虑经济和政策调控措施，建立负反馈控制模型，从而有效控制各类用水量的变化，最终实现水资源的平衡.

基于美国水资源的历史数据，建立数学模型，分析预测未来每年的用水量. 并根据模型仿真检验，分析一个国家面对水资源不足的情况时，寻求水资源消耗的控制措施并有计划地使用不可再生的地下水资源.

3.6.1 模型 3.1：美国年用水量的预测模型

数据表明，地球上超过 2/3 的地表面都覆盖着水，但其中只有 2.5% 是淡水. 幸运的是每年有大量的海水经过自然蒸发，又以降水的形式重新形成地表淡水. 然而降水并没有得到充分的利用，大多数降水又从江河流入大海. 因为降雨量的不均衡性使得局部地区的水资源匮乏，这些地区的年用水量甚至大于自然降水量. 为了填补这些空缺，这些地区只能不断地汲取地下水，这造成了地下水位的下降，可预计有些地区的地下水将在 20～50 年内枯竭.

表 3-7 给出了美国 1950—2000 年的年用水量[8]. 对每年的用水量做简单的线性函数拟合，即得到拟合函数为

$$w(t)=4.929t+60.331,$$

拟合优度指数 $R^2=0.75$，其拟合曲线如图 3-28 所示.

表 3-7　美国的年度用水量

年份	用水量（km^3/年）
1950	247
1960	347
1970	470
1980	538
1990	492
1995	503
2000	512

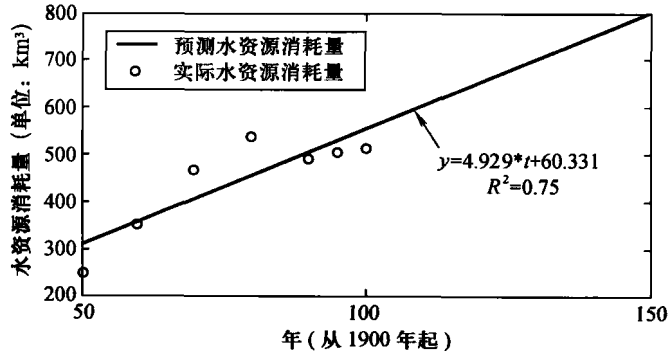


图 3-28　耗水量与时间线性拟合曲线

在线性函数拟合关系下，可以得出美国在 2050 年的耗水量将达到 800.5 km^3. 考虑到用水量不可能无限制地增长，则进一步使用阻滞增长模型做预测分析，即使用 Logistic 阻滞增长模型拟合历史数据，可得到拟合函数表达式为

$$y=375.52\ln(t)-1181.1,$$

此时的拟合优度指数 $R^2=0.82$，其拟合曲线如图 3-29 所示.

采用 Logistic 函数拟合预测，得到美国 2050 年的年耗水量为 700.5 km^3. 虽然上述模型可以看出美国水资源年消耗量的变化趋势，但仍然具有以下不足之处：

（1）该模型没有考虑任何其他因素的影响，例如：人口的增长、技术的发展以及经济和政治因素的影响. 随着水资源价格的上涨以及各种节水政策法规的制定，新增的水资源消耗量将逐步减少，当水资源的年消耗量接近于污水回收再利用量时，上述模型将无法解释这一现象.

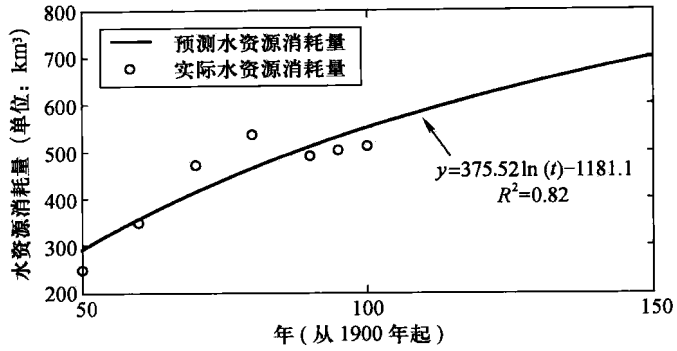


图 3-29 水资源年消耗量的 Logistic 拟合曲线

（2）模型是基于所选用的数据，也就是假设这些数据是真实准确的，能够用以准确地预测未来的发展趋势. 而实际上，即使是美国在数据的统计获取方面也具有很多的不确定性，统计数据的不准确性可能导致模型无法精确地外推以预测未来的变化情况.

3.6.2 模型 3.2：水资源分类消耗的预测模型

模型 3.1 仅仅对历史统计数据进行了简单的拟合，虽然由这种拟合的方法可以大致看出水资源消耗的变化趋势，但是它忽略了很多细节因素. 由于水资源主要有三个用途：农业用水、工业用水和生活用水，每种用途水的消耗都有其规律，并且受当时经济和政策的影响. 因此应该对三种用途的水资源消耗分别建立相应的模型. 下面分别从农业用水、工业用水和生活用水三方面讨论水资源的消耗模型.

1. 农业用水量的变化模型

首先研究农业发展速度的变化情况. 假设农业的发展规模为研究区域内完全依赖于人工灌溉的农业区域的面积. 由于可用耕地是有限的，因此农业的发展规模不会无限制地增长，于是可以认为农业的发展规模是按照 Logistic 模型增长，随着农业发展规模的不断扩大，最终趋于稳定.

表 3-8 给出了美国 1950—1995 年的农业发展规模的统计数据[8]. 根据数据的特征，我们不妨先用指数函数拟合农业发展规模的变化规律，拟合函数为

$$A = 22270 - 12060e^{-0.03897(t-1950)},$$

相应的拟合优度指数为 $R^2 = 0.8557$.

表 3-8 美国 1950—1995 年的土地灌溉面积

时间 t(年)	1950	1960	1970	1980	1990	1995
灌溉面积 A(10^3公顷)	10 400	13 800	15 900	20 600	18 400	20 400

在现有条件下，单位面积的灌溉耕地耗水量为$\frac{C_A}{A}$，单位时间新增耕地面积为$\frac{\mathrm{d}A}{\mathrm{d}t}$，因此单位时间新增灌溉耗水量所满足的方程为

$$\frac{\mathrm{d}C_A}{\mathrm{d}t}=\frac{\mathrm{d}A}{\mathrm{d}t}\frac{C_A}{A}.$$

其中C_A(km^3/年）表示农业灌溉耗水量，A(10^3公顷）表示农业耕地面积.

实际上，影响农业耗水量的因素不仅仅包括耕地面积的增加，相关政策调控与经济手段也可能影响农业灌溉的耗水量.

当水资源的净消耗量（包括农业、工业和生活用水）未达到可循环使用淡水总量时，政策调控或者经济手段对水资源消耗的抑制作用是有限的.

当水资源的净消耗量接近可循环使用的淡水总量时，为了控制水资源消耗，水价就会上涨，政府部门也会制定相应的措施来控制各类用水，以保证水资源的消耗量控制在可容许的水平之内.

当水资源的净消耗量大于可循环使用的淡水总量时，相关政策调控措施与经济手段将会促使各类用水量逐步下降，最终稳定在接近可循环使用淡水总量的水平上.

通过上述分析，可以得到政策调控与经济手段对用水量的控制作用满足如下关系：

$$PE=\left(1-\frac{C}{R}\right)^{P_A},$$

式中，R(km^3/年）是可循环使用淡水总量，C(km^3/年）是水资源的总消耗量（包括工业、农业和生活用水），P_A为政策与经济对农业用水的影响因子，通过调整P_A的值可以进行仿真，容易计算出政策调控与经济手段对水资源消耗的敏感程度.

当$C>R$时，政策调控与经济手段对水资源消耗的影响是负值，表示将限制未来一段时间内水资源的消耗量，使用水总量下降.

当$C<R$时，表明政策调控与经济手段对水资源的消耗没有限制作用，允许用水总量增加.

综合考虑政策调控和农业耕地的增加，农业用水的净增加量模型为

$$\frac{\mathrm{d}C_A}{\mathrm{d}t}=\left(1-\frac{C}{R}\right)^{P_A}\frac{\mathrm{d}A}{\mathrm{d}t}\frac{C_A}{A}.$$

2. 工业用水量的变化模型

这里用人均GDP值来度量工业发展规模，针对美国1975—1995年GDP的数据[9]和人口数据[8]. 从数据分析可知，美国的工业规模增速是稳定的，通过线性回归方法可以得到美国工业规模增长模型为

$$\frac{dI}{dt}=392.99t-12929,$$

此时的拟合优度指数为 $R^2=0.998$.

政策调控与经济手段对工业用水的调控作用与对农业用水是类似的，只是调控系数不同. 设工业用水的调控系数为 P_I，得到工业用水的变化模型为

$$\frac{dC_I}{dt}=\left(1-\frac{C}{R}\right)^{P_I}\frac{dI}{dt}\frac{C_I}{I}.$$

式中，I(美元/人) 表示工业发展规模，C_I(km^3/年) 表示工业用水消耗量.

3. 生活用水量的变化模型

生活用水量与人口总量密切相关. 假设人口增长符合 Logistic 模型，对美国 1970—1995 年的人口数据做 Logistic 线性拟合，得到人口的增长模型为

$$\frac{dM}{dt}=194163t-616608,$$

与农业用水和工业用水情况类似，假设政策调控与经济手段对生活用水的调控系数为 P_M，则生活用水的变化模型为

$$\frac{dC_M}{dt}=\left(1-\frac{C}{R}\right)^{P_M}\frac{dM}{dt}\frac{C_M}{M}.$$

式中，M(10^3人) 表示城镇人口数，C_M(km^3/年) 表示生活用水消耗量.

4. 总用水量的变化模型

总用水量包括农业用水、工业用水和生活用水三个部分，即

$$C=C_A+C_I+C_M,$$

对上式关于时间 t 求导，得

$$\frac{dC}{dt}=\frac{dC_A}{dt}+\frac{dC_I}{dt}+\frac{dC_M}{dt},$$

另一方面，根据农业用水量、工业用水量及生活用水量的变化模型，分别可以得到相应的调控系数 P_A，P_I 和 P_M：

$$P_A=\frac{\ln\left(\dfrac{\dfrac{dC_A}{dt}}{\dfrac{dA}{dt}\dfrac{C_A}{A}}\right)}{\ln\left(1-\dfrac{C}{R}\right)},\quad P_I=\frac{\ln\left(\dfrac{\dfrac{dC_I}{dt}}{\dfrac{dI}{dt}\dfrac{C_I}{I}}\right)}{\ln\left(1-\dfrac{C}{R}\right)},\quad P_M=\frac{\ln\left(\dfrac{\dfrac{dC_M}{dt}}{\dfrac{dM}{dt}\dfrac{C_M}{M}}\right)}{\ln\left(1-\dfrac{C}{R}\right)}.$$

所需要的数据可以通过美国 1990—1995 年的相关数据资料[8]获得. 其中 $\frac{dC_A}{dt}$为这 6 年中每年农业用水的增加量，$\frac{dA}{dt}$的数据在农业规模 Logistic 模型中已经用过，$\frac{C_A}{A}$的数据亦在相关文献[8]中获得. 将这 6 年的数据代入模型中，

取每年相应数据的平均值作为调控参数值，则有 $P_A=1.052$.

同理可得 $P_I=7.36$，$P_M=3.45$.

5. 模型的结果分析

针对上面的模型，利用 MATLAB 中的数值积分函数 ODE45( ) 对模型进行求解，结果如图 3-30 所示.

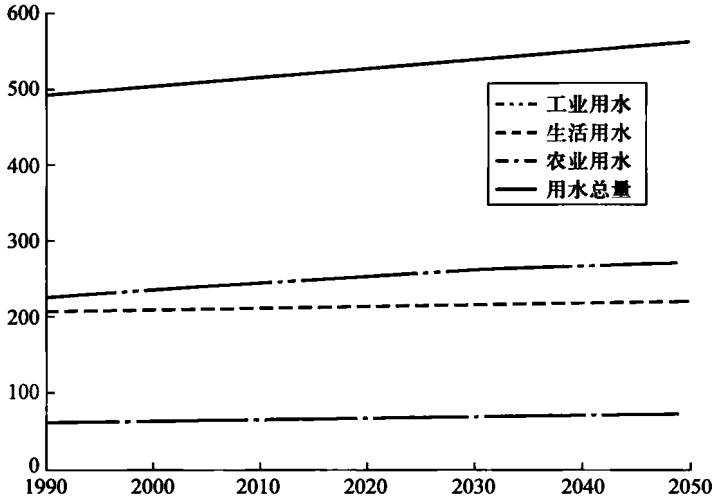


图 3-30　美国 1990—2050 年用水量的预测结果（单位：km^3/年）

从图 3-30 可以看出，美国的农业用水、工业用水和生活用水的总量都是以比较稳定的速度上升，经济和政策调控基本上不能阻止用水量的增加，这是因为美国有大量过剩的可再生淡水资源.

当淡水资源拥有量接近或者处于危险水平时，经济和政策调控作用就会显现出来，这里给出一个例子，模拟一个水资源匮乏国家 1990 年的用水量，其淡水资源使用量为 360 km^3/年，可用淡水资源总量为 430 km^3/年，则相应的农业用水量、工业用水量和生活用水量的变化情况如图 3-31 所示.

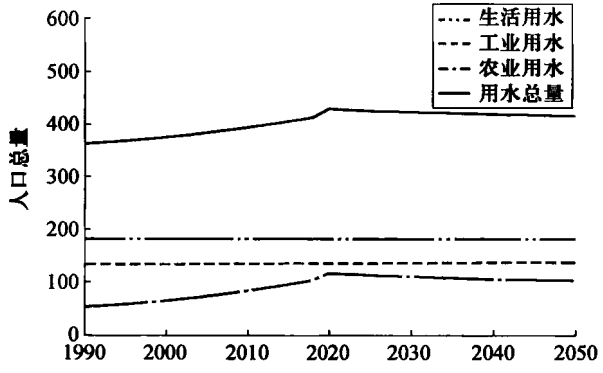


图 3-31　水资源匮乏国家用水量的模拟结果（单位：km^3/年）

从图 3-31 可以看出，当总用水量接近可用淡水资源总量时，即使人口总量还在不断增加，但是在经济和政策调控作用下，总用水量仍然能够被迫减少. 这个国家在农业用水、工业用水和生活用水总量上的变化规律与美国的情况类似，但是该国可再生水资源的总量比美国少得多，因此在后期各类用水总量受到了可用水资源总量的限制.

注意到在模型中，假设在一个特定地区经济与政策调控因子 P 为常量，该数值可以通过历史数据拟合得到. 然而实际中该因子往往是动态变化的，设定 P 为常量会使得模型无法适应时变的经济与政策调控情况. 图 3-32 给出了不同的经济与政策调控因子下水资源的消耗变化情况.

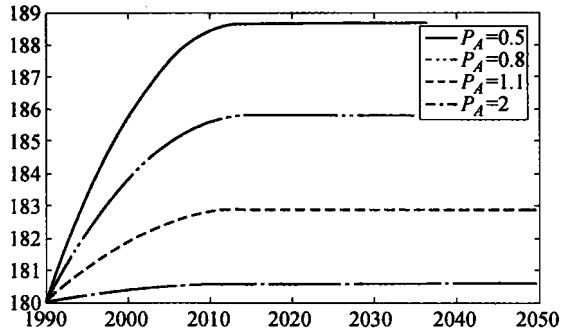


图 3-32 不同经济与政策调控因子 P_A 的耗水量的变化

从图 3-32 可以看出，调控措施越严厉，水资源的消耗量越少. 针对模型中模拟的情况和已有的初值，设定 $P_A = 1.05$. 实际中，不同的调控因子对于农业用水的调控作用并不明显，除非当水资源即将耗尽时，调控的作用才渐渐显露出来. 如果要实现早期控制农业用水量，就需要足够大的调控因子.

同样，有关工业用水调控因子 P_I 和生活用水调控因子 P_M 的分析方法与 P_A 的分析方法是类似的.

3.7 问题的综合分析与进一步研究

3.7.1 问题的综合分析

1. 模型 1 的分析与评价

模型 1 通过一阶线性常微分方程组研究石油的供应量、需求量与价格随时间变化的规律. 由研究结果可知，石油的供应量、需求量与价格随时间呈正弦函数形式波动变化，与实际情况较为相符. 为了体现经济增长导致的石油消耗量的总体增长，模型在正弦波动的基础上叠加了指数增长规律，并用历史数据进行拟合，其结果也与历史数据较为吻合. 通过分析可知，在后期价格波动对

需求量的影响并不显著. 为了简化问题，模型用简单的常量加指数函数的形式来拟合历史需求量，得到了石油资源耗尽的时间. 考虑到石油资源的消耗与GDP、人口和碳排放等因素密切相关，模型基于1995—2003年的历史数据，采用线性拟合的方法确定了不同GDP增长速率、基于Logistic人口模型预测的人口数、在不同碳排放规划下，石油资源消耗量的关系，分别预测得到了石油资源耗尽时间.

该模型的创新点是给出了石油资源在各代人之间的优化配置策略，提出通过平稳减少石油消耗量，为子孙后代预留一定的石油资源. 此外，还给出了依据石油精炼能力确定石油资源在不同国家之间的最优配置方案，这样做的结果是石油资源逐步向石油资源利用率高的国家转移，各国为了获得更多的石油资源，只能提高石油利用率.

2. 模型2的分析与评价

模型2.1参考了Hubbert的思想，将石油累积开采量使用Logistic曲线拟合，将年开采量用正态分布曲线拟合. 通过拟合的方法可以粗略地预测出石油资源耗尽时间. 在此基础上，为了得到更精确的模型，分别建立了石油累积探明储量的Logistic模型、累积开采量的Logistic模型和基于供求关系的年需求量微分方程模型. 通过历史数据的拟合得到模型的相关参数，并利用欧拉迭代法求出微分方程模型的数值解，从而得到未来一段时间内的年累积探明储量、年累积开采量、年累积需求量和剩余未开采量的变化规律.

模型2.2通过增加微分方程模型中变量的约束及改变部分变量的参数值，建立了实际中多因素影响石油消耗的模型，主要包括：控制石油开采量的响应模型、控制石油需求量的响应模型、自然灾害对石油资源的影响模型和增加需求对石油资源的影响模型等，从不同角度分析了相关因素对石油资源的影响和效果.

模型2.3通过调整控制石油产量模型中的主要影响参数，即围绕着如何逐步减少石油开采量的问题分别建立了三个模型：控制石油年产量的影响模型、降低石油开采比例系数的影响模型和石油资源的平稳控制过渡模型. 但模型只是简单地从减少年产量的角度，迫使用替代能源逐步取代石油资源，并没有从两者如何过渡的角度考虑，这一点相比模型1中在保持GDP稳定增长条件下，从替代能源与石油的过渡过程进行描述要略逊一筹.

该模型的突出优点是思路清晰，围绕要解决的问题，通过调整Logistic模型参数和约束条件，实现了各种情况下的石油消耗状况的预测分析. 这种做法与模型1是截然不同的，模型1是以回答问题为主线，针对各个问题分别建立模型，各个问题之间的关联性不强. 而模型2是以模型为主线，通过基础模型衍生出各个子模型用以完成各个小问题的解答，各个问题之间的关联性强，更像是在研究解决实际问题. 当然，如果能够将模型1和模型2的两种做法结合

起来，将会有更好的效果.

3. 模型 3 的分析与评价

模型 3 选用淡水资源作为研究对象，将淡水资源分为农业用水、工业用水和生活用水三大类，并分别针对各类用水建立数学模型进行分析研究.

首先建立了农业用水与灌溉面积的关系，工业用水与 GDP 的关系，生活用水与人口数量的关系. 然后分别通过 Logistic 模型预测未来一段时间里的灌溉面积、GDP 和人口数量. 最后通过预测得到未来的灌溉面积、GDP 和人口数量的变化规律，以及它们与用水量之间的关系，预测未来一段时间各类用水量的变化情况.

在建立农业灌溉面积、GDP 和人口数量与用水量的关系时，该模型巧妙地改造了传统的 Logistic 模型，分别使用调控因子 P_A，P_I，P_M 来表示三类用水的调控力度（当 $P_A=1$，$P_I=1$，$P_M=1$ 时，即为标准的 Logistic 模型），该因子可以反映出政府部门对于水资源危机的重视程度，该数值越大，说明政府部门对水危机越重视. 当淡水资源接近不足时，政府部门将会制定各种强制措施控制用水量. 根据美国 1990—1995 年的实际用水量，通过拟合的方法估计出三类用水途径的调控因子的具体参数值.

该模型可以针对不同国家，对于水资源危机的认识和拥有水资源量的不同，调整不同的参数，以预测各国水资源危机到来时间.

该模型并没有使用通常的“供应-价格-需求”的关系模型，而是通过“经济与政策调控因子”来实现调控力度，并通过该参数实现各类用水的最佳配置.

该模型存在如下不足之处：

（1）当三类用水方式之一发生突然变化，从而导致该地区的用水状况变得不稳定时，该模型无法得到准确的预测结果. 这是因为模型中设定在特定的区域，经济政策调控因子是个常量，该数值无法随着用水状况的改变而改变. 因此无法处理用水状况突变的情况. 如果经济政策调控因子修改成随时间变化的函数，则可以处理这种情况.

（2）由于模型的结果依赖于人口、GDP 以及农业的历史数据，所以对于历史上动荡的国家而言，这些数据往往难以精确地获取，从而无法准确预报未来的用水量.

（3）当模型的研究对象国家面积过大时，可能导致模型预测得不精确. 模型中用美国全国的统计数据代替美国各个地区的数据可能存在一定的误差. 例如，从整体上讲，美国是一个水资源富足的国家，但是美国西南部地区的用水量就大于该地区可再生水资源的储有量. 更小的地区划分，并使用该地区的历史数据进行建模分析可以让模型变得更为精确，从而得到更为准确的预测值.

3.7.2 进一步研究的问题

1. 关于建模方法的选择[10]

在理想情况下，一旦在供求关系模型建立起来后，根据历史勘探、开采和消费数据进行参数估计，就可以预测不同条件下资源消耗的状况了. 有些参赛队还针对能够提供替代资源或者提高现有资源利用率的新技术的作用进行研究. 参赛队可以使用历史数据，通过回归拟合资源的需求量和供应量的变化趋势.

各参赛队通常都注意到资源的消耗量随时间不断增加，但是增速在不断地下降. 推荐的数学模型是针对历史数据采用最小二乘估计得到的指数函数预测模型. 有趣的是，尽管各个参赛队所研究的资源有所差异，但最后的数学模型却很类似.

很多参赛队研究了世界人口发展趋势，并且建立了世界人口与资源消耗之间的关系. 但是往往忽视了不同国家的资源消耗量是不同的，发达国家需要的消耗资源量会更多，而这些国家的人口增长速度却小于发展中国家. 但不论如何，各个参赛队都能够估计出需求大于供应的时间.

很多参赛队都缺乏进一步的思考，没有考虑全球的发展状况的不平衡和资源保护意识的差异. 也就是说，假设未来的发展规律和过去一致，人们不会因为可用资源的减少而减少开支. 只有很少的几个参赛队考虑了全球发展对模型的影响. 例如，有一个参赛队根据过去石油禁运对石油消耗的影响，建立模型估计将来如果发生了石油禁运，对石油消耗量进行了预测分析.

很多参赛队研究了所讨论的资源、资源替代品和一些能够提高资源使用效率的方法. 可能是因为时间有限，并没有建立模型去估计这些替代资源和提高资源利用率的方法能够将“供不应求”的时间点推后多久.

2. 关于竞赛的结果

凝聚着参赛队辛苦劳动成果的论文都给评委留下深刻的印象. 很多论文的图表精美、公式书写规范，然而在模型的建立和模型的求解结果上差别很大. 有少数参赛队，列出的模型公式和论文描述的内容不一致，或许这只是疏忽或者排版错误. 但排名靠前的论文，在文字表述、数学模型的建立、模型的求解或者近似估计、模型的分析和结果的表述等方面都做得很好.

3. 对参赛学生的几点建议

评阅过程再次证明了大学生接受新技术的能力，并且通过独立工作提高了自己迎接挑战的潜力. 我们充分地相信，或许有一些参赛队将继续从事和挑战与资源保护相关的研究课题，或许有一天他们开发出新的技术或者方法推迟所谓的“供不应求”时间点，或者发明一种替代资源的技术，不再使用不可再生或者即将耗尽的资源. 对于将参加 ICM 竞赛的同学，我们给出如下几点

建议：

（1）论文的书写．要检查每个公式的正确性，避免排版错误导致的公式内容和文字描述不一致．

（2）明确地给出模型的假设，并且给出这些假设存在的不足之处；并通过文献说明所选用模型参数的科学性（例如模型中相关参数的取值区间等）．

（3）提供一份参考文献列表，并且在论文中清晰地标注引用．不要列论文中没有引用的文献．

（4）进行模型的灵敏度分析，针对结果的合理性充分讨论．

（5）如实地评价模型的结果，并且讨论结果的适用性．如果模型结果和文献中的结果一致，则要说明这种情况，并且给出文献出处；如果不一致，则要说明不一致的理由，并且给出文献的出处．

（6）仔细检查论文的语法和拼写错误．

参考文献

[1] Non-renewable resources [EB/OL]. [2011-09-02]. http://en. wikipedia. org/wiki/Non-renewable_resources.

[2] Hubbert peak theory [EB/OL]. [2011-09-02]. http://en. wikipedia. org/wiki/Hubbert_peak_theory.

[3] Price elasticity of demand [EB/OL]. [2011-09-02]. http://en. wikipedia. org/wiki/Price_elasticity_of_demand.

[4] Frank A Ward, King J Philip. Economic Incentives for Agriculture Can Promote Water Conservation [R]. Proceedings of the New Mexico State University Water Conservation Conference, 1997.

[5] Knapp K C. Irrigation management and investment under saline, limited drainage conditions. Water Resource Research [J]. 1992, 28 (2): 3099-3109.

[6] Colin J Campbell. The Coming Oil Crisis [M]. Essex: Multi - science Publishing Co. Ltd, 1997.

[7] U. S. Dept. of Energy. Energy Information Administration [EB/OL]. 2004. http://www.eia. gov/emeu/ipsr/t24. xls.

[8] Shiklomanov Igor A. World Water Resources and Their Use [M]. Petersburg: State Hydrological Institute, 1999.

[9] Groningen Growth & Development Centre. Total Economy Database: RealGross Domestic Product. University of Groningen [EB/OL]. [2005-02-07]. http://www. ggdc. net/dseries/gdp. shtml.

[10] Ted Hromadka. Judge's Commentary: The Outstanding Exhaustible Resource Papers [J]. The UMAP Journal, 2006, 26 (2): 175-178.

第4章　灌溉喷洒系统设置问题

灌溉喷洒系统设置问题是2006年美国大学生数学建模竞赛的A题，该问题所研究的喷洒系统，在我们日常生活中是常见的，但也是一个不引人注意的简单而有意义的问题. 事实上系统的设置不仅关系到喷灌效果，而且还会涉及水资源的浪费问题. 这是一个非常有实际意义的问题，值得我们研究和探讨.

4.1　问题的综述

4.1.1　问题的提出

目前有很多种田间灌溉的技术，从先进的滴灌系统到周期性的漫灌等. 较小的农场一般使用手动喷灌，将带有喷嘴的轻质铝管放置在田间，定时人工移动以保证每一块农田都能得到充足的水分. 这种喷灌系统比其他系统经济实用，方便灵活，也容易管理和维护，可用于各种农田和农作物的灌溉. 主要的缺点是每过一段时间，就需要花费时间和精力来移动并重新安装这些设备.

现要使用这种喷灌系统灌溉一块 80 m×30 m 的农田. 已知一套管组由若干互相连接成直线的管子组成. 每根管的内径为 10 cm，并带有一个内径为 0.6 cm的可旋转喷嘴. 管子连接在一起的总长度为 20 m. 水源处的压力为 420 千帕（kilo-Pascal，记为 kPa），水的流速为每分钟 150 L. 要求农田任何部分接受的水量不得超过每小时 0.75 cm，每 4 天至少 2 cm. 并且尽可能均匀地洒水. 需要解决下面的问题：

（1）设计一种放置管道和喷嘴的方案，需要确定喷嘴数量和各喷嘴之间的距离，在满足灌溉要求的前提下，使得农场主使用、维护该系统需要的时间最少.

（2）给出一个移动管组的方案，包括管组移动位置的工作进度表.

问题的原文如下：

Positioning and Moving Sprinkler Systems for Irrigation

There are a wide variety of techniques available for irrigating a field. The technologies range from advanced drip systems to periodic flooding. One of the systems that is used on smaller ranches is the use of "hand move" irrigation systems. Lightweight aluminum pipes with sprinkler heads are put in place across fields, and they are moved by hand at periodic intervals to insure that the whole field receives an adequate amount of water. This type of irrigation system is cheaper and easier to maintain than other systems. It is also flexible, allowing for use on a wide variety of fields and crops. The disadvantage is that it requires a great deal of time and effort to move and set up the equipment at regular intervals.

Given that this type of irrigation system is to be used, how can it be configured to minimize the amount of time required to irrigate a field that is 80 meters by 30 meters? For this task you are asked to find an algorithm to determine how to irrigate the rectangular field that minimizes the amount of time required by a rancher to maintain the irrigation system. One pipe set is used in the field. You should determine the number of sprinklers and the spacing between sprinklers, and you should find a schedule to move the pipes, including where to move them.

A pipe set consists of a number of pipes that can be connected together in a straight line. Each pipe has a 10 cm inner diameter with rotating spray nozzles that have a 0.6 cm inner diameter. When put together the resulting pipe is 20 meters long. At the water source, the pressure is 420 Kilo-Pascal's and has a flow rate of 150 liters per minute. No part of the field should receive more than 0.75 cm per hour of water, and each part of the field should receive at least 2 centimeters of water every 4 days. The total amount of water should be applied as uniformly as possible[1].

4.1.2 问题的背景资料

喷灌是当今世界上最主要的节水灌溉技术之一（如图 4-1）. 除此之外，还有微灌、负压差灌溉、坡地灌水管灌溉、渗灌等.

喷灌即是喷洒灌溉的简称，特点是利用专门的设备（水源工程、动力机、水泵、管道、喷嘴等）把水加压或利用水的自然落差将有压水送到灌溉地段，通过喷洒器（喷嘴）喷射到空中散成细小水滴，均匀散布在田间进行灌溉. 一般由水源、动力、水泵、管道系统、喷嘴、附件等组成. 根据整个系



图 4-1 茶园喷灌实景图

统的移动情况，喷灌系统可分为固定式、移动式和半固定式喷灌. 固定式是除喷嘴可移动外，其他都固定，其优点是操作方便、生产效率高、运行成本低、占地少；缺点是设备利用率低、投资大. 移动式是除水源外其他部分都可移动，其优点是设备利用率高、单位面积投资低、操作灵活；缺点是田间所需渠道多、路线多、占地多、劳动强度大. 半固定式是首部和干管部固定，支管及喷嘴可移动，其优点是设备利用率较高，投资较少，劳动强度一般；缺点是工作条件较差，有时劳动强度较大.

对于较小的农场来说，使用手动的移动式喷灌（属于半固定式）可能造价更低，也利于管理和维护，应用也较为广泛.

4.1.3 问题的分析

问题要求用一种手动的移动式喷灌系统浇灌一块 80 m×30 m 的农田. 现有管组由若干内径为 10 cm，并带有一个内径为 0.6 cm 的可旋转喷嘴的管子互相直线连接而成，总长为 20 m. 在满足喷灌要求的情况下，确定一种最优管组的组成及放置方案，即该管组需要安装多少个喷嘴，各喷嘴之间的距离是多少？该套管组如何安放在田间，每隔多长时间移动一次？这个方案要求喷灌效果要尽可能均匀，并且移动管组的次数要尽可能少.

为此，首先需要根据水压及喷嘴内径计算出每个喷嘴的射程、喷洒面积和喷水量. 通常喷嘴按工作压力和射程可分为高压远射程、中压中射程和低压近射程，其中中压中射程的工作压力为 200 ~ 500 kPa，射程为 15 ~ 40 m，流量为 0.8 ~ 31 m^3/h. 一般情况下，可旋转喷嘴的喷洒面积可以近似地看做一个以射程为半径的圆域. 由几何分析，要想浇灌大小为 80 m×30 m 的矩形田地，首先将问题转化为以最少数量的半径为 R 的圆域来覆盖整块田地的问题. 然后根据所得圆域的数量（即喷嘴的数量）来进一步规划组管的配置和移动时间表.

在规划组管时，喷嘴的组合形式除了受限于所要求的条件，主要取决于地块形状和风的影响. 管组的组合方式一般为矩形或三角形. 矩形适用于规则地块，边缘成直角的情况，这种形式设计简便，容易做到使各支管的流量比较均衡；三角形适用于不规则地块，或地块边界为开放的，喷洒范围允许超出部分边界. 这种布置抗风能力较强，喷洒均匀度要高于矩形或正方形的情况，同时所用喷嘴的数量相对较少，但不易做到各条支管的流量均衡. 有时地块形状十分复杂，或地块当中有障碍物，那么管组的组合形式就应为不规则形状. 对于题目所要浇灌的田地，选择正方形或正三角形布局方案较为合适.

土壤的允许喷灌强度是影响喷嘴选型的主要因素之一（这里喷灌强度是指单位时间内喷洒在田地的水深），一般考虑组合喷灌强度. 要求喷灌强度（记为 ρ）每小时不超过 0.75 cm，每 4 天至少要接受 2 cm，如果农场工人每

天最多工作 12 h，则每小时的洒水量要大于$\frac{2}{4\times12}=0.042$（cm）．如果把一天作为一个周期，则每天的喷灌强度为 0.042 cm/h≤ρ≤0.75 cm/h.

在计算喷嘴的水压时，还要考虑水在支管输送过程中的水压损失，然后计算喷嘴的射程以及喷灌强度来确定喷嘴的数量和间距，最后根据农田的面积与均匀度来确定管组的工作计划表，即移动时间和每次移动的距离.

4.2 模型1：基于多步迭代算法的农田灌溉优化模型

灌溉是农场工人经常要做的事情，因此希望灌溉方案规律性强、易于操作，移动管组的次数要尽量少．为此在设计管道放置位置和移动方案时，可以采用逐步优化的设计方法来实现.

4.2.1 问题的分析与假设

1. 问题的分析

根据问题的特点和要求，拟采用多步迭代的优化算法来寻求问题的最佳解决方案.

首先，运用几何分析方法将其转化为一个全覆盖问题，即如何寻求用最少数量的等面积圆域来覆盖这个给定的区域．这种方法可能会出现喷洒器的安置超出田地边界的情况，所以在满足所有田地都能够被喷洒到的前提下，适当调整喷洒器的位置．可以通过模拟方法绘制出最少移动距离的管道铺设图．采用迭代优化算法，在每一步迭代中都动态地寻求最少量的管道铺设方案，如此进行下去，经过有限次的优化迭代之后，就能够得到整个区域的最佳管道铺设方案．这个方案的基本原则是在保证整块田地能全部被浇灌到的前提下，来寻求管道铺设的最佳方案，并通过一种算法来完成喷灌日程表的编制工作.

2. 模型的假设

为了简化问题以及建立模型的需要，给出如下假设：

（1）田地足够平坦，没有坡度，各处水压相等；

（2）一般认为风向会影响喷灌的均匀度，在此暂不考虑风向的影响；

（3）喷洒器托架高度 76.2 cm（这是最普遍的高度配置）；

（4）为了灵活方便地安排喷灌日程表，不妨假设农场工可以一天连续工作 12 小时，事实上一天连续工作 12 小时是不太容易实现的；

（5）田地的边界由连着水源的管子组成，喷洒器的射程可以超出田地边界；

（6）移动安装管道和喷嘴一次需要 1 小时.

4.2.2　模型的建立

由问题的分析，要建立模型就需要根据问题所给的条件求出喷灌系统中相应的参数值，如喷洒半径、喷洒的区域等，然后就一个喷嘴的情形进行分析检验，最后建立整块田地的喷灌计划.

1. 喷灌系统参数的确定

在实际中，常用的喷灌系统的水管是铝质的，内径为10 cm，可旋转喷嘴的内径为0.6 cm，水源处的水压为420 kPa，水流速为150 L/min. 利用压力损失公式（Hanzen Williams公式）[2]：

$$\text{压力损失} = 4.55\frac{(Q/C)^{1.852}}{(ID)^{4.87}}L. \tag{4.1}$$

其中，压力损失的单位为psi，表示磅/平方英寸；Q为管道的水流速，单位为gal/min（加仑/分钟）；C是摩擦系数（铝管一般取120）；ID是管子的内径，单位为in（英寸）；L为管子的长度，单位为ft（英尺）.

喷嘴的水流速为

$$v = 29.82\sqrt{P}D^2C_d. \tag{4.2}$$

其中，水流速的单位为gal/min；P是喷嘴的水压力（单位：psi）；D是喷嘴口的直径（单位：in）；C_d是喷嘴水流量系数（一般介于0.96～0.98之间）.

已知水源的水压是420 kPa，由单位换算公式1 kPa=0.145 psi，则水源的水压为60.9 psi. 如果不考虑压力损失，则喷嘴的水压P至多是60.9 psi. 又1 in=2.54 cm，即喷嘴口的直径$D=0.6/2.54=0.236$（in）. 如果取喷嘴水流量系数$C_d=0.97$，将其代入（4.2）式，则有每个喷嘴口的水流速为12.57 gal/min，即47.58 L/min.

事实上，压力损失与管道的长度成正比，若管道总长$L=20$ m$=65.6$ ft，管道的水流速$Q=150$ L/min$=39.6$ gal/min，管子内径$ID=10$ cm$=3.94$ in. 将以上数据代入（4.1）式中，就可以得到压力损失为0.0464 psi，即0.32 kPa. 因此水在支管中流动的压力损失可以忽略不计. 假定每个喷嘴都放在高度为30 in=2.5 ft的支架上，支架为内径为1 in的钢管. 由于支架的缘故，流速为12.57 gal/min水流通过钢管支架造成的压力损失为0.058 psi，那么由此产生的水压损失也可以不予考虑.

虽然该问题给出的喷嘴配置与市场上的不太一致，但一般还是很接近的，如Nelson Irrigation公司生产的WF16（0.3 cm），是一种可转动的喷嘴（图4-2），图4-3所示为灌溉技术中心[3]提供的这种型号喷嘴的性能曲线，即喷嘴喷洒的洒水量与喷射距离的关系.

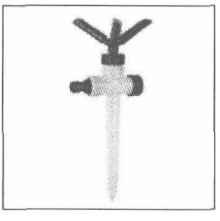


图4-2　可旋转喷嘴

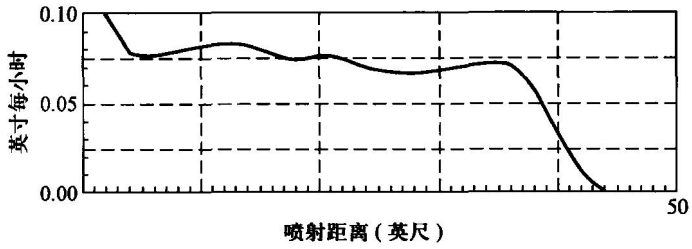


图 4-3 喷嘴的洒水量与距离的关系图

图 4-3 中横轴表示喷射的距离（英尺），纵轴表示喷灌强度（英寸/小时），喷嘴的水源水压是 60 psi，与我们问题中的 60.9 psi 非常接近. 从图 4-3 中可以看出，一个喷嘴的喷洒半径大约为 44 英尺，近似为 13.4 m. 当喷射距离 r 介于 0.1 m 与 11 m 之间时，洒水量基本均匀，保持在每小时 1.9 mm 左右. 当 r 为 0.1 m 时洒水量最大，大约为每小时 2.5 mm，而当 r 大于 11 m 以后，洒水量随着 r 的增加迅速减小为零.

假设 0.6 cm 的喷嘴与 0.3 cm（WF16）的喷嘴使用方法相同，且具有更高的水资源利用率. WF16 的流速为 3.42 gal/min，所要考虑的喷嘴水流速为 12.57 gal/min，在考虑有一定压力损失的情况下，我们所要考虑的喷嘴喷灌利用率大约是 WF16 的 3.3 倍. 此时在一个喷嘴的喷洒范围内，最大洒水量不超过每小时 0.75 cm. 也就是如果支管上只安装一个喷嘴时，可以连续喷洒，直到达到洒水量 2 cm 的要求.

2. 模拟一个喷嘴的喷灌效果

将喷嘴放置在任一位置上，给定一个时间长度，利用图 4-3 中所提供的数据，借助于 MATLAB 程序来模拟喷洒器喷灌情况. 图 4-4 显示了喷洒器运行一个小时的工作状况[1]. 因为喷洒器配置数据都是离散分布的，因此在采样点间利用样条插值来实现.

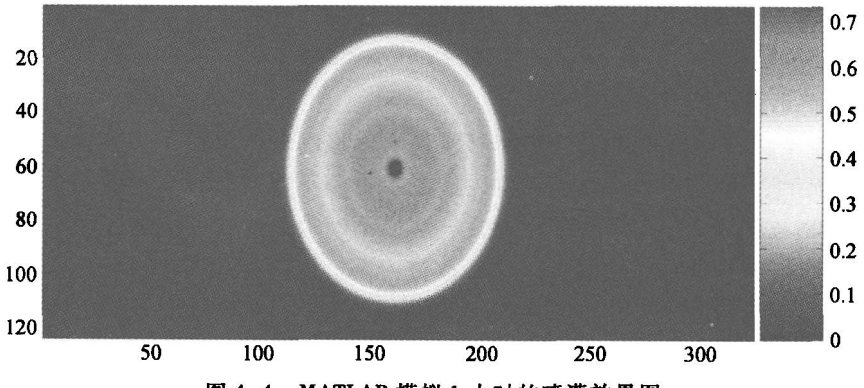


图 4-4 MATLAB 模拟 1 小时的喷灌效果图

对于整块田地而言，将其划分为单元格，每个单元格为 0.25 m×0.25 m. 在模拟中，记录下每个喷嘴的位置及其喷洒时间，并将结果填入表格中. 重复表格中喷洒器的位置，并计算出每个位置每次模拟降水的情况.

按照喷洒器的喷射范围大小划分田地网格，然后用简单的环状圆域遍历每一个网格模拟洒水情况. 计算出每一个单元格中喷洒器的位置，然后用给定的喷洒器配置和喷灌时间来计算每一个单元格接收到的洒水量. 这个数据就作为该单元格的统计洒水量.

3. 全覆盖喷洒的实现方法

按照问题的要求，每块田地每四天的喷灌强度不少于 2 cm. 为达到这个目标，将每一个喷嘴的有效喷洒面积看作一个圆域，问题就变为寻找用最少的这种圆域全覆盖整个矩形田地. 但是，由于喷嘴在每个圆域中的喷灌强度是近似均匀分布的（如图 4-4），因此径向重叠覆盖就会影响整体的喷灌均匀度，而且会增加喷嘴的数量. 因此，在保证每块田地都能被喷洒到的情况下，通过减少喷嘴的数量来减少重复喷灌的面积. 那么，这就归结为一个在一定条件下的全覆盖问题了.

4.2.3　模型的求解方法

利用单个喷嘴的喷洒半径，寻求到用最少的等面积圆域来覆盖给定的矩形田地. 如图 4-5 所示，给出了一种解决方案. 事实上可以证明，这种设置是一种有效的解决方案[4].

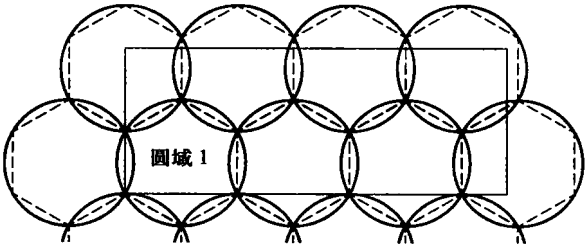


图 4-5　六角形的喷洒覆盖方案

1. 喷洒覆盖的改进方案

根据上面的配置方案，虽然能够实现全区域的喷洒灌溉，但这种方案会导致一些喷洒器的位置超出田地边界，造成水资源不能有效地利用并且影响喷灌的均匀性. 为此，需要对其进行调整. 这样不断地调整，直到得到最佳的配置方案. 图 4-6 给出了一种可行的调整改进方案[1].

由此可知，这种改进的覆盖方案使所有的喷洒器都在边界之内，并且保证覆盖尽可能均匀. 从六角形喷洒覆盖方案（图 4-5）可以看到，圆 1 位于距离田地左边界为 $X=R\sqrt{3}/2$ 的右方. 由图 4-3 喷洒水量与喷洒距离的关系，单个

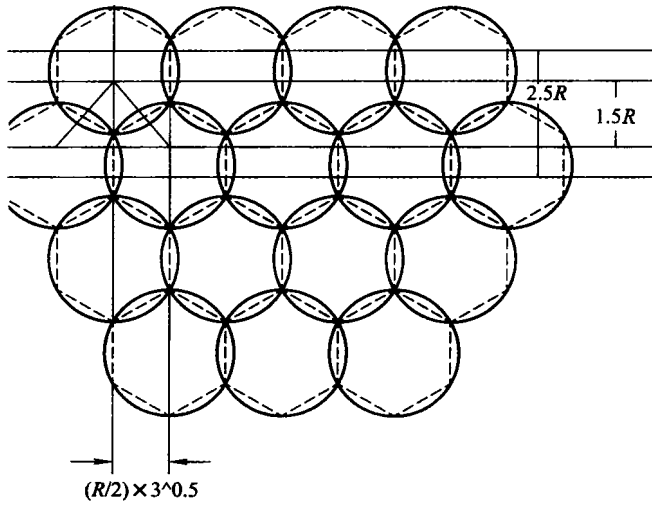


图 4-6 调整后的六角形喷洒覆盖方案

喷嘴的喷洒半径 R 约为 13.4 m，但在 13.4 m 处洒水量为零. 因为整块田地长为 80 m，$80/\sqrt{3}R \approx 3.5$，所以横向需要 4 个圆域覆盖（即需要 4 个喷嘴），整块田地最少需要 8 个圆域才能将其覆盖.

又由于两个圆心（喷嘴）之间的距离为 $R\sqrt{3} \approx 23$ m，超过了管子的总长 20 m，因此每个管组只能安装一个喷嘴. 这样工人在四天内最少要移动安装 8 次管组. 由假设知，工人每天最多工作 12 小时，每次移动管道和安装喷嘴需要用 1 小时，所以就要求在 5 小时内洒水量最少达到 2 cm，即每小时洒水量最少要达到 0.4 cm. 加上这个约束条件后，有效半径变为 11.5 m，超过的部分将得不到足够的洒水量. 因此，为了实现最大程度的覆盖，在计算洒水量时取 $R=11.5$ m，而仍以 $R=13.4$ m 的固定间距放置喷嘴. 这会导致整个系统组合向左移动 $1.9\sqrt{3}/2$（其中 $13.4-11.5=1.9$ m）. 这样，不仅保持喷洒器在边界以内，而且还可以增加喷洒的均匀度.

根据上述计算与分析，从理论上得到管道喷嘴的安放位置方案，示意图如图 4-7 所示.

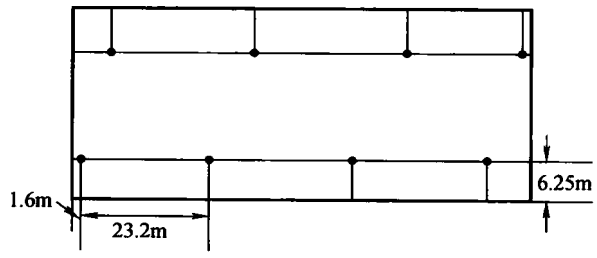


图 4-7 管道喷嘴的安放位置方案示意图

2. 喷洒均匀度的最大化方法

在实际中，尽可能把水均匀地喷洒到田里是最主要的问题，因为这与作物的产量直接相关. 在这里运用同步扰动随机逼近（Simultaneous Perturbation Stochastic Approximation，SPSA）的优化算法[5]来求解计算这个问题，这个算法主要是降低整体平均洒水量的标准偏差.

SPSA 算法经过有限步迭代后的结果如图 4-8 所示[1]，给出了喷嘴的位置和喷洒均匀度的效果.

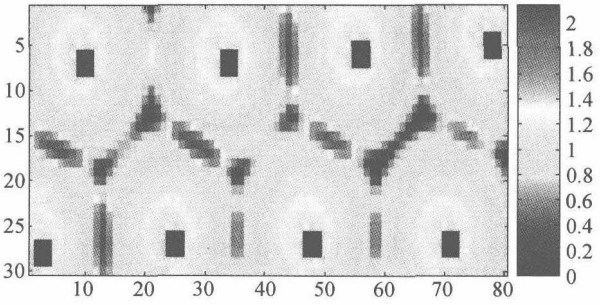


图 4-8　用 SPSA 算法迭代 5 000 次后喷嘴位置与效果图

在迭代 5 000 次之后，已经接近了覆盖问题的最优解. 也就是说，用 SPSA 算法经过多次迭代后使喷洒器移动到了最佳均匀度的位置上，得到最佳的喷洒均匀度效果.

3. 喷洒器移动优化的方法

在使用 SPSA 算法得到喷洒器设置结果的基础上，寻找使得喷洒器移动距离最小的方案. 这里采用如下的优化算法实现，算法的步骤如下：

（1）以 SPSA 迭代算法的求解结果作为喷洒器的初始坐标.

（2）任意选择一个喷洒器，不妨就选择距离田地左上角最近的喷洒器，记为喷洒器 A.

（3）测量其他喷洒器与 A 之间的距离，找到距离 A 最近的一个喷洒器，记为喷洒器 B. 由于水源主管道只沿田地横向铺设，因此只需找出横向距离最近的一个即可.

（4）检查两个喷洒器之间的距离是否小于或等于管道长度（20 m）. 如果小于或等于 20 m，则计算这两个喷洒器喷洒重叠部分的总洒水量. 如果重叠部分的洒水量不超过 0.75 cm/h，那么就将该距离设置为两个喷洒器之间的距离，转（5）. 否则，用喷洒器 B 代替喷洒器 A，转向（3）.

（5）计算结束，输出结果.

4.2.4　模型的结果分析

按照上述的方法和算法，经过计算可以得到喷洒器的设置方案，即管道喷

嘴的设置方案，如图4-9所示.

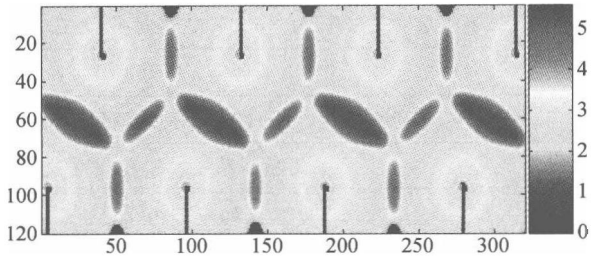


图4-9 最佳管道和喷嘴的设置方案

图4-9中的圆点表示喷嘴的位置，由支管连接到横向水平铺设在田边的主管道上. 横轴和纵轴所标数据是按0.25 m为单位标记的田地边长. 显然，这个配置方案包括8个喷洒器移动点，每隔5个小时移动一次，其中包括1小时移动和安装设备的时间. 这样每4天可累计喷灌48小时，大约每天12小时.

由图4-9可以看出，每个支管上只安装一个喷嘴，其原因是当喷嘴之间的距离小于21 m时，在喷嘴射程内的重叠部分的洒水量就会超过0.75 cm/h，不符合问题的约束条件. 同时，注意到大约有不到1%的面积洒水量不足2 cm，但这无妨大碍，因为这些大多都是在边界上的. 实际中，对于这些位于边界上的作物来说，它的生长除了受到水分的影响以外，还会受到其他诸多因素的影响，诸如农场机械设备照顾不到、劣质的土壤因素等等.

按照以上的喷洒器设置和移动方案，并根据问题的要求可以给出农场工人移动喷灌的日程表，如表4-1所示. 其中要求农场工人每天工作12 h，喷洒器在每个点连续工作5 h. 表4-1是农场工人一个周期（4天）的喷灌工作日程表.

4.2.5 模型的分析与评价

该模型通过计算基本水流量的损失，得知水平管道中的水流损失很小，可以忽略不计. 为此根据现已有的常用喷洒器，在相同水压下，给出其喷洒半径与洒水量之间的数量关系，同时得出了喷嘴的流量和射程等参数值. 在此基础上，采用计算机模拟的方法，首先模拟了一个喷嘴喷灌的效果情况（包括喷灌面积、均匀度和喷灌要求的时间），然后用等半径的圆域来覆盖整块田地，将其转化为一个最小覆盖问题，从而找出问题的最优覆盖方案. 最后通过SPSA迭代方法，经过5 000次的迭代模拟计算，确定出喷嘴的最佳放置位置和移动方案. 按照这样的方案，喷洒均匀程度可达89%以上，并且不满足灌溉要求的区域均位于田地的边缘地带，其效果很好.

表 4-1 农场工人一个周期（4天）的喷灌工作日程表

天	时间	喷嘴安装位置坐标为（x，y），坐标原点在左下角处，单位：m
第1天	8am	(1，6.25)
	1pm	(23.75，6.25)
	6pm	关闭喷灌系统，避免过度喷灌
第2天	8am	(46.75，6.25)
	1pm	(69.5，6.25)
	6pm	关闭喷灌系统，避免过度喷灌
第3天	8am	(10，23.75)
	1pm	(32.75，23.75)
	6pm	关闭喷灌系统，避免过度喷灌
第4天	8am	(55.65，23.75)
	1pm	(78.25，23.75)
	6pm	关闭喷灌系统，避免过度喷灌

该模型的优点是一条支管上只需安装一个喷嘴，且每天只需移动2次，大大减少了工人的工作量. 缺点是一次浇水即达到两天的灌溉要求，不利于农田水分的充分吸收. 另外，模型没有对算法的灵敏度进行分析，喷洒器的位置设置过于精确，在实际中可能难以达到这样的要求. 同时，还注意到以下几个方面：

（1）模型没有考虑风向或风力所产生的影响. 因为迎风一面的射程会变短，而顺风一面的射程会变长，并且风向会影响洒水量的均匀度. 如果增加这些因素，则模型和结果会发生改变.

（2）模型的结果是在工人每天工作12小时的假设下得到的，大大高于正常工作日的8小时，使得模型的结果与实际有一定的偏差.

（3）模型用到的旋转喷嘴的内径比问题中要求的小一倍，但仍然采用了问题所给的水压及水速来模拟计算，这样会与实际有一定的误差. 事实上，对于任意给定的旋转喷嘴，可以通过实地测量其喷射的距离及洒水量，设法找出二者之间差别的函数关系，对模型进行修正，就会得到与实际相符合的结果.

（4）模型中的喷洒均匀系数是一个平均值，不同的喷嘴设置方案可能会产生相同的均匀系数. 由于不同的喷嘴射程有重叠，导致有些地方洒水量过多，有些地方洒水量不足，两者平均求得的均匀系数可能会偏大，抵消了部分的不均匀性. 因此这并不是衡量均匀度及洒水量的最精确方法，尤其是我们更关心洒水过多和洒水不足的地方.

（5）为了使模型更真实有效，可以把诸如风、雨及地形等各种因素考虑进去．为了使算法更加普遍通用，可以把喷灌系统的装置、喷嘴大小或现有喷洒器型号等因素都考虑进去，效果会更好，但也会使问题复杂化．

4.3 模型2：多管组喷灌系统的优化模型

通常情况下，工人不应该每天工作如此长的时间，虽然在模型1中要求工人每天只移动两次管组，但是连续工作时间过长，而且每次喷灌的区域一次性达到2 cm，不利于作物充分吸收水分．因此建立一个多管组喷灌系统模型，即每个支管上安装多个喷嘴，且多个支管同时工作的模型，应该会提高喷灌效率，相应的效果会更好．

4.3.1 问题的分析与假设

1. 问题的分析

多管组喷灌系统优化模型主要以“花费时间最少、喷灌尽量均匀并且系统的灵敏性高”等几个方面来设计管组的喷灌实施方案．

首先，根据实验数据获取旋转雾化喷嘴的喷射半径，由实验规律确定喷嘴的数量．其次，将80 m×30 m的田地划分为0.25 m×0.25 m的网格，模拟喷嘴喷水的分布，计算网格节点处喷水分布的均匀度．最后，寻求一种设计方案，包括把管组（即在横向支管上安装喷嘴组）设置在何处以及何时移动管组，同时设计出单一管组的移动时间表和多管组的移动时间表．

为了寻求喷水均匀度更高的方案，可以通过调整喷嘴之间的距离以及各边之间的距离来实现，即使用网格对称地移动两边的管组，一个节点接一个节点地进行测试，从而获得适用于多管组的最佳设计方案．

2. 模型的假设

从实际出发，为了进一步简化问题，做如下假设：

（1）不考虑水分的蒸发和风的影响作用，不考虑移动管组所花费的时间．

（2）所使用的喷嘴为同一规格，喷嘴本身的高度可忽略不计．

（3）相对于喷灌的土地来说，喷嘴的面积很小，所以可将喷嘴看作一个质点．

（4）喷嘴处的水压等于工作水压，喷嘴处的水流量是恒定不变的．

3. 名词与符号说明

为了叙述的方便，下面先给出几个名词的解释和符号的说明．

水源水压：水源管道中水的压强，单位是kPa．

工作水压：类似于水源管道中水的压强，但单位是m，用符号 H_p 表示．转换关系为1 Mpa = 10^3 kPa ≈ 100 m高的水柱，因此，如果水源水压为

420 kPa，则工作水压近似为 42 m.

喷嘴的水流量：喷嘴处单位时间的水流量，用 Q'表示，单位为 m^3/h.

降水量：降落到土地上的水量，用 P 表示，单位为 mm.

分布模式：不同管组放置田间时的降水量.

分布的均匀度：喷洒在田地里的水的均匀度.

4.3.2　模型的建立

对于问题中所给的条件，首先需要明确所给喷嘴的喷射半径和喷嘴流量，然后根据水源水压及流量确定合适的喷嘴个数，最后以降水量要求为约束条件，来设计管组的设置和移动方案，并在此基础上，逐步调整位置使其降水的均匀度进一步提高.

1. 喷嘴喷洒半径的确定方法

喷嘴的喷洒半径与工作水压和喷嘴的型号（如尺寸、喷射角和旋转速度）有关，加大喷嘴的旋转速度会减少喷洒半径的值. 喷洒半径和上述因素的关系可以用下面的关系式表示：

$$R=f(\alpha,\ H_p,\ d).$$

当喷射角 α 一定时，有经验公式：

$$R=\xi H_p^m d^n, \tag{4.3}$$

其中 ξ，m，n 是在各种水压下所得到的测试参数，d 为喷嘴直径.

根据三种不同型号喷嘴的喷洒半径指标值，利用最小二乘估计法，针对 4 个喷射角 α 值，可以估计出对应的参数值，如表 4-2 所示. 将参数值代入 (4.3) 式，很容易求出喷洒半径.

表 4-2　计算喷洒半径的实验参数值

喷射角 α	ξ	m	n
7°	11.46	0.369	0.319
15°	5.61	0.225	0.734
22.5°	8.63	0.140	0.467
30°	4.52	0.128	0.844

2. 喷嘴数量的确定方法

由于喷灌系统中的水流量是守恒的，在同一支管上可以安装喷嘴的最大数量将由主管道的水流量和喷嘴的水流量来决定. 对喷嘴的水流量主要影响因素有两个：工作压力和喷嘴的直径. 实验结果表明，喷嘴的水流量一般与工作压力成正比，通常满足下列关系式[6]：

$$Q'=3600\mu A\sqrt{2gH_p}, \tag{4.4}$$

其中 Q' 是喷嘴的水流量（m^3/h），μ 是喷嘴的喷射系数（通常取值 0.75～0.98），A 是喷嘴的截面积（m^2），H_p 是工作水压（m）. 如果水源的水流量记为 Q，则喷嘴的数量为

$$n=\left\lceil \frac{Q}{Q'} \right\rceil,$$

其中 $\lceil\ \rceil$ 表示向上取整.

3. 喷嘴降水量分布的确定方法

根据相关文献和经验，有一个喷嘴的喷洒器在距离喷嘴为 r 处 T 时间内的总降水量计算公式[7]为

$$P=1000\cdot\frac{2Q'T}{\pi R^2}\left(1-\frac{r^2}{R^2}\right),$$

其中 P 是总降水量（mm），r 是至喷嘴的距离（m），T 是喷洒时间（h）.

根据模型的假设，这里不考虑风的影响，可以认为在等距离的地方喷洒的水是均匀的. 于是，利用上述公式，就可以得到一个喷嘴的喷洒器的在不同距离处的降水量的分布关系，如图 4-10 所示.

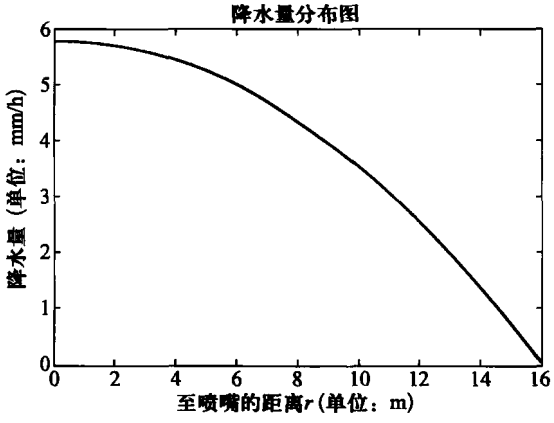


图 4-10 一个喷嘴的降水量与距离的关系图

4. 整块田地降水量的均匀度确定方法

将整个田地划分为小的网格，如取 0.25 m×0.25 m，并且每个喷嘴都是在同一地点连续喷水. 则在单位时间（每小时或每天）内每块田地所获得的平均降水量为

$$\rho=1000\cdot\frac{Q'}{\pi R^2}, \tag{4.5}$$

其中 Q' 是喷嘴的水流量（m^3/h），R 是喷射半径（m）.

在实际中，对田地实施均匀喷灌对农作物的生长起着重要的作用. 这里采用田地网格各个节点处降水量的平均值作为田地降水均匀性的度量指标，即用克里斯琴森（Christiansen）均匀系数 C_U[8]表示，有

$$C_U = \left(1 - \frac{\sum_i \sum_j (h_{ij} - \bar{h})}{N \times \bar{h}}\right) \times 100\%, \tag{4.6}$$

其中 h_{ij} 是网格节点 (x_i, y_j) 处的降水量 (mm/h)，$\bar{h} = \frac{1}{N} \sum_i \sum_j h_{ij}$ 是所有节点 $\{(x_i, y_j)\}$ 降水量的平均值，N 表示节点的个数.

4.3.3　模型的求解方法

根据问题的要求，将上述模型在问题所指定的小农场的田地中进行检验，即将实际数据代入模型中计算求解，则可找到最优的多管组喷灌系统的设置方案并进行喷灌效果分析.

已知小农场的农田面积为 80 m×30 m，支管内径为 10 cm，可旋转喷嘴内径为 0.6 cm，支管可连接的总长度为 20 m. 水源水压为 420 kPa，主管道的水流量为每分钟 150 升. 要求农田任何部分喷灌的水量不得超过每小时0.75 cm，并且每 4 天至少要喷灌 2 cm 的水量.

1. 确定喷嘴的数量

根据已知的主管道的水流量 $Q = 150$ L/min $= 9$ m^3/h，管嘴内径 0.6 cm，工作水压为 $H_p = 420$ kPa $= 42$ m 水高，喷嘴的截面积 $A = \pi (0.3)^2 = 0.282\,7$ (cm^2)，取喷射系数 μ 为 0.8，重力加速度 g=9.8 m/s^2，将其代入 (4.4) 式中得喷嘴的水流量为 $Q' = 2.336\,3$ m^3/h. 则有喷嘴的个数为

$$n = \left\lceil \frac{Q}{Q'} \right\rceil = \lceil 3.8522 \rceil = 4,$$

即这块田地需要设置 4 个喷嘴，就能够很好地完成所要求的喷灌任务.

2. 确定喷洒的均匀系数

按照上面的计算结果，由 4 个喷嘴组成的喷灌系统用于田地的喷灌作业，进一步来研究计算喷灌的均匀系数. 将田地划分为 0.25 m×0.25 m 的网格，此时节点总数 $N = 121 \times 321$. 则由 (4.6) 式，喷灌的均匀系数表达式为

$$C_U = \left(1 - \frac{\sum_{i=0}^{120} \sum_{j=0}^{320} (h_{ij} - \bar{h})}{121 \times 321 \cdot \bar{h}}\right) \times 100\%,$$

其中 $\bar{h} = \frac{1}{N} \sum_{i=0}^{120} \sum_{j=0}^{320} h_{ij}$.

3. 确定管组的位置和移动方案

根据实际情况，可以采用不同的配置方案来实施喷灌. 通过比较分析，找出最优的配置方案和实施工作日程表.

(1) 单个管组的配置方案

如果将所有 4 个喷嘴都安装到一个横向的支管上，构成一个管组系统，那

么由（4.5）式可得，近似的平均降水量 Q 应满足条件：

$$1000 \cdot \frac{4\times\frac{1}{4}Q}{\pi R^2}<7.5.$$

如果选用有30°喷射角的旋转喷雾式喷嘴，那么由（4.3）式和表4-2中的数据，可以计算得到喷射半径 R 大约为20 m.

将4个喷嘴等距离地安装在20 m长的支管上，且支管的两个端点各装一个喷嘴，喷嘴之间的间隔为6.67 m. 假设主管在田地中间横穿铺设，支管垂直安装在主管上，沿主管方向移动，每次可移动20 m（喷嘴的喷射半径）. 实际中考虑到对称性，将喷嘴对称安装在支管上. 管组的配置方案如图4-11所示.

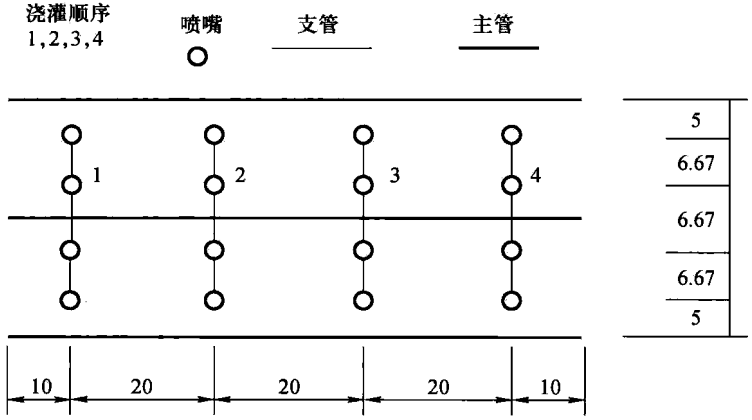


图4-11 单个管组安装位置及移动方案图

浇灌方案的实施以一天为周期，支管在每一处停留喷灌的时间受到两个限制，即对于每一个节点，每小时的降水量不超过0.75 cm，而且总降水量不少于0.5 cm.

在上述条件约束之下，利用计算机编程模拟计算，可以计算得到一个4天连续工作的时间安排表. 要完成整块田地的喷灌任务，最少需要1 228分钟，共用4个周期，喷灌均匀系数 C_U 为78%. 模拟喷灌降水量分布图如图4-12所示[1].

从模拟结果可以看出喷灌均匀系数有些偏低. 事实上，如果试图通过改变横向支管的位置和喷嘴之间的距离来提高 C_U 值，是无法得到令人满意结果的. 而且，喷射角度越大，受风的影响作用就会越明显. 因此，考虑调小喷射角，并采用多个管组系统来实现喷灌将会有更好的效果.

（2）多个管组的配置方案

首先考虑两个管组的情况，每个管组配置2个喷嘴，由（4.5）式，平均

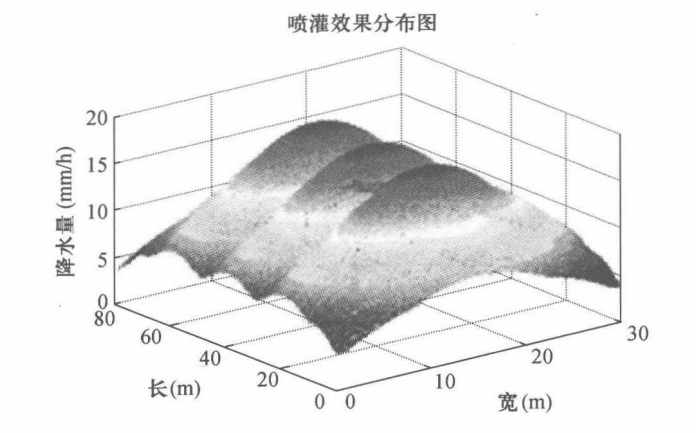


图 4-12　单个管组系统的喷灌降水量分布图

降水量 Q 应该满足如下条件：

$$1000 \cdot \frac{2\times\frac{1}{4}Q}{\pi R^2}<7.5.$$

如果选用 15°喷射角的旋转喷雾式喷嘴，那么由（4.3）式和表 4-2 中的数据，可以计算得到喷射半径 R 大约为 16 m.

假设主管道沿田地的横向边界铺设，支管垂直接入主管，每一支管安装两个相距 5 m 的喷嘴，构成一个管组系统. 两个管组之间相互间隔 10 m，让两个管组交叉移动 20 m 的距离实施喷灌作业. 具体的管组配置方案如图 4-13 所示.

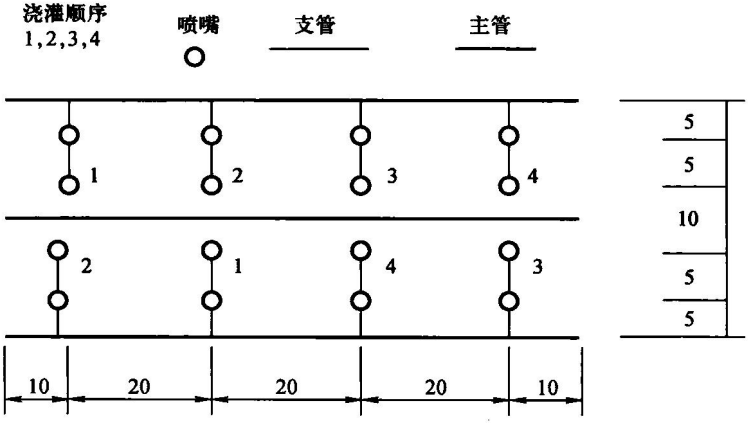


图 4-13　多个管组安装位置及移动方案图

按照这种配置方案实施作业，重叠喷灌部分大大减少，与单个管组的喷灌效果相比（图 4-12），可以大大提高喷洒的均匀度. 同样，喷灌方案的实施以一天为周期，对于每一个节点来讲，每小时的降水量不超过 0.75 cm，而且总

降水量不少于0.5 cm.

在上述条件约束之下，利用计算机编程模拟计算，可以计算得到一个四天连续工作的时间安排表. 要完成整块田地的喷灌任务，最少需要920分钟，共用4个周期，喷灌均匀系数 C_U 为83%. 模拟喷灌降水量分布图如图4-14所示[1].

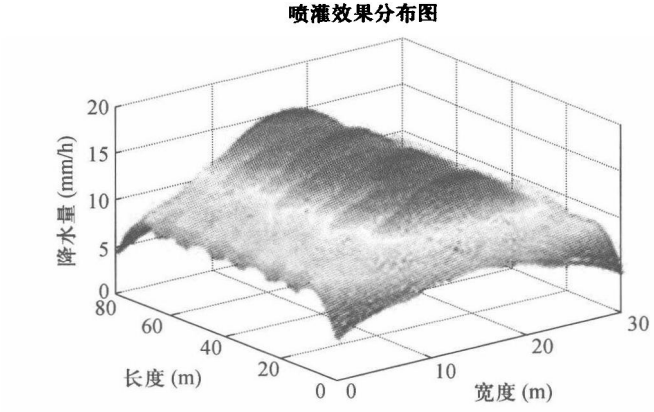


图4-14 多个管组系统的喷灌降水量分布图

对于多个管组的情况，由于中间部分会被重复喷洒，这部分的降水量有些过度. 为了提高整体喷灌的均匀度，需要对配置方案做进一步的改进.

(3) 改进多管组的配置方案

根据多管组配置方案的优点和不足，可以通过调整各支管上两个喷嘴的距离来改变喷灌的均匀度. 为此，将各支管上的喷嘴向田地两边移动，移动的步长按照网格节点逐步进行搜索，同时计算喷灌的均匀系数，则可以找到两个喷嘴的最优安装位置. 事实上，将支管上两个喷嘴分别安装在距离田地边界的3 m和8 m处，即相距为5 m时，喷灌效果是最佳的. 于是可以得到多管组系统的配置与移动方案，如图4-15所示.

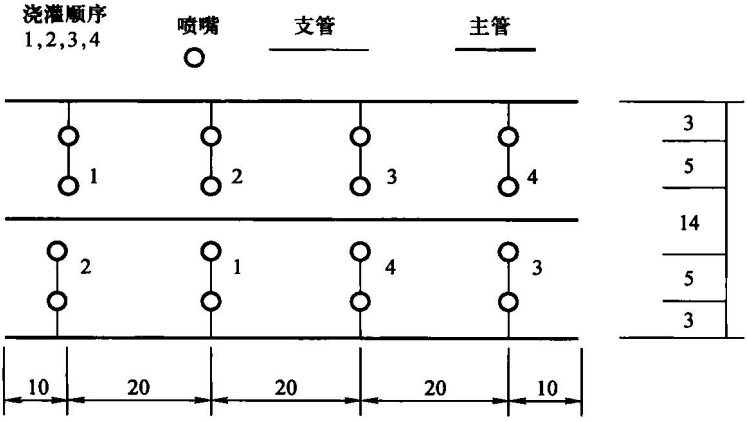


图4-15 改进的多个管组安装位置及移动方案图

按照这个配置方案实施作业，同样在关于降水量的两个约束条件之下，可以计算出每个位置上的喷洒时间长度，设计出一个四天中共进行四个周期喷灌作业时间表．结果表明，最少需要用时 732 min，喷灌均匀系数 C_U 为 90%．喷灌效果分布图如图 4-16 所示[1]．

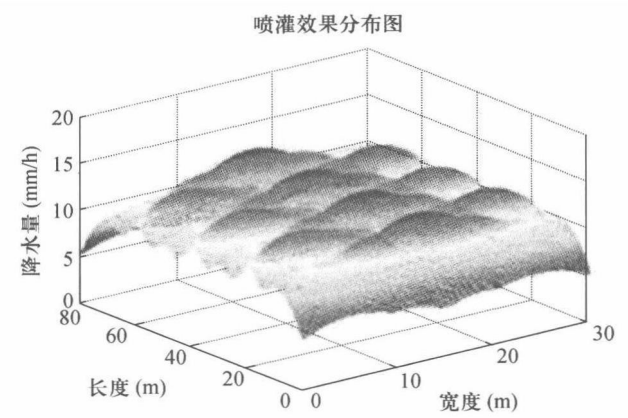


图 4-16　改进的多个管组系统喷灌效果分布图

4.3.4　模型的结果分析

上述模型给出了单个管组、多个管组和改进的多个管组的喷灌系统三个配置方案，移动作业日程表给出了 4 个周期的实施方案．按实际要求喷灌的总时间分别为 1 228 分钟、920 分钟和 732 分钟，喷灌的均匀系数分别为 78%、83% 和 90%．显然，两个管组喷灌效果要优于单个管组的情况．而且改进的多个管组配置方案用时最少为 732 分钟，均匀系数最高为 90%．

在这个模型中，还分别对不同方案的喷灌均匀系数 C_U 和依赖于不同喷射半径的最短时间作了灵敏度分析．结果表明，对于改进的多个管组两个喷嘴的情形，相应的喷灌均匀系数 C_U 随着喷射半径的增加而先增大再减少，当喷射半径为 16 m 时达到最大值，且在 16 m 附近范围内变化相对稳定．按要求实施喷灌的总时间先随着喷射半径的增加而减少，而后随着喷射半径的增加而增大，当喷射半径为 16 m 时达到最小值，而且在 16 m 附近范围所用的时间变化不大，即相对稳定．从这两个方面说明该模型具有一定的鲁棒性．

根据需求，为了获得更好的配置和作业实施方案，模型针对多管组情况，通过在相邻节点上逐步移动喷嘴的距离来进行搜索，计算相应的喷灌均匀系数，随着喷嘴与主管道距离的变化关系，找出最佳的安装位置．结果表明，当喷嘴与田边主管道 3 m 左右时，均匀系数达到了 90%，并且在附近不同网格节点的均匀系数差异不超过 1.3%，即稳定性是很好的．

4.3.5 模型的分析与评价

该模型给出的三个方案在实际中均是可行的，多个管组的配置方案效果更好一些，具有最少的喷灌时间和最大的喷灌均匀度．模型就喷灌的均匀度和依赖于喷射半径的最少时间进行了灵敏度分析，结果表明该模型具有良好的稳定性．

模型通过调节喷嘴的喷射角推导出了喷射半径的关系式，根据喷嘴的实际检验数据，利用最小二乘估计法求解得到喷射半径的计算结果．再根据主管道水源的总流量和喷嘴的水流量关系，确定出喷嘴的配置数量．进一步考虑到降水量是喷射距离的减函数，由此确定出了喷灌系统的实施作业方案．以喷灌时间最少和均匀度最大为优化目标，设计算法搜索出了最优的多管组配置方案．

该模型中采用了灌溉工程中常用的克里斯琴森均匀系数 C_U 作为喷灌均匀度指标，精度比较高．同时采用两个管组交叉移动作业方案，有利于田地对水的充分吸收，喷灌效果好、效率高．但是每天需要移动管组四次，且每次安装两个管组，相对于模型1来说增加了工人的工作量．该模型有以下的特点：

（1）利用仿真计算得到了不同配置方案的喷灌效果分布图．

（2）研究了不同的管组数量和不同的喷射半径对喷灌效果的影响．

（3）简洁直观地给出了较好的系统管组的配置结果和喷灌作业的实施方案．

（4）通过多种检验方法，并不断修正模型，从而得到最优的配置方案．

（5）没有包括现实生活中的一些实际情况，如水的渗透、蒸发以及风的影响．

4.4 模型3：手动喷灌系统的优化模型

一个喷灌系统的设置，主要依赖于喷嘴的喷射速度、喷射半径和喷洒模式等，这些参数有些可以从实际的喷嘴试验数据中获取，也可以从实际问题中通过机理分析方法确定．喷洒模式这里考虑圆盘型和圆环型两种情形．

4.4.1 问题的分析与假设

1. 问题的分析

针对这个实际问题的要求和建模目的，这里利用伯努利能量守恒方程来计算喷嘴的喷射速度、喷射半径和水流速度．先对问题进行简化，忽略一些诸如摩擦力等因素的影响．基于水流守恒原理和伯努利能量方程，首先研究只有一个喷嘴的简单模型，将求解结果应用到一般的相关参数所满足的关系中，得到单个喷嘴的喷射轨迹和射程．然后应用系统的对称性将一个喷嘴的模型推广到

多个喷嘴的系统中.

对于可旋转的喷嘴，所喷洒出的轨迹应为一个均匀的图形，可以看做是一个圆盘域或一个圆环域. 在实际中，针对不同数目的喷嘴情况，应尽量将不同喷嘴喷洒的重叠部分缩小，即减少喷洒水的重叠或不足的区域面积，从而实现喷灌的均匀度最大化. 这个优化问题的结果就决定了系统中喷嘴的间隔距离、移动管组的间隔和喷灌整块田地需要移动管组的次数等方案.

2. 模型的假设

为了合理地简化问题，从实际出发建立模型，给出如下的假设：

（1）土地均是相对平整的，不必考虑由坡度造成的水压（移动方向）等情况的变化.

（2）有一根 80 m 长的主管道纵贯整块田地，主管道上可以横向垂直安装支管道.

（3）所用的喷嘴可以 360°全方位旋转，喷射角度为 30°.

（4）系统中的水流持续稳定，系统可循环喷灌.

（5）不考虑风的影响，喷洒范围总是一个匀称的圆盘域（或圆环域）.

（6）不考虑在喷灌过程中降雨的影响；夜间不进行喷灌作业，即每天至少停开 8 小时.

4.4.2 模型相关参数的确定方法

1. 管组的布局

田地是一块长 80 m、宽 30 m 矩形区域. 主管道纵贯 80 m 长的整块田地，将其嵌入地下，横向可以安装长 20 m 的支管，支管上可以安装若干个可旋转的喷嘴，从而构成一个喷灌系统.

在实际中，将喷嘴对称安装在支管的两端，这样能够保证喷洒出的水也是对称的，而且喷灌效果也好. 根据喷嘴的数目来优化主管道上的接口数目和接口间的距离.

2. 一个喷嘴系统的参数确定

为了便于研究一般喷灌系统的喷洒效果，先考虑一个简化的模型，即只研究一个喷嘴的情况，计算出喷嘴的喷射速度，这可以通过系统内水流量守恒和能量守恒的原理来实现.

（1）水流量守恒原理

假设管道内的水流速是不变的，不考虑摩擦阻力、管道弯曲和喷嘴的影响. 由于主管道的总流量（水源处）为 $Q = vA = 150$ L/min $= 0.0025$ m^3/s，则主管道水流速度 $v = 0.318$ m/s. 根据水流量守恒原理，有

$$v_2 A_2 = v_1 A_1,$$

其中 A_1 表示主管道的截面积，v_1 表示主管道内的水流速度，A_2 表示支管道的截

面积，v_2表示支管道内的水流速度[9]．则有

$$v_2=\frac{0.318\pi\cdot(0.05)^2}{(0.003)^2\pi}=88.3\ (\mathrm{m/s}).$$

这是支管内（即喷嘴）的水流速度，相当于200 mph，即接近喷嘴所能承受的最大速度．

（2）能量守恒原理

从能量守恒的角度来考虑，假设所有经过主管道的水都被喷洒到了田地，不考虑主管道及喷嘴摩擦损失的水．根据能量守恒原理，应用伯努利方程[10]，有

$$\frac{P_1}{\gamma}+z_1+\frac{v_1^2}{2g}=\frac{P_2}{\gamma}+z_2+\frac{v_2^2}{2g},$$

其中P_1和P_2分别表示主管道和支管道（喷嘴）的水压力（单位：$\mathrm{N/m^2}$），γ是水的比重（单位：$\mathrm{N/m^3}$），z_1和z_2分别表示主管道和支管道（喷嘴）距离地面的高度（单位：m），v_1和v_2分别表示主管道和支管道（喷嘴）中的水流速度（单位：m/s），g是重力加速度（单位：$\mathrm{m/s^2}$）．将已知的参数代入公式，得

$$\frac{420}{9.81}+0+\frac{(0.318)^2}{2\times9.81}=0+z_2+\frac{v_2^2}{2\times9.81}.$$

由于喷嘴内的水从喷嘴口喷出后，水就不再受任何压力[10]，因此P_2（喷嘴的压力）为0．所以当喷嘴离地面高度为z_2时，喷嘴处的喷射速度可表示为

$$v_2(z_2)\approx4.43\sqrt{42.82-z_2}.$$

在实际中，喷嘴的高度依赖于农作物高度，一般在6英寸至4英尺（即0.183 m～1.22 m）之间[11]．为此，根据喷嘴高度的变化区间，模拟检测水速对高度的敏感度．这里为了方便计算，将喷嘴的高度从0 m到1 m之间变化．此时有

$$v_2(0)=29.0\ \mathrm{m/s},\ v_2(1)=28.7\ \mathrm{m/s}.$$

可以看出水流的速度对高度并不太敏感．因此在下面的研究中，不妨取喷嘴的高度z_2为0.5 m，则喷射水流速度v_2为28.8 m/s．

（3）喷嘴射程轨迹的确定

喷嘴的喷水射程（即喷射半径）与喷射速度和喷射角有关，为此，根据二者的变化情况就可以计算出喷嘴喷灌的范围．事实上，不管喷嘴喷射角如何，从能量守恒的角度，水的喷射速度都是一样的，为此将其作为初始的喷射速度．同时，注意到最大喷射速度与水的喷出形式无关，尽管有时喷出的水花像是扇形，而不是一股喷射流．由此，根据喷射速度可以计算出一个喷嘴的喷射范围[11]，则有喷射水的轨迹方程为

$$z=\frac{v_{z_0}}{v_{x_0}}x-\frac{g}{2v_{x_0}}x^2.$$

如果整个支管上只有一个喷嘴，利用水流量守恒方程可得到喷嘴的喷射速度为 $v_2=v_1\frac{A_1}{A_2}$. 在同一个支管上每增加一个喷嘴，有效喷灌范围都会相应地增大，而喷射速度会相应地减小. 每个喷嘴的喷射速度相同，与喷嘴个数的关系为

$$v_n=\frac{v}{n}=\frac{28.8}{n}\ (\mathrm{m/s}).$$

其中 n 是一个支管上喷嘴的个数，v 是一个喷嘴的喷射速度. 记喷射角为 θ（喷射水流与水平线间的夹角），考虑喷射水流在水平方向的速度，则可以求出一个支管上安装 n 个喷嘴在水平方向上的射程（即喷射半径）为

$$x=\frac{v\cos\theta\sqrt{v^2\sin\theta-2gn^2z}+v\sin\theta}{gn^2}, \tag{4.7}$$

其中 x 是喷嘴的喷射半径（即水平射程），v 是用能量守恒方程确定的喷嘴的喷射速度，z 是喷嘴距地面的高度，g 是重力加速度.

在实际中，大多数农用喷灌系统喷嘴的喷射角一般为 18°至 35°不等，最大的为 35°[12]，在此不妨取 $\theta=30°$.

3. 多个喷嘴系统的参数确定

根据问题要求，已知田地长为 80 m，宽为 30 m，支管最长为 20 m，所以喷嘴的射程不能少于 5 m 才能够喷灌到田地边界. 因此，在横向支管上设置喷嘴的个数最多为 4 个就可以使总射程充分覆盖到整块田地. 为了确定喷灌系统中多个喷嘴的喷水流量，必须要计算喷嘴的喷水流速. 当系统有 n 个喷嘴时，则每个喷嘴内的水流量为

$$Q_n=v_n\pi\ (0.003)^2\ (\mathrm{m^3}).$$

根据实际情况，不妨假设有圆盘型和圆环型两种喷洒模式，喷洒模式主要由喷嘴形状决定；两种喷嘴喷射出的水都是均匀分布的扇形，即每一种喷嘴都能均匀地喷灌田地. 如果记圆盘型喷嘴喷射范围的外径为 xm，内径为 0；圆环型喷嘴喷射范围的外径为 x m，内径为 $\frac{x}{2}$ m. 则两种喷嘴的喷射范围如图 4-17 所示. 对于圆环型的情况，之所以设置这样的比例，主要是从实际出发，因为如果环形带比较窄，就需要在支管上增加更多的喷嘴来使环心周围都能得到充分的喷灌，会使情况变得复杂.

按照上面的喷灌方法，降水量（随时间喷水深度）D 为

$$D=\frac{Q_n}{A_c}\ (\mathrm{cm/h}),$$

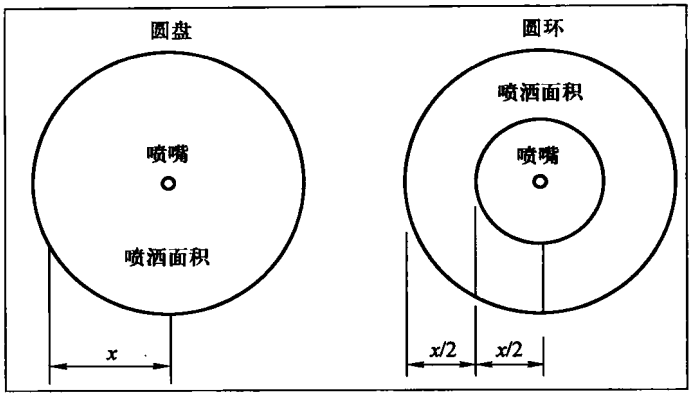


图 4-17 两种类型的喷嘴的喷灌范围的示意图

其中 A_c 是喷嘴喷灌区域的面积，即

$$A_c=\pi(x_o^2-x_i^2),$$

其中 x_o 和 x_i 分别是喷灌区域的外径和内径，圆盘区域的内径为零. 分别对这两情况对应于不同数量喷嘴的喷水量 D 的计算值如表 4-3 所示. 由计算结果可以看出，在一支管上安装多个喷嘴时，虽然每个喷嘴的喷灌覆盖区域面积减小了，但是单位时间的降水量 D 却增大了很多.

由于要求连续喷水量不能超过 0.75 cm/h，每 4 天内的喷水量不能少于 2 cm. 根据表 4-3 的计算结果，对于一个喷嘴的喷灌效率并不高，因为即使是只满足最少量的喷水要求，横向支管也得在一处停留 4 天的时间. 对于两个喷嘴的情况，喷灌效率也不是太高，因为它的喷灌覆盖范围远远超出了这块田地的边界. 不管是圆盘型还是圆环型模式，对于一个或两个喷嘴的情况，水压都接近了喷嘴所能承受的程度. 基于这个原因，这里只研究 3 个和 4 个喷嘴的情况.

表 4-3 不同喷嘴个数的喷射半径、喷嘴流量和降水量

喷嘴个数（在一支管上）	喷射半径（m）	喷嘴流量 Q_n（m^3/h/喷嘴）	降水量 D（cm/h）	
			圆盘型	圆环型
1	74	2.9	0.016	0.021
2	19	1.5	0.13	0.17
3	9	1.0	0.39	0.52
4	5	0.7	0.82	1.10

注意到，对于 4 个喷嘴的情况，喷灌的积水深度可能会超过 0.75 cm/h. 实际中可以适当地减少喷灌时间，即一旦积水深度达到 0.75 cm，可以立即停止喷灌作业.

4.4.3 模型的建立与求解

这里针对圆盘型和圆环型两种模式的情况，分别研究喷灌系统的配置和实施作业的方案.

1. 圆盘型的连续喷灌模型

安装喷嘴的支管与主管道的每一个连接点称为一个喷灌节点. 在一个节点处连续喷灌使其达到喷灌要求的过程称为一次喷灌，在每一个喷灌节点处分别连续喷灌一次后称为一个循环周期.

为了实现系统优化配置和作业，通过合理设置喷灌节点的位置，使得在每一个循环周期中，同一个地方尽量不被重复喷灌，即使喷灌重叠的区域尽量小. 这样也可以使在同一个循环周期中，在每一个节点处连续喷灌的时间最优化.

在保证能够喷灌到田地边界的条件下，一个支管上的喷嘴间距应尽可能小. 因为要使喷灌速度最大化，就会使田地边界一些小的三角形区域可能出现喷不到水或只有少量水的情况. 我们的方案使得这样的区域尽量小，同时，不妨认为土地有足够强的渗水能力，没有喷灌到水的地方可以从周围的土壤中吸收到足够的水分.

当要在一支管上安装4个喷嘴时，在支管的两端各安装一个，这样使靠近边界的区域尽可能多地喷灌到水. 即使有没有喷灌到水的地方，也要使这部分区域尽可能地小.

当要在一支管上安装3个喷嘴时，将一个安装在支管的中间，另外两个安装在距支管两端点1.1 m的地方. 比较理想的状态是喷嘴的喷射范围能够覆盖到田地的边界，并且可以触及相邻的圆盘区域，这样可以使得喷洒到边界外的水尽量少，以有效地利用水资源.

为了确定喷灌节点之间的距离，在这里应用著名的毕达哥拉斯定理的相关公式：

$$R^2=\left(\frac{1}{2}L_s\right)^2+\left(\frac{1}{2}S_s\right)^2,\quad L_s=2\sqrt{R^2-\left(\frac{1}{2}S_s\right)^2}$$

其中 L_s 是支管道沿主管道一次移动的距离，S_s 是两个喷嘴之间的距离，R 是喷嘴的喷射半径，如图4-18所示.

通常支管沿着主管道移动距离应为一个常数，并且尽可能大一些为好. 主管道与田地的长边界平行铺设，并且与边界线有适当的距离，以使主管道上每个节点都与边界线有相同的距离，系统的设置方案要使总的有效喷灌覆盖面积最大. 支管的移动距离、喷嘴的间距以及一个周期内的移动次数如表4-4所示.

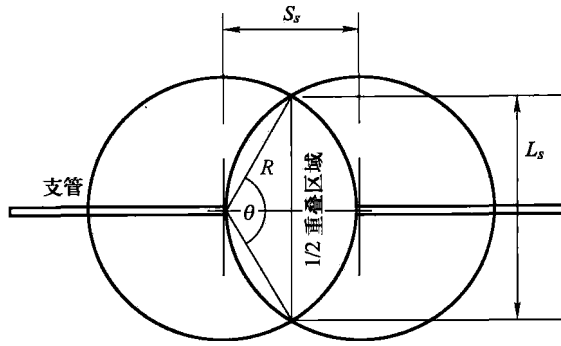


图 4-18 两个喷嘴喷灌的重叠区域示意图

表 4-4 支管的移动距离、喷嘴的间距和移动次数

喷嘴数	喷洒半径（m）	喷嘴的间距（m）	支管间距（m）	移动次数
3	8.9	8.9	15.4	6
4	5.3	6.7	8.3	10

为了计算重叠喷灌区域的面积，将等腰三角形底边的长记为 L_s，两个等腰边长正好是喷射半径 R，如图 4-18 所示. 由余弦公式 $L_s^2=2R^2-2R^2\cos\theta$，容易求得喷射角 θ；在计算重叠部分区域的面积时，可以用 S_s 代替 L_s，由 $A=\theta\pi R^2/360$ 就可以计算出图中扇形的面积，从中减去三角形的面积，然后再乘 2 倍就可得到重叠部分区域的面积.

由此计算得到的重叠部分区域面积、一次喷灌部分面积、没有喷灌到水的部分面积和喷洒到边界外部分的面积计算结果如表 4-5 所示.

表 4-5 喷灌区域的面积（单位：m^2）

喷嘴数	重叠喷灌区域	一次喷灌区域	未喷灌区域	边界外区域
3	1 131	1 250	19	717
4	1 057	1 281	59	140

在一个周期中，尽管 3 个喷嘴的重叠部分面积比 4 个喷嘴时的大一些，但没有喷灌到水的区域面积却比 4 个喷嘴时小得多. 问题中并没有要求每一寸土地都必须喷灌到，而且在一个周期中，没有一个地方会被喷灌两遍以上. 按步骤一个节点挨一个节点移动支管道，在满足喷灌要求的条件下，使其移动距离和移动时间都尽可能小.

要计算出在每一个喷灌节点处的连续喷灌时间，以免被重叠喷灌部分区域的喷水量超过 0.75 cm/h. 对于 3 个和 4 个喷嘴的情况，喷水速度都比实际允许的速度快，特别是在重叠部分区域. 所以，必要时必须及时关闭水阀门，等

待水分慢慢渗入土壤中去.

2. 喷灌时间和日程安排方案

首先确定需要喷灌的时间，即考虑两个相邻喷嘴喷灌到重叠区域的喷水量达到0.75 cm的时间. 对于3个喷嘴和4个喷嘴的情况，重叠区域的喷水量D的值应该是表4-3中相应值的2倍，因此，在一个喷灌节点处的连续喷灌时间分别为

$$t_3=\frac{0.75}{2D_3}=0.960\ (\text{h})=58\ (\text{min}),\ t_4=\frac{0.75}{2D_4}=0.455\ (\text{h})=27\ (\text{min}).$$

然后，要确定为了喷灌整块田地所需要的喷灌节点数. 事实上，当一个支管用3个喷嘴时，整块田地共需要6个喷灌节点；而当一个支管用4个喷嘴时，整块田地共需要10个喷灌节点.

最后，要确定4天内的循环喷灌的次数. 根据实际需求，对于重叠喷灌部分区域每次循环的喷水量不超过0.75 cm. 在均匀喷灌的条件下，对于非重叠喷灌的区域每次循环的喷水量就应超过$\frac{0.75}{2}=0.375$ (cm)；为了在4天内实现2 cm的喷水量，就需要在4天内进行$\left\lceil\frac{2}{0.375}\right\rceil=\lceil 5.3\rceil=6$次循环. 因此，在4天内应该进行6次循环喷灌作业，就能够实现整块田地喷灌要求.

对于3个或4个喷嘴的情况，平均移动一次管组的时间为15分钟. 则相应的喷灌时间、喷灌节点个数、循环次数等结果如表4-6所示.

表4-6　配置方案与喷灌作业计划表

喷嘴数	喷灌时间 (min)	喷灌节点数	4天的循环次数	4天的总喷灌次数	每天的喷灌次数
3	58	6	6	36	9
4	27	10	6	60	15

根据表4-6中的计划，就可以给出3个喷嘴和4个喷嘴情况时的实施作业日程表. 对于每种情况，4天内都需要进行6次循环喷灌作业，即每天需要1.5次的循环喷灌作业. 喷嘴数无论是3个还是4个，由于4天需要的总工作时间为43小时，所以每天的平均工作时间为11小时. 但是，对于4个喷嘴情况，因每天的喷灌次数较多，所以还需要更多移动管组的时间. 由于要求使移动管组的时间尽量小是问题要求的一个目标，因此，我们推荐采用3个喷嘴的喷灌系统，在同等条件下，可以省时省力，喷灌效率高，并且没有被喷灌到水的区域的面积只有19 m^2.

3. 圆环型的连续喷灌模型

实际中有些喷嘴的喷灌范围呈圆环形. 如果圆环外径是内径的两倍，则圆

环型的喷灌范围使模型变得相对复杂，主要表现在以下方面.

（1）不能保证在一次循环喷灌作业中没有被重复喷灌两次以上的区域.

（2）在一次循环喷灌作业中，为了使整块田地都至少喷灌到一次，则会有一部分区域要被重复喷灌3~4次，如图4-19所示[1].

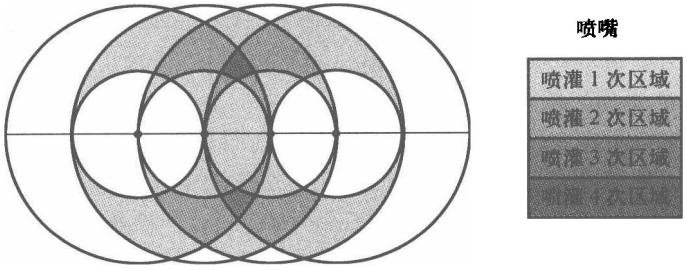


图4-19 圆环型的连续喷灌区域的示意图

（3）由于有些地方被重复喷灌的次数增多，那么每次连续喷灌的时间就要缩短，以免喷水量超过0.75 cm/h，这样也就需要更频繁地移动管组，从而增加了移动安装次数.

（4）为了使一些轻度喷灌的区域也能满足最低喷水量2 cm/4天的要求，就必须要增加循环喷灌的次数.

（5）为了防止一些区域喷水量过度，而另一些区域喷水量不足的不均匀情况，需要有一个在所有喷灌节点上交错实施喷灌作业的规划方案，显然这是一个较复杂的事情. 由此，也会导致管组的移动距离的增加和不便利.

（6）这种情况很难达到喷灌均匀的要求.

在实际中，通常喷嘴在支管上的安装是等距的，因此不考虑田地中间区域喷洒不到水的情形. 事实上，只需要已知喷灌的圆环区域的外径和内径之差，就可以此作为喷嘴的设置间距，即可以计算出管组移动的间隔距离和喷灌节点的数目等参数.

为了解决这个问题，要寻找针对一个循环周期的移动喷灌节点序列的优化算法，可以选择以下两个优化目标：

（1）一个循环周期中使管组的总移动距离尽量小. 用类似于圆盘型的情况，将管组沿着主管道上每个喷灌节点依次地移动搜索计算，一直到主管道的另一端. 因为这种情况会有大量多次重叠的区域，所以为了要满足最大喷水量不超过0.75 cm/h的要求，在规定的喷灌时间内可能需要关闭一些喷嘴令其等待一段时间. 这样就会出现间歇性的喷灌作业，不利于人工实施，同时在总的时间上喷灌效率也比较低.

（2）要求喷灌期间的间断等待时间尽量小. 这个要求是在完成一次喷灌后，要立即移向下一个喷灌节点. 有时为了避免某些区域喷水量过度，在一个喷灌节点喷灌完成后，需要跳过邻近的节点，移向下一个的喷灌节点. 为了提

高喷灌效率，这就需要寻找一个喷灌节点移动的优化算法，将喷灌等待的总时间尽量小作为优化目标．但是，这样做可能会出现在两次喷灌作业之间管组需要移动的距离要大很多（即在一个循环周期中，管组需要移动的总距离）．

在实际操作中可以综合考虑问题，既允许在两个喷灌节点之间有适当的等待时间，又允许在移动管组中跳过邻近的喷灌节点，选择下一个合适的喷灌节点实施作业．这里采用优化目标（1）．

4. 管道水压损失的讨论

在前面的模型中，没有考虑水与管道壁之间的摩擦阻力导致水的能量损失，即水压损失．在实际中，这种水压损失或多或少都是客观存在的，应该予以讨论．这里应用哈森-威廉姆斯（Hazen-Williams）摩擦方程来求解管道内的水压强损失（单位长度的管道内水压损失）．即

$$S_f=\frac{10.7}{D^{4.87}}\left(\frac{Q}{C}\right)^{1.852}, \tag{4.8}$$

其中 S_f 是单位长度（每米）的管道内水压损失，即水压强损失（单位：m），D 是管道的内径（单位：m），Q 是管内的水流量（单位：m^3/s），C 是哈森-威廉姆斯摩擦系数，对于铝制管道，一般取 $C=130$．

将管道的实际参数值 $D=0.1$ m，$Q=0.002\ 5\ m^3/s$ 及 $C=130$ 代入（4.8）式中，可得管道内的水压强损失 $S_f=0.001\ 5$ m．或者说管道内的水流过 100 m 长的管道总水压损失为 $S_f=0.15$ m．这个结果表明，在均匀的管道中水压损失几乎是微不足道的．

当在一支管上设置的喷嘴数量较少时，喷嘴之间的间距就要增大，要实现整块田地的喷灌作业，就需要在主管道上增加喷灌节点，也就是需要更多的连接支管和主管道的安装接口．由实际检测可知，每一个安装接口处的水压强损失为 0.3 m 至 0.4 m．当一个横向支管距离水源较远时，这种水压强损失累加起来就会很大，将大幅度降低水压力，从而减小喷嘴的喷射速度和喷灌半径．要解决这个问题，可以通过减少主管道上的喷灌节点来改进．

如果整个主管道只有一个出口，即所有水流都集中到一个喷嘴上，通过（4.8）式检测计算水压损失，则得到了一个惊人的结果，其水压强损失为 1 305 m．这进一步地说明了一个事实，并不是所有管道内的水流都能从喷嘴口喷出来，会有一定量的水流量按照某种特定的形式产生回流．为此，可以通过设置不同数量的喷嘴，来调整合适的水流量，使管道内所产生的水压强损失达到最小．通过计算比较分析，得到结果为 4 个喷嘴为最好，相应的水压强损失为 12.5 m．设置 3 个喷嘴时，水压强损失为 21.4 m．再次应用伯努利方程，并在方程的两端减去由管道摩擦造成的水压损失，重新计算喷嘴的喷射速度 v_2，有如下的结果：

- 当取 3 个喷嘴时，喷嘴的喷射速度为 $v=8.9$ m/s．不考虑水压强损失

时，速度为9.6 m/s.

● 当取4个喷嘴时，喷嘴的喷射速度为 $v=6.7$ m/s. 不考虑水压强损失时，速度为7.2 m/s.

当取3个喷嘴时，喷嘴的射程可以触及田地的边界；当取4个喷嘴时，因喷射速度太低，此时喷嘴的射程难以保证触及田地边界.

4.4.4 模型的模拟求解算法

这里利用计算机来模拟喷灌系统及喷灌效果，并利用遗传算法找出其近似的理想解.

1. 计算机模拟

将整块田地划分为一个由200×75个小方格组成的网格图，每一个小方格的边长为0.4 m，并且将时间离散化，每一段时间间隔为5分钟；记录每一个小方格在4天内每一次喷灌所获得的喷水量. 计算机模拟需要输入如下信息：

(1) 确定主要的移动方向是横向还是纵向（主管道的方向），纵向移动所产生的水压强损失是不可避免的.

(2) 主管道与田地边界的距离.

(3) 第一个喷嘴到支管始端的距离和最后一个喷嘴到支管末端的距离.

(4) 喷嘴的个数，喷嘴等距离地分布于支管上.

(5) 设计喷灌作业日程表，要求给出每一个支管在主管道上的安装位置；支管在一个喷灌节点连续喷灌的时间.

另外，每移动一次支管所需要的时间为15分钟，每天喷灌作业的“极限”时间为16小时，保证夜晚停工休息.

根据前面对喷嘴处的水压和喷水速度的计算方法，计算机在模拟计算的每一步都对喷嘴射程和在每一个小方格内的喷水量进行了记录. 直到模拟结束，就可以计算出每个小方格内总的喷水量. 最终结果可以分为以下三类：

(1) 喷水量过度. 如果在该方格每小时内的喷水量超过0.75 cm，即便是在4天内总的喷水量低于2 cm，也被认为是喷水量过度.

(2) 喷水量不足. 如果喷水量没有过度，而4天内总的喷水量又小于2 cm，就认为是喷水量不足.

(3) 喷水量适度. 如果在该方格内既不属于喷水量过度，又不属于喷水量不足，即可认为是在理想的喷水量范围之内.

2. 遗传算法的求解实现

在这里，利用遗传算法通过遗传基因的演变来寻求问题的最优解.

首先，随机产生100组输入数据信息，并分别对每一组信息（即基因组）进行测试，选择排在前10位的最好的基因组作为父辈，从而产生出新的100组输入信息（下一代）.

然后，再从这新的 100 组中择优选出 10 个最好的基因组，作为新一个父辈．这样两个父辈各自内部 10 个基因组的再次组合就可以产生出 90 组新的基因组（10 个基因组两两结合就可以产生出 45 种组合）；其中每一个基因组的输入值都取其父本和母本的平均值，并且掺杂小部分的随机变动．

最后选出的 10 个基因组都来自同一个父辈，并伴随有轻微的变动．从理论上来讲，计算机不断地产生出新一代子集，就可以不断地提高解集的最优性．而事实上，输入大量数据，只有较少优良解的输出，这意味着要降低解的最优性很容易，但在迭代了若干步后，再要提高其最优性却是很困难的．所以通常进行到一定的代数之后，就可以终止程序的迭代，输出优良的结果．计算机模拟结果如图 4-20 所示．

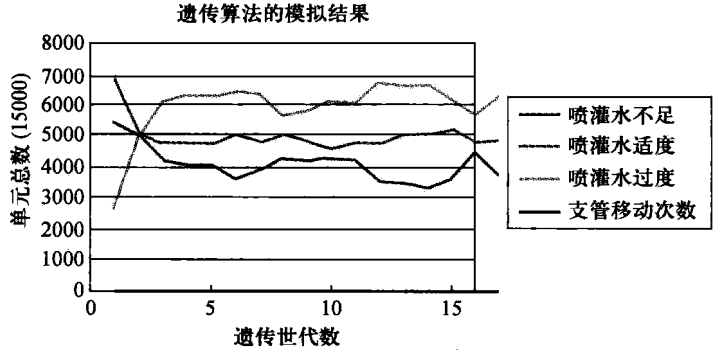


图 4-20　经过 3 ~ 4 代之后的程序运行结果图

当有足够多的子代时（几万个），从所有子代中找出一两组较优的解是可能的；由于计算能力有限，在这里只运用部分输入变量，模拟求解出了较少的一些子代．如果不限制时间，则可能计算出更多的子代，以找到更好的解．

在模拟过程中，对某些特殊的变量倾向于取某些特定的值．当设置 3 个或 4 个喷嘴时，计算机模拟结果总是会选择 3 个喷嘴作为最优值．同时，设置主管道的位置总是在田地横向距离的中间部位为最佳．在迭代 5 ~ 10 代中，第一个喷嘴和最后一个喷嘴的位置总是设在支管的两端点为好．这样设置喷嘴，能使得喷灌不到水的田地面积达到最小．

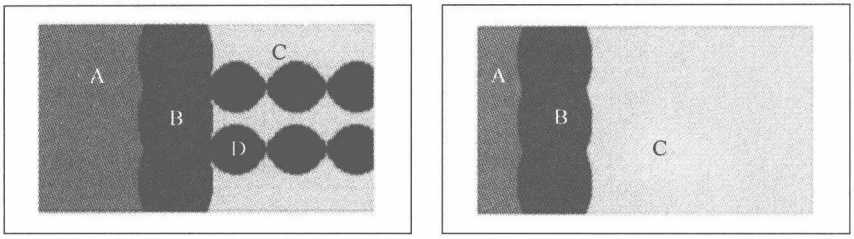
在模拟计算中，我们并没有记录喷洒出田地边界外的喷水量．如果喷嘴间距设置得足够合理，那么通过降低重复喷灌部分区域的面积和减少喷洒出边界的水量，就可以使喷洒不到水的区域达到最小．

3. 喷灌节点的移动方案

根据前面的计算分析可知设置 3 个喷嘴是合乎实际需要的．这时候即需要 6 个喷灌节点，共需要 6 次循环，每个节点每次需要平均喷灌时间不多于 57 min，如表 4-6 所示．我们分别选取在每个节点连续喷灌 60 min 和 55 min 的两种情形模拟检验，其中有 60 min 的情况会导致重复喷灌的区域喷水过量，

如图4-21所示[1].

图4-21（a）为在每个喷灌节点每次连续喷灌60分钟的情况，图4-21（b）为在每个喷灌节点每次连续喷灌55分钟的情况. 图中区域A表示喷水不足或未喷到水的方格；区域B表示正在喷灌中的方格；区域D表示喷水过度的方格；区域C表示喷水适度的方格.



(a) 每次连续喷灌60 min　　(b) 每次连续喷灌55 min

图4-21　基于节点的日程表模拟结果

图4-21（a）进一步证明了在一个喷灌节点连续喷灌60 min会有重叠区域出现喷水过度，所以在一个喷灌节点的喷灌时间不宜过长. 图4-21（b）说明了在一个喷灌节点连续喷灌55 min不会有喷水过度的方格出现. 为此，也证明了我们模型的可靠性.

现在的问题是：如何实施连续的喷灌作业呢？实际中，通常的方案是先从田地的一头开始，每隔55 min移动一个喷灌节点，直到田地的另一头；然后反方向继续下去. 如此周而复始，直到在每一个喷灌节点处都完成喷灌作业6次，共需要移动支管的距离为 $(80-4\times2)\times6=432$ m. 这种循环往返的作业方案，看起来似乎是一个高效的作业方案，但是还有没有更好的方案呢？

4. 喷灌节点移动方案的优化方法

在这里，利用遗传算法来寻求管组在喷灌节点上移动的最优方案，优化目标是使管组总移动距离最小. 算法的每一步都只对应于一个喷灌节点，对每一个喷灌节点都要实施6次喷灌作业，每次的连续喷灌时间为55 min. 但不要求喷灌节点的作业次序.

采用下面的实施方案：从田地一头的第一个喷灌节点开始，在每一个喷灌节点处连续喷灌作业55 min后，立即关闭喷嘴阀门，并等待15 min后再打开喷嘴阀门并连续喷灌55 min，然后再关闭喷嘴阀门等待15 min. 这样连续进行6次作业，所有管组都没有移动，而是在原地等待下一次喷灌作业，直到在一个喷灌节点处完成6次喷灌，才移动管组到下一个喷灌节点上. 依次类推，直到完成整块田地喷灌作业. 每个喷灌节点移动一次即可完成全部的喷灌作业.

通过计算机编程计算，经过5代遗传过后，遗传算法的计算结果很稳定，如图4-22所示. 按照这种方案，在4天的喷灌作业中总的移动距离仅为

72 m，比前面的移动距离的“最优值”432 m 有了很大的改进.

当然，这种方案可能不利于土壤在较短的时间内吸收足够多的水分，但是，它却完全符合这个问题提出的要求. 每次在每个喷灌节点处连续喷灌时间仅为 55 min，并且只有 15 min 的等待时间，任何一个重复喷灌区域的喷水量都不会超过 0.75 cm/h.

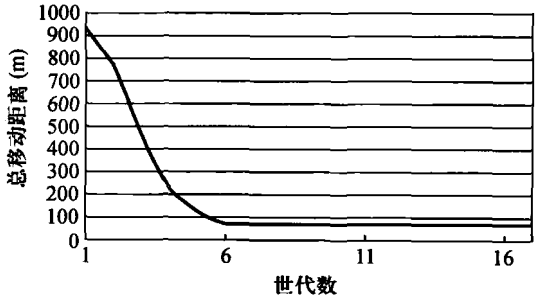


图 4-22 喷灌作业总的平均移动距离

4.4.5 模型的结果分析

该模型针对整个喷灌系统所设置的喷嘴数目，确定出了喷嘴的喷灌射程以及多个喷嘴的设置位置. 在不考虑管道摩擦所产生的水压损失的情况下，发现喷嘴距离地面的高度对喷嘴的水流速度影响不敏感. 分别考虑了圆盘型和圆环型两种喷灌模式的区域覆盖模型，同时利用喷灌区域面积和喷水流量计算出整块田地的平均喷灌量. 对于 1～2 个喷嘴的情况，因管道内水压过大会导致喷嘴的射程过远、喷灌的时间效率低、喷射速度过快和水的浪费过大，所以排除了这两种情况. 而对于 5 个喷嘴情况，管内的水压过小，喷嘴的射程太近，不能触及田地的边界处. 因此，模型只研究了设置 3 个或 4 个喷嘴的情形.

利用直观几何方法，对于圆盘型喷灌区域的情况，模型确定出了 3 个喷嘴和 4 个喷嘴的最佳距离间隔. 当取 3 个喷嘴时，距支管末端 1.1 m 处设置一个喷嘴，另外两个喷嘴设置间隔均为 8.9 m，即两端的喷嘴距离边界均为 6.1 m. 当取 4 个喷嘴时，在支管的两端点处各设置一个喷嘴，另外两个均匀地设置在中间，4 个喷嘴的距离间隔均为 6.7 m.

类似地，对于横向支管的间距、喷灌节点个数和移动管组的次数，也都只针对 3 个喷嘴和 4 个喷嘴两种情况进行分析讨论. 对于 3 个喷嘴的情况，喷灌节点的间距为 14.5 m，共需要在 6 个喷灌节点上作业；对于 4 个喷嘴的情况，喷灌节点的间距为 8.3 m，共需要在 10 个喷灌节点上作业. 其中 3 个喷嘴较 4 个喷嘴的优越性，主要是在每一次循环作业中需要移动管组的次数较少，同时喷灌不到水的区域面积相对较小.

为了满足问题中最低喷水量的要求，那些在每一次循环中只被一个喷嘴喷

灌到的区域的喷水量是决定循环次数的主要因素，不管对3个喷嘴还是4个喷嘴均如此. 为了平衡每天的工作量，每天需要完成1.5个循环喷灌周期. 为了使重复喷灌区域的喷水量不超过0.75 cm/h，每个喷灌节点每次连续喷灌时间和移动管组的次数不同，对于3个喷嘴，每个喷灌节点连续喷灌58分钟，每天移动管组9次；对于4个喷嘴，每个喷灌节点连续喷灌27分钟，每天移动管组15次. 事实上，从计算机模拟结果来看，在每个喷灌节点每次连续喷灌时间减少2～3分钟，也能达到喷灌的基本要求，在实际中也有利于人工操作，因为人工操作不能保证那么准时.

在针对圆环型喷灌区域模型中，为了确保所有区域适度的喷水量，需要频繁地移动管组和更多的循环喷灌次数，并且使管组的移动和喷灌作业的程序变得复杂.

模型在考虑了由于管道内摩擦阻力所引起的水压损失之后，应用伯努利能量方程重新求得喷嘴喷射速度，并与先前未考虑水压损失的喷射速度作了比较分析. 即当取3个喷嘴时，考虑水压损失的喷射速度为8.9 m/s，没有考虑水压损失的喷射速度为9.6 m/s；当取4个喷嘴时，考虑水压损失的喷射速度为6.7 m/s，没有考虑水压损失的喷射速度为7.2 m/s. 这也充分说明了管道内的摩擦阻力所引起的水压损失对喷射速度的影响是明显的.

最后，模型利用3个喷嘴最优的配置参数，在计算机上进行了仿真模拟，给出了一个为期4天的理想的喷灌作业实施日程表. 另外，我们发现把两个喷嘴安装在支管的两端，并相应地调整喷灌节点的间距，可以使得喷灌不到水的区域面积减少到0. 这是很有实际意义的一个结果.

4.5 模型4：基于不同特性的喷灌系统模型

实际中，通常是用一种低成本的铝管和可转动的喷嘴组成可移动的喷灌系统. 对于这样一个小规模的喷灌系统，如果没有一个有效的喷灌作业实施计划，则会对作物的浇水不当，不利于作物的生长，或喷灌效率低下，浪费灌溉用水. 为此，我们通过数学模型来确定一个手动的喷灌系统，尽可能在最短的时间内完成田地的均匀喷灌. 同时评价并测试各种喷灌算法的有效性，选择最有效的喷灌策略.

4.5.1 模型的预备知识

从实际出发，基于问题的具体要求，下面给出模型所需要的一些主要参数的确定方法和系统的评价指标.

1. 喷嘴的喷水速度的确定

针对要解决的问题，根据能量守恒原理，我们可以推导出喷嘴的喷水速

度、喷水量及水滴的风阻系数的关系式，并针对70H1型黄铜制的喷嘴进行计算. 根据上述假设，利用伯努利方程，有

$$P_1+\frac{1}{2}\rho v_1^2+\rho g y_1=P_2+\frac{1}{2}\rho v_2^2+\rho g y_2,$$

其中 P_1，P_2 分别表示管道内水压和喷嘴内的水压；ρ 表示水的密度；v_1，v_2 分别表示管道内的水流速和喷嘴的水流速；g 表示重力加速度；y_1，y_2 分别表示管道和喷嘴距地面的高度.

注意到，所考虑的田地是平的，所以不妨设 $y_1=y_2$，即管道和喷嘴的高度对水流速度没有影响. 于是有

$$v_2=\sqrt{\frac{2}{\rho}(P_1-P_2)+v_1^2}.$$

为此，要确定喷嘴的水流速度 v_2，就要先给出水源（即管道内）的水流速度 v_1，即

$$v_1=\frac{150}{1000}\frac{1}{\pi(0.05)^2}\ (\mathrm{m/min})=\frac{1}{\pi}\ \mathrm{m/s},$$

于是喷嘴处的水流速度为

$$v_2=\sqrt{\frac{2}{1000}\cdot 420\times 1000+\frac{1}{\pi^2}}\approx 28.9\ (\mathrm{m/s}).$$

从这个结果可以看出，喷嘴的水流速度为60 mph，即每小时60英里，实际中是不可能的，这是因为没有考虑管道摩擦的影响. 因此引入衰减因子 C_s 的概念.

2. 喷水量和衰减系数的确定

喷嘴的喷水量 Q 为水流速度 v_2 乘以喷嘴的截面积 A_s，再乘以衰减因子 C_s，则

$$Q=C_sA_sv_2=C_sA_s\sqrt{\frac{2}{\rho}(P_1-P_2)}\ (\mathrm{m^3/s}).$$

根据70H1型黄铜喷嘴的参数[13]，得到其衰减因子为

$$C_s=\frac{Q}{A_s\sqrt{\frac{2}{\rho}(P_1-P_2)}}=\frac{\frac{3.17}{3600}}{\pi(0.003175)^2\sqrt{800}}\approx 0.983.$$

显然，衰减因子对水流速度的影响是很小的，即

$$v_2=0.983\times 28.9\approx 28.5\ (\mathrm{m/s}),$$

则喷嘴单位时间的喷水量为

$$Q=C_sA_sv_2=28.5\pi(0.003175)^2=0.0008058\ (\mathrm{m^3/s})\approx 48.35\ (\mathrm{L/min}).$$

因此，如果使用3个喷嘴，则总喷水量不会超过150 L/min. 喷嘴数超过3个时，将会使得水压下降. 为了确定增加喷嘴后新的水压变化，根据能量守

恒原理，即流入的总水量等于流出的总水量，有 $A_s v_s = nA_0 v_0$，其中 A_s，v_s 分别表示主管道的横截面积和水流速度，n 为喷嘴个数，A_0，v_0 分别表示喷嘴的横截面积和喷水速度. 于是有单个喷嘴的喷水速度为

$$v_0 = \frac{r_s^2}{n\pi r_0^2} = \frac{(0.05)^2}{n\pi (0.003)^2} \approx \frac{88}{n} \ (\mathrm{m/s}),$$

这里 $n>3$，r_s，r_0 分别表示主管道和喷嘴口的半径. 显然当 $n=4$，即有 4 个喷嘴时，每个喷嘴的喷水速度为 $v_0=22$ (m/s)，相应的水压为 252 kPa. 而 70H1 型黄铜喷嘴所需要的水压大约为 280 kPa. 因水压相差太多，所以，我们认为最多只能使用 3 个喷嘴.

3. 喷嘴降水量的衰减分布

根据问题的具体要求和条件，运用伯努利方程求出喷嘴的水流速度为 28.5 m/s，单位时间内喷嘴的流量为 48.35 L/min (模型 3). 同时，也已说明在 20 m 长的支管上安装 3 个喷嘴的最好的方案 (参见模型 2). 由资料可知喷嘴的喷射半径可以达到 19.5 m，并且根据喷射半径、喷射角和喷射速度的关系，在一定的范围内调整喷射角可以得到不同的喷射半径.

虽然喷嘴的喷射半径 R 可达 19.5 m，但是在以喷嘴为中心，半径为 R 的圆域内，各点的降水量是不同的，且随着距离的增加而减少. 在一般情况下，可以认为是直线型衰减或指数型衰减. 根据 Louie 等人[12]的测试结果，规格为 4.73 mm 的喷嘴的喷水情况，在喷嘴附近的两米范围内水的分布先是呈尖刺状快速递减，然后基本保持稳定，在接近区域边缘时快速递减到 0. 为此，我们用指数函数对曲线进行拟合，喷嘴的降水量随距离 r 的变化函数为 $f(r)$，即喷嘴降水量的分布为

$$f(r) = (3\times 0.00267\times e^{-0.7r} + 0.00267)\times e^{-(r/19.5)^{20}}.$$

其曲线如图 4-23 所示.

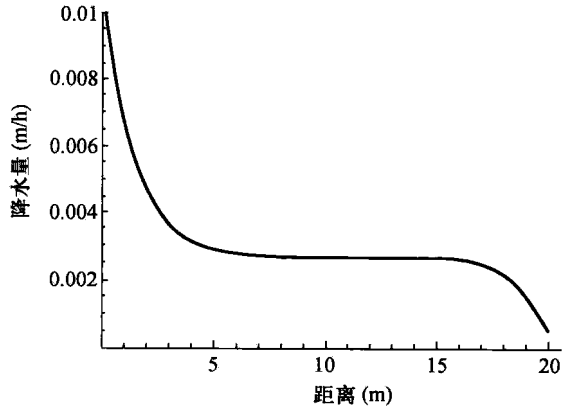


图 4-23 喷嘴降水量与距离的分布

按照这种指数衰减的喷水分布规律，距离喷嘴很近范围内（如 $r<1$ m）的降水量将会超过 7.5 mm/h，最多可达 10 mm/h. 因此，根据问题的要求在设计喷灌方案时，连续喷灌作业的时间不应超过 1 小时，大致应为不超过 45 分钟为好. 同时，对于距离喷嘴 4 m 到 18 m 的区域，喷嘴的降水量基本都在 2.25 mm/h 左右，如果按照连续喷灌 45 分钟计算，这部分区域的降水量可以达到 1.68 mm.

4. 喷灌系统配置方案的评价

一个喷灌系统配置方案的优劣，在满足问题的要求和约束条件下，可以用管组的移动次数、系统的作业时间、水的有效利用率、对于风与其他干扰因素的稳定性和喷灌的均匀性等指标来评价.

喷灌系统的作业时间应该由接收降水量最少的区域来确定；水的有效利用率可用整块田地均达到 2 cm 的降水总量与喷灌系统降水总量的比值来表示，比值越高，则表明水的利用率就越高；或者用喷灌系统 4 天内的降水总量来表示，这个数值越高，则表明水的利用率就越低；喷灌的均匀性用降水分布的方差来确定，即把整块田地划分为规则的网格（三角形或矩形），一个单元格区域记为 D，平均降水量和方差分别为

$$\overline{P}(\mathrm{D})=\frac{1}{A(\mathrm{D})}\iint_{\mathrm{D}}P(x,\ y)\,\mathrm{d}x\mathrm{d}y,$$

$$\mathrm{Var}(P,\ \mathrm{D})=\frac{1}{A(\mathrm{D})}\iint_{\mathrm{D}}(P(x,\ y)-\overline{P}(\mathrm{D}))^2\,\mathrm{d}x\mathrm{d}y,$$

其中 $\overline{P}(\mathrm{D})$ 表示区域 D 上的平均降水量，$\mathrm{Var}(P,\ \mathrm{D})$ 表示不同区域 D 上降水量的方差，$P(x,\ y)$ 表示 $(x,\ y)$ 处的总降水量，$A(\mathrm{D})$ 表示区域 D 的面积.

显然，方差 $\mathrm{Var}(P,\ \mathrm{D})$ 越小，说明喷洒的均匀度就越高. 进一步还可以计算出区域 D 上的最大降水量 $\max(P,\ \mathrm{D})$ 和最小降水量 $\min(P,\ \mathrm{D})$，用二者的差（即不同的区域 D 上降水量的极差）来定义方差，其大小也可以说明喷洒的均匀性.

在实际中，应该如何设置喷嘴的布局？两次移动管组的最佳距离应该是多少？我们从两种布局分别考虑.

当喷嘴之间形成矩形布局时，我们将其在 0.8 ~ 1.2 倍的喷射半径范围内进行连续变化，分别计算其降水量的方差变化情况. 一般来说，方差随着距离的增加先是减少，后是增加，距离变化到喷射半径的 1.1 倍时，其方差达到最小值. 即当喷嘴之间呈矩形布局时，管组横向移动的距离取 1.1 倍喷射半径的效果最佳.

当喷嘴之间形成三角形布局时，类似地，可以得到不同管组之间距离为喷射半径的 0.85 倍时，降水量的方差达到最小. 即当喷嘴之间呈三角形布局时，管组横向移动的距离取 0.85 倍喷射半径的效果最佳. 这为我们设置管组的作

业方案提供了基础.

4.5.2 喷灌作业方案的确定模型

考虑到这块田地形状是规则的，因此在设置主管道时应充分利用其对称性. 这里针对工人的不同特点，分别给出三种不同的喷灌模型，即懒惰型管组配置模型、创新型管组配置模型和保守型管组配置模型.

1. 懒惰型管组配置模型

懒惰型管组配置模型的主要特性是，在喷灌作业中移动管组的次数越少越好. 然而，由于支管的长度为 20 m，喷嘴的喷射半径为 19.5 m，所以对于 80 m×30 m 的田地来讲，只安装一次是无法实现整块田地喷灌的.

如果支管上只安装一个喷嘴，将其分别安放在距离田地左下角（20，15）和（60，15）的位置上，那么喷嘴到四个角的距离均为 25 m，这样也会出现很多喷不到水的地方. 因此，要想使移动管组的次数最少，支管上至少需要安装两个喷嘴，并且移动管组的次数不少于 2 次. 因此，下面只针对在支管上安装两个喷嘴的情况，结合喷嘴之间的放置距离，给出相应的管组配置方案. 图 4-24 给出了管组移动两次时支管与喷嘴的放置方案示意图.

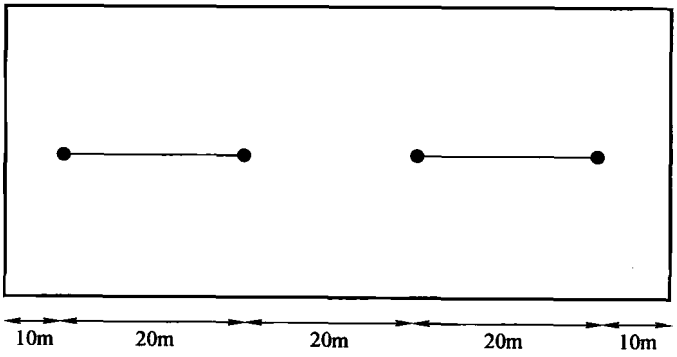


图 4-24 懒惰型管组配置方案示意图

按照这种配置方案，距离喷嘴最远的地方为 18.02 m，计算相应的降水量，最少的部分能达到 1.68 mm/h（实际上一个小时内只喷灌作业 45 min）. 要想在 4 天内达到田地需要的降水量 2 cm，只需要在一个地方连续作业 13 h，那么每点的降水量最低可以达到 2.18 cm. 即使有不超过 5% 由蒸发等因素带来的损耗，实际也是可以达到问题要求的.

该模型在 4 天内只需要移动管组 2 次，每次连续喷灌作业 13 h 即可，共需要 26 h，总用水量为$2\times26(\mathrm{h})\times45(\mathrm{min/h})\times48.35(\mathrm{L/min})=113\ 139(\mathrm{L})$，降水量的方差为 1.7×10^{-6}，其均匀度还是比较高的. 并且喷洒到田地外的水量也是较少的. 虽然移动管组的次数少了，但每个小时却需要开与关一次阀门；抗风性能差，一旦有风的影响，邻近边界的田地就有可能达不到喷水量的

要求.

2. 创新型管组配置模型

在这个模型中，对每一种可能的配置方案，都要计算出田地的边角与最近的喷嘴之间的最大距离，记为 d_{max}，要寻求一个最佳的方案使 d_{max} 尽可能的小. 实际中，这个距离越小，其抗风性能就会越好. 如图 4-25 所示，给出了一种具体的配置方案，共需要移动管组三次.

在这个配置方案中，$d_{max} \geqslant 16.5$ m，抗风性能改变不大，可以计算出需要喷灌时间达 39 h，总降水量为 170 340.3 (L)，水的利用率相对不高，喷洒出区域之外的水较多. 这种不符合常规的方案并不是一种好的方案，不建议使用.

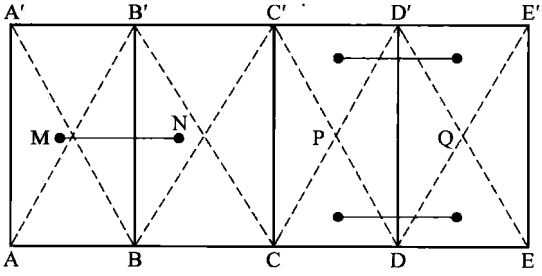


图 4-25　三次移动管组的配置方案示意图

基于上面方案的不足，可以考虑改变管组移动的次数，譬如增加到 4 次，此时可行的方案会有很多种，但不可能用枚举法来寻找最优的方案. 然而，可以从实际出发来寻求一个更加均匀、稳定的配置方案. 尽可能让喷嘴靠近田地的四个边界角点 A，A′，E，E′. 为了保证喷灌效果的均匀性，在设置管组方案时，充分考虑对称性，为了减少过度喷灌的情况，取喷嘴呈三角形的设置布局，如图 4-26 所示，三角形的边长均为 20 m.

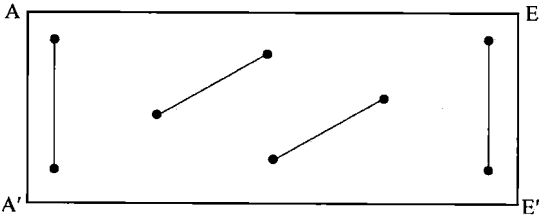


图 4-26　四次移动管组的配置方案示意图

在这个方案中，田地的边角与最近喷嘴之间的最大距离 d_{max} 为 5 m，此时为了能使不稳定性达到最小化，左右两边的管组放置的距离不超过 14 m. 这意味着只要能保证风的影响不会造成 25% 的误差，这个方案就是有效的，其降水量的方差为 3.35×10^{-6}，喷灌时间共需要 52 小时.

3. 保守型管组配置模型

为了减小喷灌的非均匀性，喷嘴放置方式如图 4-27 所示. 在这种情况

下，田地的边角与最近喷嘴之间的最大距离 $d_{max} \leqslant 13$ m，两个管组的中间部分均可以被4个喷嘴覆盖，因此无论风向朝哪个方向都可以被喷灌到. 只要风向使喷嘴喷洒改变的覆盖范围不超过6 m，就能达到良好的喷灌效果. 这种布局的标准差为 4.17×10^{-6}，其喷灌所需总时间为52小时.

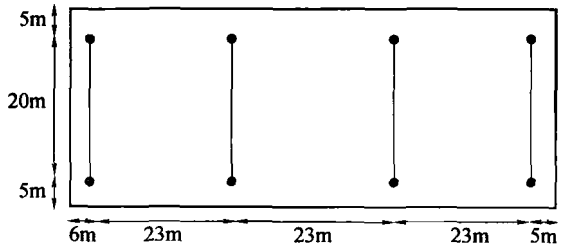


图4-27 保守型管组配置方案

在这里，我们对移动4次管组的配置方案进行了模拟检验，结果表明具有很好的稳定性. 对于多于4次的移动方案，由于重复浇灌的面积增加，可能会导致高度的不均匀性. 另外，移动管组也需要一定的时间，而且降水的标准差也会因为距离太短而增加，因此，不考虑移动管组多于4次的配置方案.

4.5.3 模型的分析与评价

将田地分割成20 m×30 m的矩形，每个矩形再分成两个相等的三角形，假设将整块田地放置在以（0，0），（80，0），（0，30）和（80，30）为端点的坐标系中. 对上述三种配置方案分别讨论移动次数、均匀性、稳定性、操作时间和喷水量等指标，计算结果如表4-7所示.

表4-7 三种方案的比较分析

方案	移动次数	均匀性	稳定性	操作时间	喷水量
懒惰型	2	高	低	26 h	113 000 L
创新型	4	中上	中上	52 h	226 000 L
保守型	4	中	高	52 h	226 000 L

由上述结果表明，懒惰型管组配置方案是实用的，整个系统的操作时间少，移动次数也最少，而且具有较好的均匀性和最少的用水量.

保守型管组配置方案稳定性较高，这对那些可能受到外界因素影响而无法充分喷灌的田地很实用. 但操作时间和用水量都是懒惰型方案的2倍之多，并且有些地方可能被多次喷洒，均匀性相对较差.

创新型管组配置方案的均匀性和稳定性都介于上述两个方案之间，用时和用水量与保守型方案相一致. 但该方案更有创意，管组呈三角形放置，适用于不喜欢循规蹈矩工作的人.

针对这个问题，我们提出了三种管组的配置方案，其中保守型和懒惰型两种配置方案的实用性更强. 保守型配置方案适用于有风的天气情况或是田地不太平坦的地块. 一般情况下，使用懒惰型的配置方案更好，因为它操作时间短、方便，且均匀性较好.

管组的配置模型和求解结果都是建立在现有喷嘴实验数据的假设之上来讨论的，在实际中可能会有一定的局限性. 但是，建模方法和评价指标都具有普遍意义，这些分析方法和评价指标可以推广到其他的喷灌系统中去.

4.6 问题的综合分析与进一步研究

4.6.1 问题的综合分析

从以上介绍的几个模型容易看出，虽然所采用的建模方法有所不同，但解决问题的思想和方法步骤基本上是一致的.

1. 问题的总体研究思路

综合分析上面的几个模型，问题的总体研究思路和基本步骤可归纳为以下三个方面：

（1）研究一个喷嘴的喷灌情况. 对于一个喷嘴的情况，主要是要确定喷嘴的喷射半径.

模型1根据市场上现有喷嘴的型号与相关参数，依据通常的水源压力，分析得到喷嘴的喷射半径以及喷水量与喷射半径的关系图.

模型2利用了经验公式，喷嘴的喷射半径通过喷嘴内径、水源水压和喷射角的关系来确定. 实际中可以近似地认为降水量与喷射半径无关.

模型3和模型4都是利用伯努利方程和能量守恒定律来分析计算喷嘴的喷射速度，从而确定喷嘴的喷射范围，即喷射半径. 四个模型都近似地视可旋转喷嘴的喷洒范围为圆型区域，其中模型3考虑了圆环型区域的情形. 对于喷洒范围为圆环型区域情形，设计更加复杂，移动管组的频率太高，增加了工作量.

（2）研究在一根支管上安装喷嘴的个数和管组设置的初步方案.

模型1是在每根支管上安装一个喷嘴，每天只需要移动两次管组，也就是一次连续喷灌就让田地达到所需要的喷水量. 虽然这样喷灌操作简单，省工省时，但是不利于田地的渗透吸收.

模型2是在10 m长的支管上安装两个喷嘴，两个管组交叉喷灌，喷灌操作的工作量增加了，但喷灌的均匀度也增加了.

模型3分别分析计算出了3个和4个喷嘴的管组移动次数和喷灌的均匀度，通过比较分析，在20 m长的支管上安装3个喷嘴较为合理.

模型4通过综合分析，确定在20 m长的支管两端各安装一个喷嘴. 而且，不同管组的移动方案，喷灌的效果不同.

（3）利用计算机模拟来调整管组的配置方案. 上述几个模型都是以喷灌的均匀度最高为最优目标，确定喷嘴的安装位置和管组的移动次序与次数，从而得到最优的喷灌方案. 模型都给出了具体的喷灌实施方案，并且讨论了管道内的水压损失对模型的影响. 同时利用不同的算法对所提供的方案进行了模拟检验，通过适当的调整得到最佳的喷灌实施方案. 但是，对于管组移动位置可能偏差的灵敏度分析不够，没有做相应的分析.

从建模的角度，模型都应该有确定的管组设置、移动喷灌方案，但在实际操作时只能是近似的，不可能达到很高的操作精度. 因此，管组安装位置能够精确到1 m左右就可以了，无需精确到厘米. 由此，在利用计算机模拟确定最佳位置时，一定要考虑管组移动位置的偏差对结果影响的灵敏度分析，这是判断模型可靠性的重要指标.

2. 竞赛论文的评价分析

农田喷灌系统在实际的生产灌溉中被广泛地应用. 针对这个竞赛题目所提出的问题，在设计喷灌计划时，应该要考虑喷灌节约用水问题，即在整块田地满足最低喷灌要求的前提下，保证喷灌不能过度，而且管组的组合方式和移动要具有相对的稳定性.

（1）评委重点关注的问题. 事实上在这个问题的建模过程中，所用的方法和参数指标都可以从实际中获取，如反映喷灌均匀性的指标、喷嘴的喷射半径和喷洒的分布特征等都是容易获取的. 因此，评委更为关注的是对问题的理解、处理方法以及假设的合理性等.

特别是以下几个方面的问题：对问题中确定的限制条件的使用情况，如降水量的要求；对问题中主观的限制条件的使用情况，如最优性具体含义的解释；对问题本身产生的限制条件的使用情况，如单个喷嘴的喷灌半径、喷射角和水压等；判断评价结果优劣的标准；求解模型最优解的方法、步骤和程序等；论文的写作是否简练、准确、层次分明等.

（2）竞赛论文中存在的具体问题. 首先，一个普遍的问题就是对最优设置方案中的“最优”的理解问题. 题目要求设计管组和喷嘴的配置喷灌方案，即在满足农田灌溉要求的前提下，应使得农场主管理、操作和维护系统所需要的时间最少. 也就是说，以管组的移动时间或移动距离最少为目标，或者其他类似的优化目标都应是合理的. 如果把实际喷水时间最少作为优化目标，则似乎就没有什么意义了，而且这个问题似乎与费用也没有多大关系.

其次，由于题目对于有些细节没有详细给出，如主管道和水源出口的位置、喷嘴的具体型号等，对于这些细节问题，只要假设合理即可，不必过分关注. 评委更看重的是如何解决本问题的难点，即喷嘴的个数设置和管组的移动

方案. 无论是理想模型，还是简单模型（喷嘴的喷灌区域为均匀的圆域）都是可以的. 但在实际中，喷嘴的喷水分布、风向、风力及其他因素对喷灌效果的影响是不可忽略的，优秀的竞赛论文应该讨论这些因素对模型的影响，这也是能否获奖的一个重要方面.

（3）有的参赛论文对问题的理解不够准确，主要是混淆了两个概念. 首先，对于每小时的喷水量不得超过 0.75 cm 的限制条件，将其当做瞬时降水速度，错误地将其等价于 0.0125 cm/min，当喷灌时间在 1 h 内，且不少于 30 min时，就用降水速度不超过 0.025 cm/min 作为限制条件，这显然不是问题本身的意思.

其次，当在一根支管上设置多于一个喷嘴时，应注意喷嘴喷水的速度和喷嘴处水压的变化情况. 有的参赛论文没有考虑这些情况，是不够完善的.

（4）优秀的论文，尤其是获特等奖的论文一般都具有以下的特点：论文的层次结构分明、建模思路清晰、分析准确翔实、模型结果详细、配置方案可操作；利用数值计算得到管组移动的计划表，求解方法可以采用模拟退火算法、遗传算法、逐步搜索法等.

4.6.2 进一步研究的问题

农田灌溉是保证农作物生长的必要条件. 手动的喷灌系统对于小型农场、花园和草坪是很有优势的. 对于这个实际的问题，还有以下问题值得进行研究.

1. 喷嘴的喷水分布规律

要想设计更加省时省力高效的喷灌方案，喷嘴的喷水分布规律是必须要已知的，因为这直接影响到喷灌的效率和喷灌的均匀性效果. 在上述几个模型中，采用不同的喷水分布，而且根据不同的喷水分布设计出了相应的喷灌方案. 事实上，这些都是在比较理想的假设下来实现的，因为实际中喷嘴的喷水分布，除了与喷嘴的物理结构参数有关以外，还与要喷灌田地的物理环境和自然条件有一定的关系，这些关系往往是复杂的，在每一个具体的问题中都是值得研究的. 最直接的方法是在不同的物理环境和自然条件下，通过实验检测喷嘴的喷水量与喷射距离的关系，然后分析比较找出物理环境和自然条件对喷嘴喷水分布规律的影响关系，进一步通过修改不同的参数来调整模型和喷灌方案.

2. 风对喷灌方案的影响

喷灌作业总是在室外的自然环境条件下进行的，风向和风力的影响不可避免. 理想的模型可以不考虑风的影响，但在实际实施中不可能达到预期的效果. 因此在设计管组的配置方案和移动方案时，都应该考虑风的影响，在模型中引入风向和风力的参数，从而得到更一般的喷水分布规律，根据不同的风向

和风力大小确定管组的配置方案，以保证最佳的喷灌效果. 这是一个很需要进一步研究的问题.

3. 水的效用问题

虽然在上面的几个模型中，多数都考虑到水的效用问题，但从主管道流出的水，除了正常的管道接头损失和在喷洒过程中蒸发造成的损失外，还应考虑从管道中流出的水究竟有多少浇到了需要喷灌的田地里. 实际中，需要综合考虑这些可能影响喷灌效果的因素，合理地定义水的效用，譬如可以用整个田地接收到的喷水量占喷嘴喷出水的比例来刻画. 水的效用越高，喷洒到田地外的水量就越少，则表明所建立的模型和喷灌方案的设计就越合理，喷灌效率也就越高.

4. 喷灌模型和方案的推广应用问题

这个问题是针对相对规则的一块矩形田地提出的，从某种意义上来讲，建立模型的方法和求解算法都可以直接推广到更一般的田地、花园或草坪的情况. 对于不规则喷灌区域，喷嘴的设置和管组的移动方案都是值得研究的问题，同时需要进一步研究喷灌效率和水的效用等问题，这些都是很有实际意义和理论价值的应用问题.

参考文献

[1] COMAP, Inc. Sprinkler Systems for Irrigation [J]. The UMAP Journal, 2006, 27 (3): 237-332.

[2] Rain Bird Agri-Products Co. Sprinkler and micro irrigation engineering formulas-U. S. units [OL]. http://www.rainbird.com/pdf/ag/FormulasUS.pdf,2001.

[3] Zoldoske, David F, Kenneth H Solomon. Coefficient of uniformity - What it tells us. Irrigation Notes (Center for Irrigation Technology, California State University, Fresno, CA) [OL]. http://cati.csufresno.edu/cit/rese/88/880106/,1988.

[4] Kershner R. *The number of circles covering a set* [J]. *American Journal of Mathematics*, 1939, 61: 665-671.

[5] Spall James C. Simultaneous perturbation stochastic approximation: SPSA, A method for system optimization [OL]. http://www.jhuapl.edu/SPSA/,2006.

[6] Zhao J C. Technology of irrigation [J]. Agricultural Science and Information, 1999, 17: 59-61.

[7] Mateos L. Assessing whole-field uniformity of stationary sprinkler irrigation systems [J]. Irrigation Science, 1998, 18: 73-81.

[8] Wilcox J C, Swailes G E. Uniformity of water distribution by some nder-tree orchard sprinklers [J]. Journal of Scientific Agriculture, 1947, 27 (11): 565-583.

[9] Walski Thomas, et al. *Computer Applications inHydraulic Engineering* [M]. *6th* ed. Watertown, CT: Bentley Institute Press, 2004.

[10] Finnemore John E, Joseph B Franzini. Fluid Mechanics with Engineering Applications [M]. New York: McGraw-Hill, 2002.

[11] National Resources Conservation Service. National Engineering Handbook. Part 652: Irrigation Guide [OL]. ftp://ftp.wcc.nrcs.usda.gov/downloads/irrigation/neh652/ch6.pdf, 2006.

[12] Louie Michael J, John S Selker. Sprinkler head maintenance effects on water application uniformity [J]. Journal of Irrigation and Drainage Engineering, 2000, 126 (3): 142-148.

[13] Rain Bird Agricultural Products. n.d. Performance data: 70H1″ (25mm) full circle, brass impact sprinkler [OL]. http://www.rainbird.com/pdf/ag/imptmtrc70h.pdf,2006.

第5章　机场轮椅配置问题

机场轮椅配置问题是2006年美国大学生数学建模竞赛（MCM）的B题[1]，研究的是社会公共事业管理领域的资源优化配置与科学运营方面的实际问题. 该问题看似一个不引人注意的简单问题，但事实上是一个非常有实际意义的问题，也是一个非常开放的竞赛题，充分反映出MCM赛题代表性的特点，很值得我们研究和思考.

5.1　问题的综述

5.1.1　问题的提出

乘飞机旅行，一个令人头疼的事情就是中途需要在多个机场转机，而且每到一个机场通常都需要转乘另外一架飞机. 对于行动有困难的旅客，从一个候机区走到另外一个候机区就更是困难了. 解决办法之一是航空公司为需要帮助的旅客提供轮椅和服务人员，帮助这些旅客快速方便地中转. 通常情况下，航空公司都能预先知道哪些乘客需要帮助. 但也有些旅客到机场登记时才提出需要帮助的请求. 也有少量情形，直到飞机即将降落，航空公司才收到旅客需要帮助的请求.

航空公司面临着降低运行成本的压力. 机场所配置的轮椅价格昂贵，时间长了会磨损，并且还需要日常的管理和维护，同时，还需要提供服务陪同人员，这些都需要费用. 另外，航空公司为了给需要帮助的旅客在航班到达时提供及时的帮助，轮椅和陪同人员需要一直在机场内巡回. 在一些大的机场，巡回时间也是不容忽视的. 轮椅在不用的时候需要存储，机场候机大厅场租昂贵，而且地方非常有限. 还有，若把轮椅留在客流繁忙的通道上，旅客经过时必须得绕过轮椅，这会妨碍行人通行. 最后，如果某位旅客因为等候服务陪同人员而耽误了时间，从而导致飞机因等待这位乘客而使航班延误，将使航空公司付出巨大的代价. 这对航空公司来说是不希望出现的，因为它将影响航空公司的航班平均延误时间，有些乘客会因此而不选择这个公司的航班，造成该航

空公司机票销售量的减少，带来航空公司的经济损失.

Epsilon 航空公司决定请求第三方为轮椅及陪同人员的调度管理、维护中的各种问题以及经济成本进行详细的分析. 航空公司希望得到每一天节省成本的轮椅调度方法，制定出短期及长期的成本预算规划.

Epsilon 航空公司要求参赛小组通过分析形成一个投标书，以帮助航空公司解决问题. 投标书应该包括对实际情况的概述和分析，以便航空公司能够确定你们是否完全理解了问题. 需要提供可执行算法的详细叙述，该算法能够确定轮椅和陪同人员应该安置在哪里，以及每天应该怎样移动. 目标是使总成本尽可能低. 参赛小组的投标书将是 Epsilon 航空公司所考虑的许多投标书之一，你们必须提供一个强有力的案例，以说明为什么你们的解决方案是最佳的，并且能够处理不同环境下的各种机场的问题.

投标书还应该包括算法如何有效处理大型（至少 4 个候机大厅）、中型（至少 2 个候机大厅）和小型（1 个候机大厅）机场在客流高峰与低谷时段的示例. 你们应该确定所有的成本，并权衡它们的权重. 最后，因为老年旅客在旅客总数中所占的比例将会越来越多，因为他们有较多的时间外出旅行，可能提出帮助的需求也会越来越多，所以你们的报告还应该包括对未来潜在成本和乘客需求的规划，以及怎样满足未来需求的建议.

问题的原文[1]如下：

Wheelchair Access at Airports

One of the frustrations with air travel is the need to fly through multiple airports, and each stop generally requires each traveler to change to a different airplane. This can be especially difficult for people who are not able to easily walk to a different flight's waiting area. One of the ways that an airline can make the transition easier is to provide a wheelchair and an escort to those people who ask for help. It is generally known well in advance which passengers require help, but it is not uncommon to receive notice when a passenger first registers at the airport. In rare instances an airline may not receive notice from a passenger until just prior to landing.

Airlines are under constant pressure to keep their costs down. Wheelchairs wear out and are expensive and require maintenance. There is also a cost for making the escorts available. Moreover, wheelchairs and their escorts must be constantly moved around the airport so that they are available to people when their flight lands. In some large airports, the time required to move across the airport is nontrivial. The wheelchairs must be stored somewhere, but space is expensive and severely limited in an airport terminal. Also, wheelchairs left in high traffic areas represent a liability risk as people try to move around them.

Finally, one of the biggest costs is the cost of holding a plane if someone must wait for an escort and becomes late for their flight. The latter cost is especially troubling because it can affect the airline's average flight delay which can lead to fewer ticket sales as potential customers may choose to avoid an airline.

Epsilon Airlines has decided to ask a third party to help them obtain a detailed analysis of the issues and costs of keeping and maintaining wheelchairs and escorts available for passengers. The airline needs to find a way to schedule the movement of wheelchairs throughout each day in a cost effective way. They also need to find and define the costs for budget planning in both the short term and in the long term.

Epsilon Airlines has asked your consultant group to put together a bid to help them solve their problem. Your bid should include an overview and analysis of the situation to help them decide if you fully understand their problem. They require a detailed description of an algorithm that you would like to implement which can determine where the escorts and wheelchairs should be and how they should move throughout each day. The goal is to keep the total costs as low as possible. Your bid is one of many that the airline will consider. You must make a strong case as to why your solution is the best and show that it will be able to handle a wide range of airports under a variety of circumstances.

Your bid should also include examples of how the algorithm would work for a large (at least four concourses), a medium (at least two concourses), and a small airport (one concourse) under high and low traffic loads. You should determine all potential costs and balance their respective weights. Finally, as populations begin to include a higher percentage of older people who have more time to travel but may require more aid, your report should include projections of potential costs and needs in the future with recommendations to meet future needs.

5.1.2 问题的背景资料

在实际中，一般航空公司在各大机场都配备有航空轮椅，以满足特殊人群的乘机或转机需要，主要是方便老年人、残疾人和特殊病人的需求，这也是国际惯例. 由航空公司负责提供日常的管理、维护和服务，而不是由机场提供. 乘客在有需求时，需要提前向所乘运的航空公司提出申请，办理相关手续后航空公司就可以提供服务.

1. 轮椅的申请使用方法

乘客申请使用轮椅有三种途径：

(1) 通过互联网在线申请.

(2) 通过电话向航空公司提出预订申请.

(3) 预订或购买机票时提出使用申请，办理乘机手续时办理轮椅的使用手续.

航空公司在接到乘客的需求申请之后，根据乘客的实际需求和所乘航班，会在机场安排轮椅服务. 正常情况下，提前预约的申请都能得到满足，而且在机场内部的服务都是免费的. 当然，特殊情况下，因为航空公司在一个机场配备的轮椅数量和服务人员的数量都是有限的，如果在同一时间段内需求量过大，可能会出现不能保证及时服务，甚至使得个别乘客误机的情况.

2. 轮椅的配置数量

航空公司在每一个机场都会配备一定数量的轮椅，为什么不能多配备一些轮椅满足乘客的需求呢？主要的原因是以下几个方面：

(1) 航空轮椅的价格比较昂贵，管理维护、维修的工作量和成本也都比较高，配备轮椅越多，所需要的服务人员就越多，成本也就越高.

(2) 在机场配备轮椅，存放保管和使用都需要占用机场的候机厅资源，配备轮椅越多，占用资源就越多，费用也就越高.

(3) 一般都是特殊的乘客需要轮椅，需求量并不是很大. 如果配备过多轮椅而没有足够多的利用，那么会出现资源浪费，造成航空公司和机场无效的运行成本.

然而，如果配备的轮椅过少，那么就可能会出现不能保证满足乘客需求，甚至使乘客误机的情况. 一旦出现这种情况，就会导致乘客不满，从而影响航空公司的声誉，对航空公司会造成潜在的损失. 如果因此造成乘客误机，那么航空公司还要承担乘客的经济赔偿责任. 所以，对于一个航空公司，在一个机场应该配备多少轮椅和多少地面服务人员，既能够尽量满足乘客的需求，又不会出现资源浪费的情况，是一个值得研究的问题.

3. 航空公司配备轮椅的缘由

为什么是由航空公司而不是由机场提供轮椅服务？这主要是基于航空公司的运行机制，航空公司根据乘客的订票情况制定和调整相应的飞行计划. 航空公司掌握着乘客的具体情况，包括乘客的实际需求，而机场没有乘客的情况信息.

5.2 基于多候机厅的配置模型

这里主要研究航空公司处理联程旅客从上一个到达出口到转乘下一次航班的规程. 根据美国运输部门的规定，航空公司地面服务人员有责任帮助需要帮助的乘客从上一个到达口到下一个航班的登机口. 我们建立了多候机厅配备轮椅的数学模型，模拟多候机厅之间的乘客、轮椅、机场中的服务工作人员间相

互协调作用的关系. 跟踪了每名乘客转机过程与可能的航班延误情况. 需要轮椅服务的乘客以三种方式中的一种在机场中运动.

5.2.1 名词解释与符号说明

(1) 登机口: 用一对有序数 (i, j) 来表示第 i 个候机厅中的第 j 个登机口.

(2) 候机厅: 由若干个登机口组成的集合. 如果第 i 个候机厅包含 k_i 个登机口, 则它用向量 $c_i=[(i,1),(i,2),\cdots,(i,k_i)]$ 来表示.

(3) 机场: 由若干候机厅组成的集合, 将其视为一个图, 用 A 表示. 如果一个机场共有 n 个候机厅, 即 $A=\{c_1, c_2, \cdots, c_n\}$, 有 n 个节点.

(4) 乘客: 一名在机场中移动的旅客, 与某一到达航班或离港航班有联系. 我们将需要轮椅服务的乘客 (wheelchair passenger, WP) 与不需要轮椅服务的乘客 (non-wheelchair passenger, NWP) 区分开来.

(5) 轮椅存放站: 轮椅在不使用时的存储地点. 将所有的候机厅视为一个相互关联的运输系统, 在这个运输系统中, 假设每个候机厅都设有一个轮椅存放站.

(6) 陪同人员: 航空公司配备的雇员, 他们的工作是到登机口迎接 WPs, 并将其送往相应的离港登机口.

(7) 误机: 如果一位乘客在航班离港前 15 分钟还未到达指定登机口, 航班将不会再等待, 该乘客就会错过该航班, 即误机.

(8) 策略: 表示陪同人员与轮椅在机场中的移动方案.

5.2.2 问题的分析与假设

(1) 机场的布局

假设一个机场包括 1 到 10 个候机厅, 每个候机厅包括 2 到 50 个登机口. 一般同一个候机厅中的登机口是相邻排列的, 而不同候机厅之间的距离则比较长. 因此, 假设乘客在候机厅之间移动的时间比在一个候机厅之内移动的时间长得多.

(2) 需要轮椅陪同服务的乘客

假设需要轮椅且需要陪同服务的乘客占乘客总人数的比例为 ω. 美国某机场 1996 年的统计数据为 1.5%, 将其作为仿真的初始数据, 到 2006 年值为 1.6%[3].

在所有 WPs 中, 有比例为 P_{inf} 的乘客在到达机场之前向航空公司提出了需求申请, 在此假设 $P_{\text{inf}}=95\%$. 由 $P_{\text{inf}}=0.95$ 及 $\omega=0.015$, 用二项分布来描述一个航班中 WP 数量及可能的情况如表 5-1 所示.

表 5-1 在一个 120 人的航班中，WP 没有预约申请的情况

没有预约申请的人数	可能性（%）
0	91.39
1	8.23
2	0.37
3	0.01

在实际中，有两种可能的情况使一名 WP 错过转接的航班. 第一种是进港的航班晚点，据统计大约有三分之一的航班会晚点，大约有 5% 的航班会晚点一个小时以上；第二种是航空公司服务陪同人员（带轮椅）的迟到造成 WP 误机. 当然，实际要尽量避免这种风险.

（3）在登机口之间的移动

一般人快步行走的速度大约是 250 英尺/分钟（即 3 英里/小时），但是当人的手臂被固定（如推着一辆轮椅行走）时，行走速度只有 180 英尺/分钟（即 2 英里/小时）[4]. 假设服务陪同人员就是以这个速度在机场行走.

通常一名陪同人员一次只能操作一辆轮椅，陪同人员需将每一位 WP 送到转接航班上，并且一直等到飞机起飞. 陪同人员的报酬为 11.80 美元/小时，按年计算，每名员工每年的收入为 $p_E = 40\ 000$ 美元[5]. 轮椅的购买成本价格是 135 美元/辆[6].

（4）市场竞争力和航班延误

对于航空公司来讲，为 WP 提供轮椅陪同服务，不仅是一种法律责任，也是从顾客服务角度增强航空公司市场竞争力的好办法.

如果由于航空公司轮椅配备方案不合理，致使 WP 错过其转接航班，这将对航空公司信誉和服务产生很不好的影响.

5.2.3 多候机厅模型的建立

1. 一般模型

如果机场 A 共有 n 个候机厅 c_1，c_2，…，c_n，并且有若干登机口 $[(1, k_1), (2, k_2), \cdots, (n, k_n)]$，其中 k_i 表示候机厅 $c_i(i=1, 2, \cdots, n)$ 中登机口的个数. 另外，E_0 表示服务陪同人员与 WP 总数的比，W_0 表示轮椅数与 WP 总数的比. 服务陪同人员与轮椅的运动方式按照某种策略 χ 进行.

在实际机场系统中，以下因素是我们在模型中不能控制的：

（1）乘客中 WP 的数量所占的比例 ω；

（2）客流高峰与低谷的日期；

（3）每年的客流总量；

(4) 轮椅的成本与服务陪同人员的薪金.

对于我们的模型，这些都是不可控的外部因素，所以在这里不考虑，只考虑可控变量χ，W_0和E_0.

我们的目标是通过减少直接成本（包括航空公司陪同人员的薪金及轮椅成本）和隐性成本（主要是航班延误造成的客户流失）来达到航空公司的最小运行成本.

多候机厅模型（Multi-Concourse Airport Model，MCAM）包括与直接成本和总的航班延误有关的3个可控变量. 航班延误损失包括到达航班的晚点与错过离港航班造成客户流失所造成的损失两个部分. 用C表示航空公司每天的运行成本费用，用D表示因航班延误带来的损失. 我们试图找到一种函数f来描述如下变量之间的关系：

$$f(\chi, W_0, E_0) \mapsto (C, D).$$

这就是一般的MCAM模型，实际中难以得到其精确的形式. 这里应用蒙特卡洛算法对机场几天运行情况进行了仿真运算. 通过仿真能够预期每天的可能延误损失和运行成本，仿真结果能够较好地近似代替函数f.

2. 总的延误时间与无用功效

每一个航空公司都有相应的规定，如果有乘客没有到达登机口，则飞机会等候一段时间x^*分钟. 如果通过增加x^*的值来减少乘客误机情况，则会增加其他乘客的等待时间；同样的，如果减少x^*的值可能会造成较多乘客误机，也会造成相应的损失. 航空公司都在努力优化选择合适的x^*值，以平衡乘客因为等待时间过长的抱怨与晚到乘客误机的可能性.

我们可以认为无效时间与不希望出现的延误时间成线性关系. 这样，一个载客量为120人的航班等待了15分钟，那么总公共效应损失正比于$120\times15=1\ 800$分钟的延误. 同样的，设航班计划离港时间为$t=0$，如果有一名乘客在$t=0$时刻没有到达登机口，但是我们知道他正在前来的路上，那么就应等候他x^*分钟. 令因有人误机而引起的公共效应损失为$-U_m<0$.

设L为晚到乘客的晚到时间，$P(L\leqslant x^* \mid L\geqslant 0)$为一名乘客在$t=0$之后能到达登机口的概率. 它表明晚到乘客的获益量是$U_m P(L\leqslant x^* \mid L\geqslant 0)$，其他乘客平均需要等候的时间为$E[T \mid 0\leqslant T\leqslant x^*]$，并且他们等候的时间至少是15分钟.

于是，我们可以选择最优的x^*值，使得等待乘客的公共效应损失量等于晚到乘客的获益量. 所以，当有N名乘客在飞机上等待的时候，最佳的选择是：

$$NE[T \mid 0\leqslant T\leqslant x^*] = U_m P(L\leqslant x^* \mid L\geqslant 0).$$

根据经验，假设取$x^*=15$，因而有

$$U_m=\frac{NE[T\mid 0\leqslant T\leqslant 15]}{P(L\leqslant 15\mid L\geqslant 0)}.$$

对于一个机场，如果统计出了 $P(L\leqslant x^*\mid L\geqslant 0)$ 和 $E[T\mid 0\leqslant T\leqslant x^*]$ 值，那么这个机场就可以安排配备恰当的服务陪同人员和轮椅，使得每一个 WP 在下飞机后都能立刻被送往转乘航班的登机口. 对于三个机场的统计情况如表 5-2 所示.

表 5-2 因晚到乘客等候 15 分钟的情况

机场	延误天数	$P(L\leqslant 15\mid L\geqslant 0)$（%）	$NE[T\mid 0\leqslant T\leqslant 15]$（分）
Midland	延误较少	31.5	5.0
Columbus	延误较少	23.2	69
St. Louis	延误较多	19.8	8.5
Midland	延误较多	25.9	2.5
Columbus	延误较多	26.9	6.6
St. Louis	延误较多	26.3	7.7
平均值		25.6	6.2

如果平均每架飞机能载 120 名乘客，上座率为 0.695（即机上座椅的占用率为 69.5%），所以实际乘客人数 $N=120\times 0.695$，于是有公共效应损失 $U_m=6.2N/0.256\approx 2\ 000$，即表明错失一次航班的后果等同于 2 000 次一名乘客在登机口等候 1 分钟，这应该是一个较为合理的结果.

由此可知，一架航班延误 1 分钟影响到 N 名乘客，相当于

$$\frac{N}{2000}U_m=0.0005NU_m$$

的公共效应损失，其中 U_m 表示错失一次航班所带来的无用功效.

3. 第三种无用功效

如果将 WP 使用过的轮椅闲置在登机口附近，这些轮椅将对其他乘客造成不便，会给乘客留下对航空公司服务质量不好的印象，并且对航空公司来说也有一种潜在的责任风险. 我们可以认为一个轮椅在登机口每停留 1 分钟所产生的无用功效等同于一名乘客被延误 1 分钟的 20% 无用功效，也就是相当于 $0.000\ 1U_m$ 的公共效应损失. 对航空公司来说，应该尽量避免这种无用功效的出现.

4. 服务陪同人员和轮椅的配备策略

服务陪同人员和轮椅的配备策略应该能够做出怎样跟随 WP 运动的决定. 主要包括以下几方面：

（1）服务陪同人员怎样选择需要接送的 WP？选择哪一个 WP？

(2) 服务陪同人员怎样找到一个可用的轮椅?

(3) 服务陪同人员将 WP 护送到指定地点后怎么做?

(4) 服务陪同人员在使用完轮椅后把轮椅放到哪里?

MCAM 模型尝试使用了三种可行策略,一个是随机策略,另两个属于智能策略.

(1) 随机策略

当一名 WP 到达机场时,在空闲的服务陪同人员中随机地选择一名将其送往转接航班的登机口. 陪同人员与轮椅一直在所到的登机口处等候直到下一次被呼叫,再去为下一个 WP 服务.

对于大多数的 WP 来说,在他们到达机场前,航空公司都会提前知道相关信息. 每一名服务陪同人员与每个登机口的代理(航空公司在登机口工作的代表)都有一张已提出服务申请的乘客到达顺序列表. 等候陪同者将其送往转接航班登机口的 WP 排序的算式为

$$H=\text{航班离港的时间}-\text{乘轮椅到达指定登机口的时间}.$$

如果一位 WP 提出了申请,陪同人员就会提前在相应的登机口等候. 在模型中,对于提前预约的 WP 会在他们的航班到达前 20 分钟被列入等待的队列. 当一名没有提前预约的 WP 出现时,登机口的代理人员会通过无线电立刻通知航空公司的地面工作人员,以便每名工作人员都可以及时更新他们手中的 WP 名单.

当一名陪同人员完成陪护工作之后空闲时,需立即向组织部门报告. 在 WP 名单中,排在最前面的 WP 会被通知给相距最近且空闲的服务陪同人员去陪护,这位服务陪同人员会通过无线电找到距离 WP 最近的或在接 WP 路上能够找到的可用轮椅.

(2) 直接转移策略

直接转移策略是基于登机口之间的运行策略,服务陪同人员的所有工作都是在登机口之间进行的. 服务陪同人员将一名 WP 送到指定登机口 (i, j) 后,服务陪同人员和轮椅都留在那个登机口,直到被呼叫去服务下一名 WP. 下一次服务也就是到另一个登机口 (i', j') 接送另一名 WP,服务陪同人员都是通过无线电联络,并在距登机口 (i', j') 最近的地方找到可用轮椅. 这种策略是将轮椅分散在各登机口之间,以便于每个登机口附近都有可用的轮椅. 这种策略的不足之处是在各登机口附近都有闲置的轮椅,会给在候机厅的乘客通行造成不便,为机场的管理和安全也会带来一定的影响.

(3) 周转式运输策略

在每一个候机厅都设一个轮椅存放点. 当服务陪同人员将一名 WP 送到登机口 (i, j) 之后,将轮椅返还到候机厅 i 的轮椅存放点. 这种策略避免了把轮椅留在登机口可能引发的交通拥堵和潜在的安全问题,并且服务陪同人员知

道所有可用轮椅都在轮椅存放处，也避免了服务陪同人员在需要时为了找到可用轮椅而花费时间.

5.2.4 模型的求解实现

利用 MCAM 模型可以仿真机场一天的运行情况，虽然能够反映出实际的情况，但是还应该考虑长远的影响因素，如 WP 占乘客的比例、服务陪同人员的薪金、轮椅的成本、存放轮椅的费用和航空公司经营成本等因素，在未来的时间里都可能发生变化. 模型 MCAM 是在理想状态下，利用过去这些因素的常数值，仿真了未来几年航空公司的运营情况. 而实际上 MCAM 模型中的这些常数值，在未来 10 到 40 年的时间中都可能会发生变化.

特别是 WP 的比例问题，随着人口老龄化问题的变化，这个比例会越来越大. 据统计资料，2000 年年龄在 60 岁以上的人占美国人口的六分之一[5]，而且需要乘坐轮椅的人中有 72% 都在这一个年龄段，60 岁以上的需要轮椅的可能性是 60 岁以下人的 13 倍之多[1]. 根据美国人口调查局的人口统计数据，在随后 20 年里，65 岁以上的人口数量会增长 40%[7]. 在 MCAM 模型中，假设 2006 年乘客总数的 1.6% 是 WP，由此计算出了未来的乘客中 WP 的数量.

现在不清楚的问题是：随着老年人口的增加，未来航班出行的人数是会成比例增加，还是会减少？这是一个与节约开支、节省时间和安全保障等有关的问题. 在 MCAM 模型中，用美国使用轮椅的人数比例代替了在机场需要轮椅服务的人数的比例，即选取了这个比例的中间值为 ω 的取值.

另一方面，关于航空公司的利润问题. 在 2004 和 2005 年，燃料费用高使得航空公司的收入减少. 如果燃料价格居高不下，航空公司将被迫提升机票价格或降低利润率. 应该考虑航空公司的长期经营，为此，可以通过考虑降低航空公司在每名乘客身上的利润率，当然也可以考虑提高机票价格来考虑这一问题.

5.2.5 模型的结果分析

针对一个航空公司的实际情况做了仿真，分别应用了随机策略、直接转移策略和周转式运输策略，就 5 名服务陪同人员与 5 辆、10 辆、20 辆轮椅的情况，以及 10 名服务陪同人员与 20 辆轮椅的情况做仿真计算，得到如下结论：

(1) 在只有 5 名服务陪同人员工作的情况下，在低延误期与高延误期内，对于不同配置策略所产生的无用功效是基本相同的.

(2) 当配有 10 名服务陪同人员时，高延误期所产生的无用功效要比低延误期产生的无用功效多 40. 一种解释是不论在低延误期还是在高延误期内，5 名服务陪同人员一直都在忙碌中. 在低延误期内 10 名服务陪同人员已经足以满足所有轮椅的陪护需求了，但在一个高延误期内，迟到的乘客即便是在有

10 名可用的服务陪同人员情况下，也有可能会出现误机的情况.

（3）对于随机策略与直接运输策略，所有的轮椅与服务人员的配置方案，在航班延误的情况下，运营状况是基本相同的，即两种策略没有多少区别.

（4）周转式运输策略在只有 5 名服务陪同人员的情况下，相对另外两种策略的效率要低一些，但是在有 10 名服务陪同人员时，其效率相对就高了. 周转式运输策略关系到轮椅运动的流程，所以在没有足够的服务陪同人员的情况下，不是很有效. 事实上，周转式运输策略的高效主要是建立在有一定数量服务陪同人员的基础之上.

这个模型应用了航空公司的真实数据，包括 2005 年 365 天航班延误的分布情况，根据市场占有率与利润来区分轮椅的配备策略的优劣. 从长期运营角度看，这个模型也能够适应一个航空公司 WP 发生变动的情况.

5.3 基于航空公司间竞争的配置模型

MCAM 模型模拟了服务陪同人员、轮椅和 WP 的运动情况. 但是，为了寻找一个航空公司最小运行成本的策略，需要知道航空公司基于顾客满意度的市场占有率发生变化的情况以及在这种运营策略下所得到的利润变化情况. 如果将航空公司长远的潜在损失也视为一种成本，则航空公司的利润最大化也就等同于成本的最小化.

航空公司之间的竞争模型（Airline Competition Model，ACM）要应用 MCAM 模型所得到的结果，来决定存在竞争关系的若干航空公司的市场占有率和盈利情况.

5.3.1 航空公司的市场占有率模型

影响一个航空公司长远盈利的主要因素是市场占有率. 如果一个航空公司因为可用轮椅太少而导致航班延误或乘客误机事件，WP 或者 NWP 可能会去选择其他航空公司的航班. 将这种“客源流失”看成是对航空公司客源的负效用函数.

令 m 表示航空公司的个数，$N_t^{(i)}$ 表示第 i 个航空公司在 t 时间的市场占有率，有 $f_t^{(i)}$ 的乘客在 t 时刻末转乘其他航空公司的航班. 假设选择其他航空公司的乘客会以相同的机会选择其他 $m-1$ 个航空公司中的任一个，市场占有率的变化率为

$$\Delta N_t^{(i)} = N_{t+1}^{(i)} - N_t^{(i)} = \left(\sum_{j \neq i} \frac{1}{m-1} N_t^{(j)} f_t^{(j)} \right) - N_t^{(i)} f_t^{(i)}. \tag{5.1}$$

另外，也可以给出另一个假设，对于要转乘其他航空公司航班的乘客，根据市场占有份额的多少在其他 $m-1$ 个航空公司中选择，即有

$$\Delta N_t^{(i)} = \left(\sum_{j \neq i} \frac{N_t^{(i)}}{1 - N_t^{(j)}} N_t^{(j)} f_t^{(j)} \right) - N_t^{(i)} f_t^{(i)}.$$

在没有实际乘客选择航空公司数据的情况下，假设机票价格为决定性的因素. 因为实际中机票价格的变化情况比较复杂，但是受市场占有率的影响并不大[8]. 正是因为机票价格影响的重要性，这里用（5.1）式可能更为符合实际.

如果航空市场的乘客总量以 r 的比率增长，更多的新乘客选择乘飞机出行，航空公司的市场占有率不会发生大的变化. 于是，可以用一个矩阵较好地表示这个假设. 用向量 $\boldsymbol{N}_t$ 表示航空公司的市场占有率，并定义一个 $m \times m$ 阶的矩阵 $\boldsymbol{A} = (a_{ij})_{m \times m}$ 如下：

$$a_{ij} = \begin{cases} \dfrac{1}{m-1}, & i \neq j, \\ -1, & i = j. \end{cases}$$

这样将（5.1）式可以简化为

$$\boldsymbol{N}_{t+1} = (\boldsymbol{I}_m + \boldsymbol{A}\boldsymbol{F}_t)\boldsymbol{N}_t,$$

其中 $\boldsymbol{F}_t$ 为一个 m 阶的对角矩阵，对角线上的元素为 $f_t^{(j)}(j=1, 2, \cdots, m)$，并且非对角线上的元素均为 0，$\boldsymbol{I}_m$ 为 m 阶的单位矩阵. 对上式进行迭代得到闭合表达式：

$$\boldsymbol{N}_T = \left[\prod_{t=1}^{T} (\boldsymbol{I}_m + \boldsymbol{A}\boldsymbol{F}_t) \right] \boldsymbol{N}_0.$$

其中多元随机变量 $\boldsymbol{F}_t$ 表示 WP 所占乘客的比例，因为社会上残疾人的比例在短期内变化不大，所以可以认为它的分布在短期内是一个常数. 因而易知，$\boldsymbol{N}_T$ 在短时间内基本上是常量，它是 T 个服从均匀分布值为 $\boldsymbol{I}_m + \boldsymbol{A}\boldsymbol{F}$ 的随机变量的乘积. 这里把 $\boldsymbol{F}$ 下角标略去，主要是因为在短期内 $\boldsymbol{F}$ 值与时间没有太大关系.

5.3.2 航空公司的乘客流失模型

航空公司为了追求最大的利润主要是通过降低运行成本，或增加乘客的满意度从而增加客源，并且权衡二者之间的关系. 否则，单方面追求航空公司的利润而降低成本，使得乘客的满意度下降，则就会导致乘客流失，即航空公司的市场占有率的减少.

J 表示不同类型的日期，其中每种类型的日期航班延误情况可能不同. 例如，在节假日或周末，因为客流量加大，可能会有较长时间的航班延误. 用 $f_t^{(i,j)}$ 表示第 i 个航空公司的 j 种类型的日期每天乘客的流失率，即第 i 个航空公司在当天顾客流失总量占客流总量的比率，它也是一个随机变量. 假设同一天中机场的交通状况是不变的，但是不同类型日期之间因客流量的不同，其交通

状况也会有不同. 这些交通状况的差别，潜在地会影响到乘客流失率 $f_t^{(i,j)}$ 的数值变化. 令 d_j 表示一年中第 j 种类型日期的总天数.

这里利用 MCAM 模型，在给定参数 (χ, W, E) 的情况下，模拟出顾客流失的总量. 即针对第 i 个航空公司的第 j 种类型日期的第 t 天，给定相应的参数 (χ, W, E)，就可以模拟得到该航空公司每天顾客流失率 $f_t^{(i,j)}$，则顾客的保留率为 $1-f_t^{(i,j)}$. 这样，就可以给出在天数 t 变化的情况下，随机变量 $g_t^{(i,j)}=\log(1-f_t^{(i,j)})$ 的变化情况. 其分布的均值和方差用一个数对 $(\mu(i, j), \sigma^2(i, j))$ 来表示，即有

$$g_t^{(i,j)} \sim (\mu(i, j), \sigma^2(i, j)).$$

第 i 个航空公司的第 j 种类型日期的第 t 天的保留率由 $1-f_t^{(i,j)}$ 给出，所以该航空公司一年的保留率为

$$1-f_t^{(i)}=\prod_{j=1}^{J}\prod_{k=1}^{d_j}(1-f_k^{(i,j)}).$$

其对数形式为

$$\log(1-f_t^{(i)})=\sum_{j=1}^{J}\sum_{k=1}^{d_j}g_k^{(i,j)},$$

根据中心极限定理，如果 d_j 的数值足够大的话，则有如下的近似分布：

$$\sum_{k=1}^{d_j}g_k^{(i,j)} \sim N[d_j\mu(i, j), d_j\sigma^2(i, j)],$$

其中 $N[\mu, \sigma^2]$ 为正态分布. 在模型的具体实现中，采用随机抽样来确定变量 $g_k^{(i,j)}$ 的取值，这样可以得到变量 $f_t^{(i)}$ 近似服从于如下的分布：

$$f_t^{(i)} \sim 1-\exp\Big(\sum_{j=1}^{J}N[d_j\mu(i, j), d_j\sigma^2(i, j)]\Big).$$

上述模型是在轮椅的配置策略不固定的情况下，其他所有因素都是不变的. 由此可知，如果有乘客误机，就是因为不恰当的轮椅配置策略，而不是其他原因. 于是，假设如有一名乘客误机，他就会以 $p_0=1/4$ 的概率去选择其他航空公司的航班. 如果是一名 WP 误机，他会认为是航空公司的轮椅配备不到位或服务不好，没能帮助他到达转接航班的登机口，所以，就会促使他改乘其他航空公司的航班. 为此，取这样高的概率是合理的.

在第 t 天第 i 个航空公司有 $P_tN_t^{(i)}$ 名乘客，其中 P_t 表示第 t 天航空市场的总乘客量，$N_t^{(i)}$ 表示第 t 天第 i 个航空公司的市场占有率. 乘客在误机后还不转乘其他公司航班的可能性是 $1-p_0$. 于是，在有 m 人次误机后，转乘其他公司航班的可能性是 $1-(1-p_0)^m$.

假设有乘客误机，并转乘其他公司航班产生的负效应为 D_t，这种负效应对于所有乘客的影响是均匀分布，所以每名乘客所受的负效应为 $u_t=D_t/N_t^{(i)}$，这是对错过一次航班所产生的负效应的一个度量. 从而，每一名乘客改转乘其他公

司航班的概率为 $1-(1-p_0)^{u_j}$. 根据大数定理，第 i 个航空公司在第 j 个类型日期的第 t 天，转乘其他航空公司航班的乘客比例（乘客的流失率）可近似等于随机变量的期望值. 即有

$$f_t^{(i,j)}=N_t^{(i)}[1-(1-p_0)^{u_j}].$$

5.3.3 航空公司的利润函数模型

假设航空公司每名乘客的成本年增长率为 r_C，毛利润年增长率为 r_D. 用 $\pi_t^{(i)}$ 表示第 i 个航空公司在第 t 年的实际利润（需要考虑通货膨胀的影响）. 如果单位商品的利润与成本都以常速增长，利润 $\pi_t^{(i)}$ 就是零通胀的情况. 要计算航空公司的长期利润，通货膨胀的因素应该考虑进去. 所以，第 i 个航空公司的最大期望利润要有相应的折扣率 δ，即有

$$\pi(i)=\sum_{t=0}^{T}\delta^t\pi_t^{(i)}.$$

不妨取总的时间年限 $T=40$，并且考虑2046年的情况. 现在的通货膨胀率近似为2%，可以取 $r_C=2\%$，并且 r_D 为远期没有考虑风险的收益率，假设它为常量4.5%，于是有航空公司期望利润的折扣率为

$$\delta=\frac{1}{1-r_C+r_D}=\frac{1}{1.025}\approx 0.975.$$

对于航空公司来说，在考虑轮椅的配备策略时，要充分考虑与此有关的影响利润的因素. 相关的因素有来自乘客的利润（与市场的占有率有关）、轮椅的购置与维修费用、在机场存放轮椅的费用和服务陪同人员的薪水与保险金等. 根据这些因素构造航空公司的利润函数模型：

$$\pi_t^{(i)}=\begin{cases}P_tN_t^{(i)}q^{(i)}-(p_w+p_S^{(i)})W_t^{(i)}-p_E^{(i)}E_t^{(i)}, & t=0,\\ P_tN_t^{(i)}q^{(i)}-p_wR_t^{(i)}-p_S^{(i)}W_t^{(i)}-p_E^{(i)}E_t^{(i)}, & t>0.\end{cases}$$

其中，$q^{(i)}$ 表示第 i 个航空公司平均从每名乘客获得的利润；$R_t^{(i)}$ 表示在 t 年第 i 个航空公司需要更换的轮椅数；p_w 表示轮椅的单位价格；$p_S^{(i)}$ 表示第 i 个航空公司每年存放一辆轮椅需要的费用；$p_E^{(i)}$ 表示第 i 个航空公司每年雇佣一名服务陪同人员的成本；$W_t^{(i)}$ 表示第 i 个航空公司第 t 年所需要的轮椅数；$E_t^{(i)}$ 表示第 i 个航空公司第 t 年所需要的服务陪同人员数.

5.3.4 基于利润的轮椅配备模型

航空公司最初购置轮椅的费用是不同的. 随后每年将会有大约20%的轮椅需要更新，所以在第 t 年末将只有 $0.8W_t$ 的可用轮椅. 由于航空公司市场占有率的变化，在第 $t+1$ 年中可能会需要更多或较少的轮椅，则第 i 个航空公司的需求轮椅的期望数量为

$$W_{T,t+1}^{(i)} = W_0^{(i)} \cdot P_t N_t^{(i)}.$$

类似的，第 i 个航空公司的需求服务陪同人员的期望数量为 $E_{T,t+1}^{(i)} = E_0^{(i)} \cdot P_t N_t^{(i)}$. 在实际中，即便是乘客的需求量减少了，也不会扔掉那些完好的轮椅，所以在第 t+1 年使用的轮椅数量为

$$W_{t+1}^{(i)} = \max \{0.8W_t^{(i)}, W_{T,t+1}^{(i)}\},$$

其中 $W_0^{(i)}$ 与 $E_0^{(i)}$ 分别表示第 i 个航空公司的乘客中需要轮椅和服务陪同人员的比例. 于是第 i 个航空公司第 t+1 年需要替换的轮椅数量为

$$R_{t+1}^{(i)} = \max\{0, W_{T,t+1}^{(i)} - 0.8W_t^{(i)}\}.$$

5.3.5 模型的求解实现

应用计算机对 ACM 模型进行模拟求解，主要依赖于 MCAM 模型所得结果. 这里就 m 个航空公司的情况进行了模拟计算，根据每一个航空公司的参数（χ, W_0, E_0）与起始年数都可以计算出他们的总成本与总的负效应.

从第一年开始，利用 ACM 模型模拟了第 i 个航空公司运用一个确定的轮椅策略直至第 T 年后的运营情况，并将其结果作为 MCAM 模型的输入值. 注意到，高延误日（出现较长时间的航班延误）与低延误日（出现较短时间的航班延误）的天数会对航空公司的运营产生不同的影响. 根据 2005 年（美国）西南航空公司每天航班延误统计情况的真实数据，在 3 个机场中每年都有 10% 的天数会出现高延误和 90% 的天数出现低延误[9]. 所以，将一年天数分为两类：一类是高延误日的天数最多占 10%，即每年有 35 天；另一类是低延误日的天数最低占 90%，即每年有 330 天. 由此，可以模拟计算出每年 35 天高延迟日和 330 天低延迟日的具体情况.

同时，根据西南航空公司的统计报告，平均每天的航班乘坐率为 69.5%. 因为高延误通常都是发生在出行的高峰期，可以假设此时航班都是满载的，即乘坐率为 100%，而在低延误日的乘坐率为 66%. 故在计算时高延误日与低延误日的乘坐率分别取 100% 与 66% 的加权平均值.

这样就能得到航空公司总的无用功效，进而，在第一年每个航空公司的市场占有率相等的条件下，再通过（5.1）式计算得到航空公司的市场占有率升高或降低的变化情况. 同时，也能够计算出每个航空公司在 t 年的利润，并且再应用 t 年的市场占有率计算下一年度的情况.

利用这种方法对未来 40 年的情况进行模拟计算，可以得到每个航空公司每年的利润向量. 如果未来的折扣率为 δ，则可以与现在的各种配置策略（χ, W_0, E_0）所得到的结果进行比较分析. 在这里只考虑航空公司在机场使用轮椅的有关利润，相对的折扣率只是用来判断各配置策略的优劣.

5.3.6 模型的结果分析

对 3 种轮椅运动策略与 4 种轮椅和服务陪同人员的组合策略，分别模拟计

算了航空公司每年的市场占有率以及最后的利润.

对于一个航空公司，采用周转式运输策略，有10名服务陪同人员和20辆轮椅，到2046年时最大的市场占有率为11.1%. 在同样的运输策略下，如果只有5名服务陪同人员和20辆轮椅，则到2046年时市场占有率仅为7.2%.

从航空公司的利润方面分析，10名服务陪同人员和20辆轮椅，利用周转式运输策略很有优势，使得航空公司的最大利润达到106，第二位为101.9. 有趣的是，对于同样数量的轮椅和服务陪同人员，如果采用随机运输策略，将使该航空公司虽有第三位的市场占有率，但利润倒数第二. 因此，最佳的配置策略应该是20辆轮椅和10名服务陪同人员，并采用周转式运输策略. 当然，如果采用周转式运输策略，在每个候机厅中安排两名陪同人员与两辆轮椅，会使航空公司的市场占有率与利润都达到最大.

这个模型是将乘客的无用功转换为了乘客的“流失”，航空公司在制定规划时，特别要重视这个问题，对乘客无论是长期的影响因素，还是短期的影响因素，都会影响航空公司未来对市场的占有率.

5.4 基于最小成本的配置模型

5.4.1 问题的分析与假设

建立模型的目标是：提供一个使得航空公司总体成本最小的人员与轮椅配置和存储方案，以及一个能够产生轮椅与陪护人员最优调度策略的算法.

总体思路是：建立一个实际机场的几何结构模型，根据一个确定的轮椅和服务陪同人员调度方案，来仿真需要轮椅服务乘客（WP）的抵港和离港的具体过程，计算航空公司所需要的费用. 并且，还要考虑随机到达的需要轮椅服务的乘客情况，以及可能出现的航班延误. 目的是给出每一个服务陪护人员在每一个时间段内的运动过程，即给出对WP的服务规则. 根据实际需求，通过这些规则进行比较选择，找出使得航空公司成本最小的轮椅配置和服务陪护人员的调度方案. 为了研究这个问题给出如下的假设：

（1）机场候机厅的布局是确定的，一般候机厅的个数不会太多，不妨假设有一个或两个，候机厅各登机口的位置也是确定已知的.

（2）每个候机厅内都有信息服务台，可以与服务陪护人员通过对讲机进行交流.

（3）服务陪护人员的服务范围是被限制在一个范围内，每名服务陪护人员护送一名WP从一个到达航班登机口到转机航班的登机口.

（4）每一个航班上需要轮椅的乘客数量相对于总乘客数量总是比较少的，即对某一个航班来说只需要研究个别乘客的需求，而不是所有乘客的

需求[10].

(5) 大部分需要轮椅服务的乘客航空公司都是提前预知的，但是也有少数个例是在航班到达后才获知需求信息. 当有乘客需要轮椅时，总是有服务陪护人员陪护.

(6) 当服务陪护人员空闲时都可以从事其他服务工作，比如在信息咨询处提供信息咨询等，目的是减少航空公司提供轮椅服务的费用[10].

5.4.2 单项因素的费用模型

首先考虑与轮椅服务和陪护人员的费用相关的因素，主要因素有以下几个方面：

(1) 因为没有及时为需要轮椅的乘客提供轮椅服务，让其没能按时转机，使得航班发生延误，造成航空公司的损失.

(2) 由于航空公司的原因让需要轮椅的乘客必须等待，航空公司的信誉受到影响，导致售票量的减少，收益下降.

(3) 雇佣陪护人员所需要的费用.

(4) 轮椅的购置、折旧、维修和存储费用.

(5) 分配空闲的服务陪同人员从事其他工作，减少航空公司成本，增加收益.

如果时间步长 Δt 均取为 1 分钟，下面分别考虑这些费用的计量方法.

1. 航班的延误费用

由于航班延误，对于航空公司都会产生一定的经济损失，其大小与延误的时间长短有关. 根据实际，航空公司因航班延误所产生的延误费用为

$$C_D = \sum_{t_{fd}=1}^{\infty} A(t_{fd}) c_A(t_{fd}), \tag{5.2}$$

其中 $A(t_{fd})$ 为延误的航班数，t_{fd} 为延误时间（单位为分钟）；$c_A(t_{fd})$ 为一个航班延误 t_{fd} 分钟所产生的费用（单位为美元），具体取值可以根据航空公司的实际确定.

2. WP 的等待费用

WP 因为轮椅服务不到位或不及时需要等待一定的时间，使得 WP 对航空公司的不满意程度增加，也可能影响到航班的延误，都会使航空公司产生一定的等待费用. 可以得到等待费用为

$$C_W = \sum_{t_{pw}=1}^{\infty} a(t_{pw}) c_a(t_{pw}), \tag{5.3}$$

其中，$a(t_{pw})$ 为 WP 中等待 t_{pw} 分钟的乘客数量，$c_a(t_{pw})$ 为一位 WP 等待 t_{pw} 分钟所产生的费用（美元）.

3. 服务陪护人员的总费用

航空公司为 WP 雇用的服务陪护人员需要的总费用为

$$C_E = K \cdot T \cdot \omega(t), \tag{5.4}$$

其中，$\omega(t)$ 为时刻 t 正在陪护服务的陪护人员数量，T 为每个陪护人员平均完成一次陪护任务所需要的时间（分钟），K 为每个服务陪护人员单位时间所需要的平均费用（美元/分钟）.

4. 轮椅的成本费用

将航空公司为轮椅的购置、维护、折旧和存放所需要的费用记为 W_0. 注意到 W_0 的数值取决于机场的具体情况，即与航空公司在机场所配置轮椅的数量和类型有关. 所以相对于确定的服务陪护人员数量来说是一个不变的常量，当然这与服务陪护人员多样化的工作转换方式有关.

5. 服务陪护人员兼顾其他工作所产生的效益

服务陪护人员兼顾其他工作为航空公司所产生收益的计算是比较复杂的. 根据微观经济学的相关理论，假设航空公司增加一名工作人员从事某项工作所得的收益与正在从事这项工作的人数成反比. 反映在数学上，这意味着有 n 个雇员从事某一项工作所得的收益可以近似地表示为与 $\ln(n+1)$ 成正比. 如果不安排空闲的服务陪护人员来做这项工作，航空公司就必须再雇佣一位专职的工作人员来做这一项（次要的）工作. 对于由专职雇员从事这一项工作，航空公司所得的“收益”应该等于这名雇员的工资（即所付出的成本）. 这也就意味着一名空闲的服务陪护人员一次工作转换后，航空公司的收益（即负成本）为

$$C_K = -\frac{K}{\ln 2}\sum_{t=0}^{T}\ln(\omega(t)+1). \tag{5.5}$$

5.4.3 航空公司的总费用模型

1. 航班延误与乘客等待的总费用

根据美国联邦航空管理局 2000 年统计数据，一个航班平均延误 1 分钟的费用为 44 美元，使一名乘客等待 1 分钟的平均费用近似为 0.25 美元[11]. 由于额外等待 1 分钟的边际成本与总的延误无关，于是可得

$$c_A(t_{fd}) = 44t_{fd},\quad c_a(t_{pw}) = 0.25t_{pw}.$$

$A(t_{fd})t_{fd}$ 表示所有航班延误 t_{fd} 分钟的总延迟时间的和，因此，所有航班的总延误时间为

$$T_{fd} = \sum_{t_{fd}=1}^{\infty} A(t_{fd})t_{fd}.$$

类似地，乘客总的等待时间为

$$T_{pw}=\sum_{t_{pw}=1}^{\infty}a(t_{pw})t_{pw}.$$

由（5.2）式和（5.3）式中所定义的费用的表达式，航班延误总费用和乘客等待总费用分别为

$$C_D=44T_{fd},\quad C_W=0.25T_{pw}.$$

2. 一个班次服务陪护的总费用

利用（5.2）式至（5.5）式，机场内的所有服务陪护人员完成一个班次的工作所需要的总费用为

$$c=C_D+C_W+C_E+W_0+C_K,$$

即

$$c=44T_{fd}+0.25T_{pw}+\omega KT+W_0-\frac{K}{\ln 2}\sum_{t=0}^{T}\ln(\omega(t)+1).$$

3. 航空公司的年度总费用

问题研究的目标不仅仅是需要降低航空公司每天的运行成本，还要使全年的成本尽量小. 以过去3年为例，不妨设在过去3年的1 095天里，每天都是8小时轮班工作制，即服务陪护人员每天每8小时轮班一次. 每一天不同时间段的航班流量是不同的，通常航班流量繁忙程度按时段可分为轻度、重度和中度三种情况：

（1）轻度繁忙. 从下午4点到午夜12点和从午夜12点到第二天上午8点两个班次，这两个班次至少占了一年中通航日的90%，即有329天. 占三年（1 095天）中的60%，即有657天. 保守估计在这些时间段里航班流量大约占总航班流量的一半左右.

（2）重度繁忙. 按照8小时换班的规则，重度繁忙的时段最多有10%，即大约有110天. 相关资料数据表明，在这些天里的航班流量大约为平均流量的2.5倍.

（3）中度繁忙. 不属于上述轻度和重度繁忙两种情况的时段就是中度繁忙的时段（即剩余的328天）. 相关资料数据表明，在这些天里的航班流量大约为平均流量的1.5倍.

在这里，取航空公司全年的总费用函数为全年每天不同时段费用的加权和. 以 C_l，C_m，C_h 分别表示航班流量轻度繁忙、中度繁忙和重度繁忙时段的总费用，则航空公司年度的总费用为

$$C_a=657C_l+328C_m+110C_h. \tag{5.6}$$

5.4.4 模型的求解算法

为了利用航空公司最小成本费用模型，这里设计了一种动态调配轮椅和陪护人员的算法，算法流程图如图5-1所示. 整个模拟过程可以分为6个步骤，

下面对每一步说明如下：

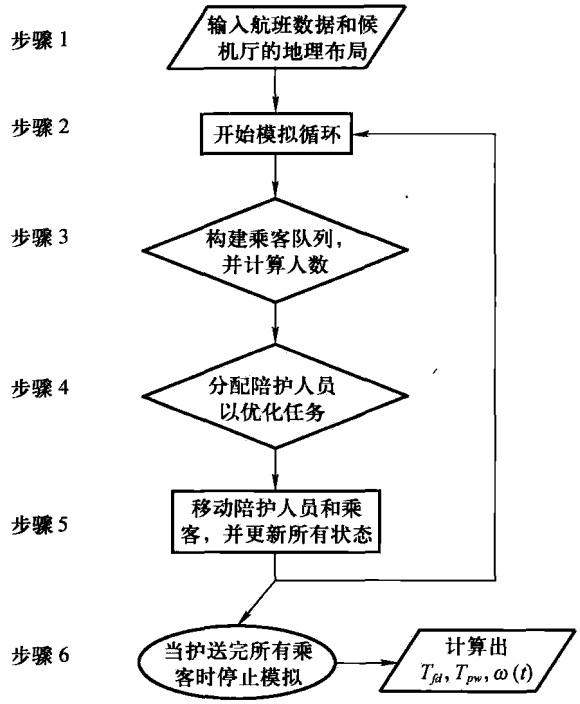


图 5-1 模型求解算法流程图

步骤 1：输入基本数据

该算法需要已知机场候机厅的布局以及 WP 所乘航班的时刻表. 候机厅的布局包括服务台的位置、每一位 WP 到达和转机的登机口的位置信息等. 如果候机厅有两层楼的话，则还需要有一部电梯. 航班时刻表包括到达航班与转机航班的时间和相应的登机口等有关信息.

步骤 2：主循环

对于每一个时间步长 Δt，确定 WP 的队列和相应的停留时间（分钟数），并制定最佳的陪护人员的分配方案.

步骤 3：建立 WP 的服务队列

首先考虑需要轮椅但没有事先预订的少部分 WP 的可能性，这部分 WP 可以视为占事先预订 WP 的一个固定的比例，在这里不妨设为 5%，即有 5% 的 WP 没有事先预订需要轮椅和陪护服务.

为每一个 WP（包括提前预订的和没有事先预订的）分配一个时间（分钟数），即到他们转乘的航班起飞前的时间. 根据分钟数的大小重新安排服务队列，优先安排分钟数小的 WP，并可插入新来的 WP，同时为其安排服务陪护人员. 这一功能将会尽可能地减少航班的总延误时间 T_{fd}，以减少航空公司的

损失.

步骤 4：优化分配服务陪护人员

这里利用海明距离 $|\Delta x|+|\Delta y|$ 来表示两点间的距离，选择距离最近、又处于空闲的服务陪护人员分配给 WP 中排在队伍最前面的乘客. 一个服务陪护人员护送完一个 WP 之后，在被安排下一个陪护任务之前，都要先回到信息服务台从事其他服务工作.

步骤 5：服务陪护人员的运动和更新状态

在为每一个 WP 安排确定了一个服务陪护人员之后，对于还没有完成护送任务的陪护人员，将前往各自陪护的目的地，在每一个时间步长 Δt 之内，每名服务陪护人员都将移动一段距离 Δd，这就确定了一个自然运动速度 $\Delta d/\Delta t$. 在这里不妨取 $\Delta t=1$ 分钟，服务陪护人员的运动速度为 2.5 ft/s，主要是考虑到候机厅内可能有一些其他障碍和候机乘客的影响，这个假设应该是合理的. 于是可以得到 $\Delta d=150$ ft.

当服务陪护人员护送一个 WP 到达转机登机口时，服务陪护人员的状态信息将相应地更新. 当一个服务陪护人员带轮椅到达一个 WP 入境登机口，而因航班延误该 WP 可能还没有到达，用 p_d 表示该航班延误的可能性（概率），实际上根据相关的统计数据有 $p_d\approx 0.29$[12]. 对于延误了的航班，其延误时间长度服从指数分布[13].

步骤 6：输出计算结果

循环仿真计算，直到已知航班时刻表上的所有 WP 都被护送到各自的转乘航班登机口为止. 一旦这个条件得到满足，循环计算终止，输出所有 WP 的等待总时间 T_{pw}. 由于个别 WP 没有及时到达转乘登机口，而导致航班延误的总时间 T_{fd}和在 t 时刻服务陪护人员的总数 $\omega(t)$.

5.4.5 模型的应用与分析

以 Dalta 航空公司在 3 个机场候机厅的数据为案例进行了应用分析，这 3 个机场分别为芝加哥奥黑尔国际机场（IL）、纽约约翰肯尼迪国际机场（NY）和亚特兰大国际机场（GA）. 这里设陪护人员的工资为 $K=3.5$ 美元/h. 这是因为有一些航空公司为陪护人员平均付 7 美元/h，但对志愿者是不付费的，取平均 $K=3.5$ 美元/h.

在每个机场，用该模型仿真产生一个 8 小时工作制的人员换班调度表，该调度表与 Delta 航空公司实际航班的抵达和转机频率基本相吻合[14,15]. 基于抵达和离开的时间间隔、航空公司所涉及的机场数量、机场候机厅的数量和航班承载的旅客数量等，用泊松过程来模拟每个时间间隔到达的事先预订的 WP 数量，该调度表不包含事先未预订的 WP，但仿真过程将考虑. 此外，假定这些 WP 只与 Delta 航空公司的航班有联系，即可以认为 Delta 航空公司在独立地

使用这些机场的候机厅，当然这是理想的情况.

对于所考虑的机场的分布情况，可通过谷歌地球获得各个候机厅的卫星图像，用来找出登机口与信息服务台之间的距离. 同时，在 Delta 航空公司的网站上可以找到 2006 年所使用登机口的具体信息，将候机厅的这些信息转化为一个网格布局结构图，作为仿真过程的输入信息. 网格间距要比 $\Delta d = 150$ 英尺小很多，每一个登机口都可以被视为网格图内的一个普通的节点. 对于约翰肯尼迪机场候机厅，电梯也被看做一个节点，在这个候机厅中抵达与离开的登机口分布在不同的楼层，所以每次从抵达登机口到离开登机口都包括一段在电梯上运动的距离，即需要花费一个时间步长（1 分钟）的时间.

1. 约翰肯尼迪国际机场

对于一个有两个候机厅的中型机场候机楼，Delta 航空公司在约翰肯尼迪国际机场的候机厅拥有 20 个登机口. 根据 2006 年 Delta 航空公司的报告，轻度、中度、重度繁忙的航空流量分别为每 8 小时有 10 架次、30 架次和 50 架次的航班抵达[16]. 尽管约翰肯尼迪机场的候机厅大于芝加哥奥黑尔机场的候机厅，但抵达约翰肯尼迪机场的航班远少于抵达奥黑尔机场的航班. 预计在 8 小时中服务陪护人员的费用曲线如图 5-2 所示[1].

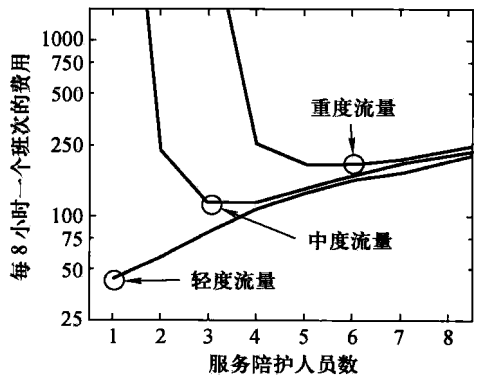


图 5-2 2006 年约翰肯尼迪国际机场服务陪护人员的模拟费用曲线

在三种航空流量下，所需要的最少服务陪护人员分别为 1 人、3 人和 6 人，每护送一次 WP 所需费用分别为 40 美元、115 美元和 190 美元. 于是，根据（5.6）式计算可得每年服务陪护人员的最小花费为 121 000 美元.

2. 奥黑尔国际机场

Delta 航空公司在芝加哥奥黑尔国际机场的候机厅只有 5 个主要的登机口. 根据 2006 年 Delta 航空公司的报告，在 8 小时内的轻度、中度和重度航空流量情况下，抵达航班数分别为 43 架次、129 架次和 215 架次[16]. 预计 8 小时中服务陪护人员的费用曲线如图 5-3 所示[1].

在这三种航空流量下，所需要的最少服务陪护人员分别为 3 人、6 人和 9

人，每护送一次 WP 所需费用分别为 90 美元、180 美元和 300 美元. 于是，根据（5.6）式计算可得每年服务陪护人员的最小花费为 152 000 美元.

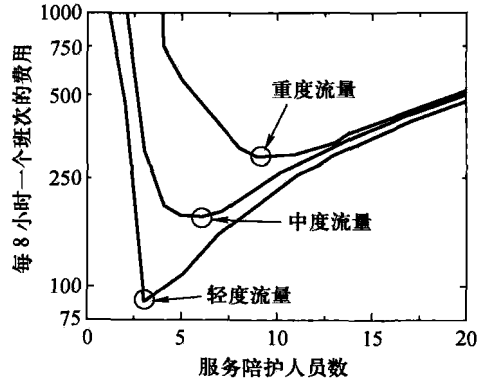


图 5-3 2006 年奥黑尔国际机场服务陪护人员的模拟费用曲线

3. 亚特兰大国际机场

Delta 航空公司总部设在亚特兰大，候机楼设有 4 个大型候机厅和 20 个主要的登机口. 尽管与约翰肯尼迪国际机场有相同数量的登机口，但亚特兰大国际机场登机口的航空客流通过量更多. 根据 2006 年 Delta 航空公司的报告，在 8 小时内的轻度、中度和重度航空流量情况下，抵达亚特兰大国际机场的航班数量分别为 107 架次、321 架次和 535 架次[7]. 预计在 8 小时中服务陪护人员的费用曲线如图 5-4 所示[1].

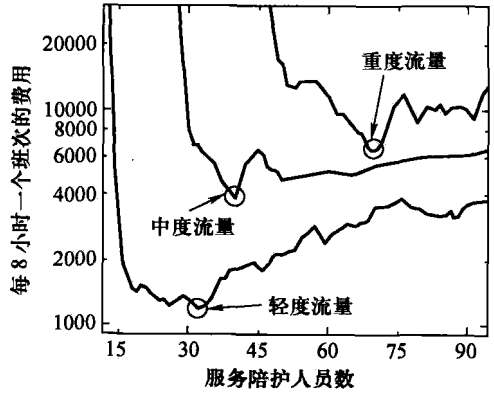


图 5-4 2006 年亚特兰大国际机场服务陪护人员的模拟费用曲线

在三种航空流量下，所需要的最少服务陪护人员分别为 32 人、40 人和 70 人，每护送一次 WP 所需费用分别为 1 180 美元、3 880 美元和 6 430 美元. 于是，根据（5.6）式可以计算得到每年服务陪护人员的最小费用为 3 977 000 美元. 这个数字比奥黑尔或约翰肯尼迪机场的费用大得多，因为亚特兰大国际机场的候机厅大约是其他机场的 4 倍，航空客流量大约是其他机场的3～10倍

之多. 航空公司在亚特兰大国际机场的花费主要是由航班延误造成的. 为了满足较大候机厅的实际需要，假设在亚特兰大国际机场转乘航班的平均时间为75分钟，而在其他机场设为45分钟. 这也是与实际的航空公司的航班时刻表相吻合的.

5.4.6 模型的结果分析

在机场最需要轮椅服务的是高龄人群，美国有关统计资料显示，未来5年中高龄人口将增长7%，未来25年将增长87%，未来45年将增长112%[17]. 假设高龄人群快速增长的模式保持不变，机场设施的布局也不变，利用上面模型的算法，就可以预测出航空公司未来在提供轮椅服务方面的费用变化情况.

未来30年中，在中度航空流量的情况下，飞抵约翰肯尼迪国际机场的航班数将从50架次增长到100架次以上. 利用模型的仿真算法进行仿真，发现所需服务陪护人员的最优数量将从6人增长到9人，相应的每8小时所需费用增长到600美元. 结果如图5-5所示[1]，同时仿真计算了轻度航空流量和重度航空流量情况下所需费用的变化趋势.

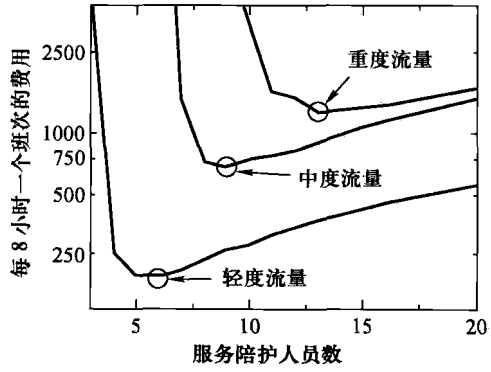


图5-5 2036年约翰肯尼迪机场候机厅服务陪护人员的模拟费用曲线

这个模型提出了一个航空公司如何利用最少的费用来管理航空轮椅及陪护服务的方案. 所给出的算法可以根据不同的候机厅布局和航班时刻表，预测未预订需要轮椅乘客数量、可能的航班延误和适当地分配服务陪护人员在空闲时去从事其他的服务工作，以减少航空公司的成本. 将这个模型和相应的仿真算法直接应用于Delta航空公司在芝加哥、纽约和亚特兰大机场候机楼的实际情况，其结果完全与实际相吻合，充分说明了模型和算法的有效性以及方案的可行性.

针对Epsilon航空公司，利用一个优化的earliest-due-date-first（EDD）算法来求使总费用最小的方案，这个算法适用范围广，适用于不同的机场候

机厅和航班时刻表，以及航班延误的情况. 并且允许服务陪护人员在空闲时从事其他服务工作. 此外，还可以为每一位工作人员设计一个工作过程的调度表.

对于航空公司的任一种轮椅配置方案，如果可以通过雇佣最少数量的服务陪护人员能够保证将每一个 WP 护送到他们需要的登机口，则这个方案是可行的. 如果这种方案不是最优的，则可以在预计客流量的基础上，利用这个算法，通过分配不同的陪护人员数量和轮换班次来改善. 比如，Delta 航空公司在亚特兰大国际机场利用一种没有进行优化的方案，每年所需要的费用超过 500 万美元，然而，利用这模型所给出的方案可以把费用减少到每年 400 万美元. 对于 Epsilon 航空公司来说，如果应用这个模型来设计轮椅配置方案，同样也会有类似的效果.

由上面的分析可知，这个模型具有以下的特点：

(1) 模型简单易行，使用非常灵活. 对于相关参数，比如机场布局、乘客数量和服务陪护人员的行走速度、成本费用和其他参数等，都可以方便地更新设定，并且能够对仿真算法做相应的调整改动.

(2) 即便是对最忙碌的机场，仿真计算所需要的时间也是很小的.

(3) 对于候机厅内部的设施利用网格方法，视为“城市街区”距离，显得更加自然和逼真.

(4) 算法简单且直观，便于交互式操作和判别.

(5) 算法实用有效，能够为每一个服务陪护人员制定最优的调度表.

5.5 基于成本效益的配置模型

5.5.1 问题的分析

在美国虽然大约只有 0.6% 的人坐轮椅，但有关方面的统计表明，在乘飞机旅行的乘客中，需要轮椅服务的人数大约在这个数字的两倍以上[18]. 对于航空公司来说，既要考虑服务成本，又要考虑服务效益. 影响服务成本和服务效益的最重要的两个问题是：

(1) 需要配备多少轮椅和服务陪护人员？

(2) 如何调度所配备的轮椅和服务陪护人员？

如果轮椅的服务陪护人员配置多了，则会增加航空公司的成本，服务效益也会降低. 如果配置的数量太少，则会出现不能满足乘客需求的情况. 如果对服务陪护人员没有统一的调度方案，可能导致有些区域无人管理，每个人的工作量也会不均衡. 同时也会影响到服务质量和服务效益. 因此，我们提出航空公司在机场合理配置轮椅和服务陪护人员的优化调度方案的问题，目标是减少

航空公司的运行成本，提高服务效益. 通过优化设计调度算法，从而最大限度地利用每个陪护人员和有效的服务时间，最后确定出了使得成本最小的轮椅配置和服务陪护人员的数量.

这里给出一个随机仿真驱动优化算法，即将问题的解决过程分为3个阶段：预仿真阶段、仿真算法实现阶段与优化调度阶段. 预仿真阶段是生成用于仿真的必要输入参数和数据，比如机场的布局和WP的需求清单，其中包括一天中轮椅的需求量和陪护服务人员的需求量. 仿真算法实现阶段是通过一个连续陪护事件为基础实现仿真，即包括乘客的到达与离开时间，以及轮椅与服务陪护人员的运动过程. 最后，在优化确定服务陪护人员数量的基础上，使得航空公司的总成本费用达到最小化.

5.5.2 模型的建立方法

1. 确定机场的设施布局

利用仿真的方法确定机场的设施布局. 将机场描述为一个双向图，其中节点代表登机口，即入口和出口，以及其他类似设施的位置点. 点与点之间的边表示可行走的通道，通常为穿过候机厅的主要通道.

假设从一个登机口走到另一个登机口所需的时间有的短些，有的很长. 也就是说两个登机口可以是相邻的，也可能是在通道的两头，在这里可以估计出在两个登机口之间的通道上行走所需要的时间.

对于一个机场的网络图，总是可以通过编码将其表示为一个 $n\times n$ 的邻接矩阵 $\boldsymbol{A}$，用 a_{ij} 表示从节点 i 到节点 j 所用的时间. 用实际的行走距离除以3英里/小时的行走速度得到行走所用的时间. 将从一个节点到另一个节点行走的最短时间（可用迪杰斯特拉算法得到）保存在矩阵 $\boldsymbol{D}$ 中，即用 d_{ij} 表示从节点 i 到节点 j 所用的最短时间.

假设所有服务陪护人员都知道任意两个登机口（节点）之间的最短距离. 在预仿真过程中，为了简化不考虑从登机口到飞机之间的距离. 在同一个机场，一般从登机口到飞机间的廊桥长度都是相等的，所以在这里不加区别.

2. 确定轮椅与服务陪护人员的数量

将轮椅与服务陪护人员视为一个统一的整体，统称为陪护. 航空公司可能有一些备用轮椅，当使用中的轮椅发生故障时应急使用，这些备用轮椅的费用包含在轮椅的保养和存储费用之中. 关于服务陪护任务，这里给出一个重要假设：在仿真过程中陪护的数量保持不变. 在现实中，陪护的数量会按一定的规律发生变化. 在仿真一天的过程中，假设在一个陪护开始工作时，另一个陪护立刻结束工作来代替他的位置. 同时，在仿真过程中，不会新雇佣或解雇服务陪护人员，也不会新购置或淘汰旧轮椅. 在总成本费用函数中，用一个下凹的函数表示与这些活动有关的成本费用.

3. 确定乘客需求的计划表

针对一个航站楼，确定一天的经过该航站楼的航班时刻表. 统计一天中经过该航站楼的所有乘客的数量，平均每架降落的飞机按 125 名乘客计算，用这个数据就可以估算出一天中起飞和降落的航班数量. 对于像亚特兰大这样忙碌的机场，通过观察起降的航班有关信息可以证明这一点[19]. 同时也可以了解其他方面的信息，如上午 6 点与下午 10 点之间的情况是比较规律的，晚上的航班相对较少[20].

首先，根据航班起飞的情况，按一定的规律分为若干个时间段. 然后，让其在不超过 1 小时的时间内随机变化，从而，利用相应的仿真算法不断地进行调试. 最后，将这些航班分配到特定的登机口. 接下来，建立一天中需要陪护帮助需求的乘客数据量，试图按照正常需要轮椅人数占总人群的比例来模拟总客运量和 WP 所占百分比，由此来产生 WP 的数量. 事实上，轮椅的需求比例不是 0.6%，而是 1.2%[18].

假设每一个 WP 在到达半小时内都能被安排一辆轮椅和服务陪护人员. 同时假设有一定比例的 WP 是提前预订的，所有请求的等待时间为 0. 随机产生 WP 请求需要帮助的时间，从乘客到达前的半小时到 5 小时之间随机变化. 根据请求时间进行分类，给出详细的乘客需求时间安排表. 这样当仿真算法用到该安排表时，它将模仿调度员在安排一天中接收 WP 帮助请求的情况，包括没有预约但需要轮椅服务的情况.

仿真算法可以通过改变每天的需求安排来实现每天不同的情况. 需求安排表模拟了仿真过程中的乘客流量负荷变化，特别要关注重度繁忙时段的详细安排. 此外，还要注意到某一天需要帮助的请求增加了、旅游旺季航空枢纽机场的情况以及在每年的旅游旺季旅游热点地区的航班运行情况等.

4. 确定陪护人员的调度计划

假设每名服务陪护人员都使用对讲机与调度员保持联系，调度员时刻掌握所有服务陪护人员的调度安排情况. 当一名服务陪护人员完成一次陪护任务后，会及时与调度员取得联系. 调度员随时掌握服务陪护人员在网络节点（机场候机厅登机口）x 到 y 之间行走所需要的时间，记为 δ_{xy}. 调度员在任何时间都可能收到需要轮椅服务的预约请求，设预约请求包括四部分信息：WP 的到达时间 t_a、到达地点 a、离开时间 t_d 以及离开地点 d. 为了使每一位 WP 能够顺利地转机，服务陪护人员必须在 WP 到达之前的某个时间离开上一个服务点，前往下一个要服务的指定位置，即预定的时间段为

$$t_f = t_d - \delta_{ad}.$$

按照“先到先服务”的原则，即按 WP 的到达时间先后依次排列. 同时，也要考虑登机截止时间，这样使得轮椅和服务陪护人员的调度计划表更加紧凑.

现在的问题是：如何为所有 WP 分派合适的服务陪护人员呢？

为了给一个 WP 分派一个服务陪护人员，航空公司的调度员首先需要确定在某一时间是否有人能够完成陪护任务，然后从所有可能完成这项陪护任务的陪护人中找出一位最适合的人，即能够使得从一个节点到另一个节点所需要的时间最短，并且有最优的行走路线. 按照这种方法依次进行下去.

服务陪护人员的安排可能会有下面四种情况：

（1）服务陪护人员不能按时护送乘客到目的地，但只要等待时间不超过 δ_w，就不会出现误机. 为了使这种情形尽量不出现或少出现，调度员在确定服务陪护人员 e 能否完成这项陪护工作时，先要确定他所在的位置 l_e、完成前一项护送任务的结束时间和完成这项任务需要的时间 t_e. 应满足

$$t_e+\delta_{ea}<t_f+\delta_w,$$

满足这一要求的服务陪护人员记为第 O_1 组.

（2）虽然服务陪护人员护送一位 WP 准时到达了目的地，但在去护送下一位 WP 时却迟到了. 与前面情况类似，但没有上面的延迟时间 δ_w. 即有

$$t_e+\delta_{ea}<t_f,$$

符合这一要求的服务陪护人员记为第 O_2 组.

（3）按照实际中的轮椅和陪护人员的调度计划表，希望在每一位服务陪护人员完成一项任务之后，还能够按时完成后面的另一项任务，并且不受航班延误事件的影响，这是一种理想的情况. 将符合这一要求的服务陪护人员记为第 O_3 组.

（4）一名服务陪护人员在完成一项护送任务之后，紧接着去执行下一项护送任务. 这个过程不需要调度员重新安排就可以连续执行. 如果服务陪护人员 e 在执行完一项任务后，要求他在 t_s 时刻于 s 位置开始执行下一项陪护任务，即该服务陪护人员就能够在时间 t_s 之前从位置 d 到达位置 s 以完成这项护送任务. 则有

$$t_e+\delta_{ds}<t_s,$$

将符合这一要求的服务陪护人员记为第 O_4 组.

航空公司调度员在确定陪护方案时，要在最适合的非空组中选择. 从这个组中选择一名前一项陪护任务已经完成，并且距离下一个需要陪护乘客位置最近的服务陪护人员. 如果这项新任务与另一项或多项任务在时间上有冲突，调度员必须要调整调度计划表，以适应后面的陪护服务任务.

如果乘客没有预订服务请求，到达时服务陪护人员或者都忙于其他陪护任务，或者是因距离太远，而不能保证准时到达所需要陪护的地点. 在这种情况下，这些 WP 可能会导致航班延迟或误机，这些特殊情况不在本模型考虑的范围之内.

5.5.3 模型的仿真算法

我们将这个问题抽象为一个时间“装箱”问题，调度员要尽可能使服务陪护人员的调度安排符合实际任务的要求.

仿真计算的主程序算法：

```
主程序:mainsimulation(D,R,N_E)
    创建陪护人员的任务集合 E;
    I=findindexunfulfilledrequest(R);
    当 I>0 时,执行:
    O=makeoptimalitymat(D,R_I,E);          %说明现在执行任务
    (R,E)= executerequest(D,R,E,I,O);
    I=findindexunfulfilledrequest(R);
    计算总共误机事件=sum(R_误机事件);
    计算总共延误时间=sum(延误时间);
    返回(总共误机事件,总共延误时间).
```

输入参数：最短行走时间矩阵 D；请求轮椅和陪护服务人员列表 R；航空公司服务陪护人员的数量 N_E.

创建服务陪护任务集合（数组）E，陪护任务队列中的 e_i 是指第 i 个服务陪护人员的陪护任务.

程序包含两个子程序：makeoptimalitymat( ) 和 executerequest( ). 由这两个子程序决定分配哪一个服务陪护人员去执行哪一项服务陪护任务，并且在实际过程中应如何调整当前的调度表以适应实际需要.

子程序 makeoptimalitymat( ) 的执行结果形成一个 $N_E \times 4$ 的矩阵，第 i 行表示第 i 个服务陪护人员 $e_i(i=1, 2, \cdots, N_E)$ 是否属于第 j 个最优组 $O_j(j=1, 2, 3, 4)$.

由子程序 makeoptimalitymat( ) 执行的结果矩阵的每一行都是由另一个子程序 optimaltycheck( ) 产生的. 子程序 optimaltycheck( ) 的计算过程如下：

```
子程序:optimalitycheck(D,R_I,e_i)
    初始化 O_j=(0,0,0,0)(j=1,2,3,4);
    如果服务陪护人员 e_i 在第一组 O_1 中,即可以在航班延迟期限内完成陪护任务 R_I,
    则取 O_{i,1}=1;
    如果服务陪护人员 e_i 在第二组 O_2 中,即能够完成 R_I,平均延误时间 w/o,
```

则取 $O_{i,2}=1$；
如果服务陪护人员 e_i 在第三组 O_3 中，即能够完成 R_l，不会出现延误，
则取 $O_{i,3}=1$；
如果服务陪护人员 e_i 在第四组 O_4 中，即能够完成 R_l，有延误可改变任务，
则取 $O_{i,4}=1$；
返回 O_j

对于每一个需要轮椅帮助的请求，子程序 optimaltycheck( ) 将每一个服务陪护人员分配到某一个最优的组里，创建一个最优矩阵 O 中一行，$O_{i,j}$表示服务陪护人员 e_i 是否是被分配在最优组 O_j 中，也就是说 e_i 是否是 $O_j(i=1, 2, \cdots, N_E, j=1, 2, 3, 4)$ 组中的元素.

最后，在可能的情况下，用子程序 executerequest( ) 为每一个 WP 分配一个服务陪护人员. 随着服务陪护人员分组的不断优化，调度员将收到越来越多的陪护任务分配调整工作，来试着满足所有的 WP 请求.

子程序：executerequest(D,R,E,I,M)
如果第四组 O_4 中包括一个 1，即矩阵 O 中的第 4 列包含有一个 1，则执行：
找出要完成的任务 e^* 之前离得最近的服务陪护人员；
将该任务插入到 e^* 的调度计划表中；
将需要陪护帮助的需求 R_l 标记为已完成.
如果第三组 O_3 中包括一个 1，即矩阵 O 中的第 3 列包含有一个 1，则执行：
找出要完成任务 e^* 之前离得最近的服务陪护人员；
找出该任务的插入点 s；
将 s 点后的任务尽可能远的反向移动；
将该任务插入到 e^* 的调度计划表中；
将需要陪护帮助的需求 R_l 标记为已完成；
将 s 点后的任务尽可能远的正向移动.
如果第二组 O_2 中包括一个 1，即矩阵 O 中的第 2 列包含有一个 1，则执行：
找出要完成任务 e^* 之前离得最近的服务陪护人员；
找出该任务的插入点 s；
将 s 点后的任务尽可能远的反向移动；
将该任务插入到 e^* 的调度表中；
将需要陪护帮助的需求 R_l 标记为已完成；
移动调度计划表中有冲突的任务；
将相应的需要陪护帮助请求标记为未完成；
将 s 点后剩下的任务尽可能远的正向移动.
如果第一组 O_1 中包括一个 1，即矩阵 O 中的第 1 列包含有一个 1，则执行：

找出要完成任务 e^* 之前离得最近的服务陪护人员；
找出该任务的插入点 s；
将 s 点后的任务尽可能远的反向移动；
将该任务插入到 e^* 的调度计划表中；
将需要陪护帮助的需求 R_i 标记为已完成；
移动调度表中有冲突的任务；
将相应的陪护请求标记为未完成；
将 s 点后剩下的任务尽可能远的正向移动；
记录 R_i 的延迟时间.

对于其他的情况，则执行：

将任务标记为已完成；
将 R_i 记录为误机的乘客数量.

返回 (R,E).

5.5.4 运行成本的计算

我们以美元为单位来计量所需要的费用，问题共有以下两类成本费用：

(1) 轮椅与服务陪护人员的费用，即由轮椅的购置费、维护费、存放费和服务陪护人员的工资等成本费用组成.

(2) 航班延误带来的损失和乘客误机造成的损失等.

在仿真计算过程中，轮椅和服务陪护人员的成本费用在一天中被视为固定不变的. 在成本函数中，N_E 表示服务陪护人员的数量，R 表示每天预先提出需要轮椅和陪护帮助请求的清单，D 表示机场的布局分布图，X 表示误机的乘客数量，Y 表示航班在登机口等待迟到乘客的总时间. 则问题的目标函数为

$$[X,\ Y]=\text{mainsimulation}(D,\ R,\ N_E),$$

$$C(X,\ Y)=C_{\text{固定}}+C_{\text{误机}}X+C_{\text{延误}}Y.$$

假设航班在登机口等待乘客平均每小时要损失费用 $C_{\text{延误}}=1\ 018$ 美元，1986 年的平均损失费用受到了通货膨胀的影响[21]. 每名乘客误机的费用为 $C_{\text{误机}}=500$ 美元，对票价和住宿的赔偿都是必要的. 固定不变的费用（常量）包括轮椅的购置费用、维修费用、存放费用和服务陪护人员的工资等，假设每年用来维修、存放和更新轮椅所需要的费用为 130 美元，服务陪护人员的工资为每小时 10 美元[22]. 于是，单位时间的固定费用可表示为

$$C_{\text{固定}}=\left(10+\frac{130}{365\times24}\right)HN_E,$$

其中 H 表示仿真模拟的时间，即为每小时产生的固定费用.

5.5.5 模型的检验与分析

在对上述仿真模型进行模拟计算，过程如下：

（1）针对每个调度策略有代表性的数据进行了10次检验.

（2）计算误差采用标准差.

（3）半数需要陪护的乘客在航班到达以前都向航空公司提交了预订申请.

1. 最优调度方案的检验

在这里针对有一个候机厅，中等客流量（如每天有15 000位乘客）机场的情况进行模拟检验，在模拟检验过程中，允许使用优化组 O_1，O_2 和 O_3 中的服务陪护人员. 也就是按照优化分组 O_1，O_2 和 O_3 的顺序依次分派服务陪护人员. 换言之，利用这种最简单的算法是能够保证在 O_4 组中没有服务陪护人时，不能错过一个需要陪护帮助请求；O_4 或 O_3 中没有服务陪护人员的情况下，也不会错过任何一个需要陪护帮助的请求，等等. 从而证明了这个调度方案的合理性和有效性.

模拟计算是基于有1.2%的乘客需要轮椅陪护帮助的假设，每一个服务陪护人员都可以循环执行陪护任务，也包含 O_3 组中的人员，这样对于航空公司来说可以节省成本，从而可以获得更大的收益. 模拟结果表明，航空公司可以少雇佣5～6个服务陪护人员，每天大约可以节省成本1 000美元，如图5-6所示. 通过对几个机场的具体情况进行模拟，运用这个方案所节省的费用将会随着乘客流量和机场规模的增大而增加. 就Epsilon航空公司而言，只要合理地使用这个方案，每天能够节省的费用在10 000美元以上.

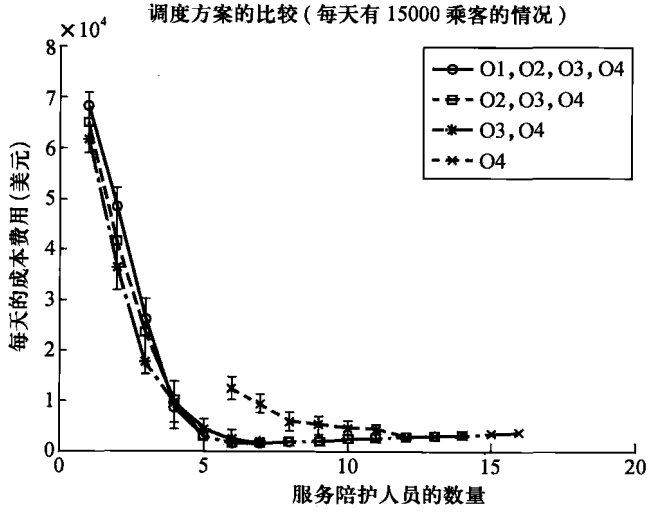
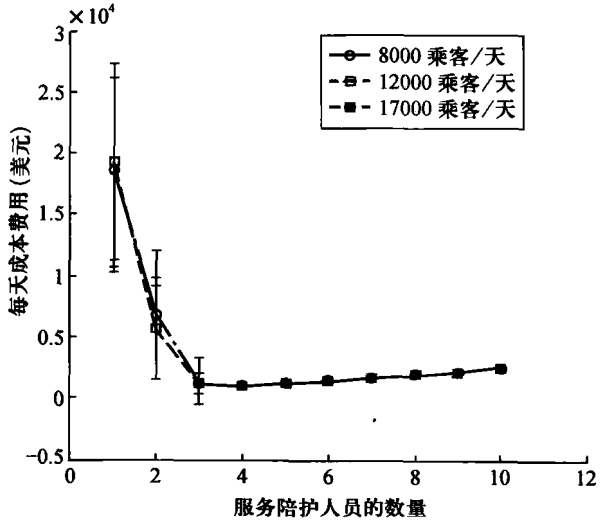


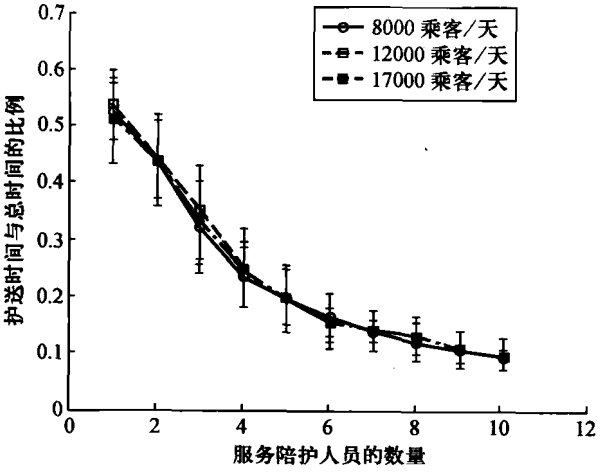
图5-6 调度方案的比较

2. 针对不同候机厅的模型性能与灵敏度分析

将芝加哥奥黑尔机场的航站楼作为测试点来检验在不同距离情况下的模型性能. 分别就 1 个、2 个和 4 个候机厅的情况进行模拟计算，每一种情况都针对航站楼 3 个不同级别的客流量进行了模拟仿真，计算结果如图 5-7、图 5-8 和图 5-9 所示[1].



(a) 一个登机口的费用



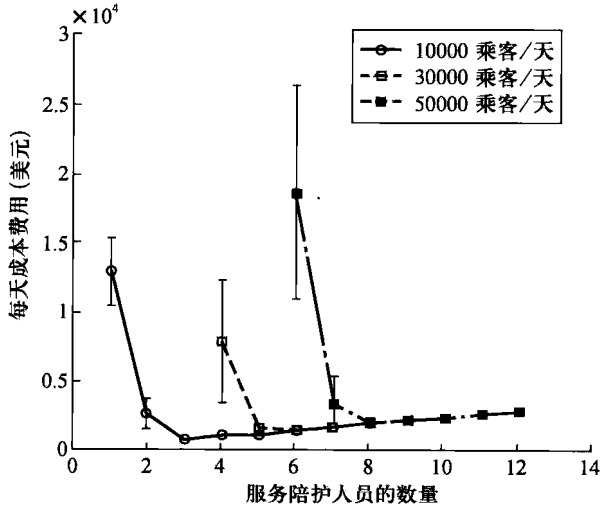
(b) 一个登机口的护送时间比例

图 5-7 奥尔黑 3 号候机楼

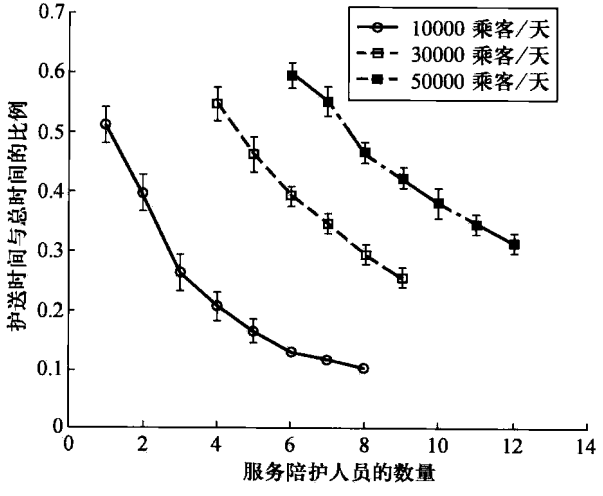
这里假设有 0.6% 的乘客需要轮椅和陪护服务帮助，需要服务陪护人员的最优数量大约与候机厅数量的增长成正比例关系. 此外，随着机场客流量的增长，需要服务陪护人员的数量有所增加，大体上来看航空公司的最优成本是稳

定的. 从图 5-7 可以看出，在只有一个候机厅的情况下，客流密度的变化几乎不会引起最优服务陪护人员数量的变化.

但是，图 5-8 和图 5-9 表明当有 2 个和 4 个候机厅时，服务陪护人员的需求量有大幅度的增加，这是因为在各登机口之间移动所需要的时间会增加.



(a) 成本费用



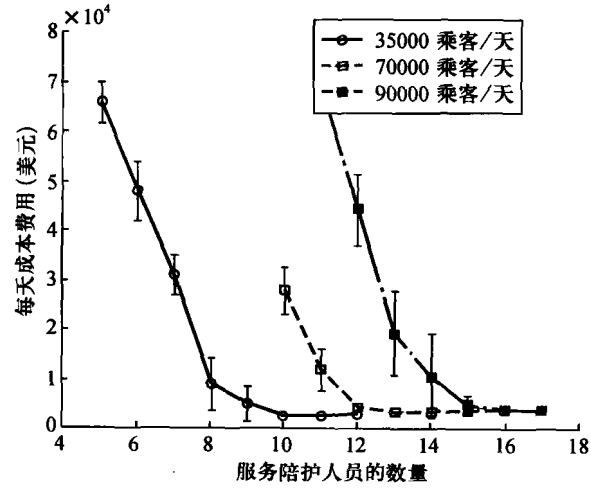
(b) 护送时间比例

图 5-8 奥黑尔 2 号候机楼

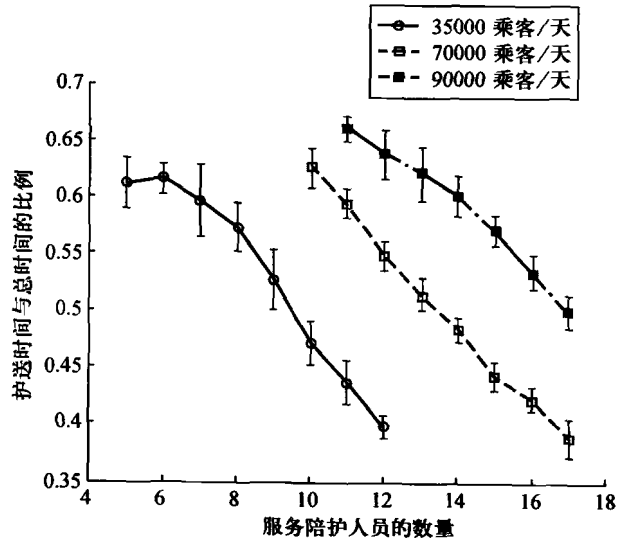
3. 针对不同机场的模型性能与灵敏度分析

本仿真模型算法适用于不同机场候机厅和多登机口的情况，利用机场卫星图像可以将其转化为由网络节点与边构成的网络图. 针对两种不同的客流量水平，假设有 1.2% 的乘客需要轮椅与陪护帮助，我们分别对纽约 LaGuardia

(LGA) 机场的两个候机厅、芝加哥奥黑尔机场和 Dallas/Fort Worth (D/FW) 机场进行了模拟对比分析.



(a) 成本费用

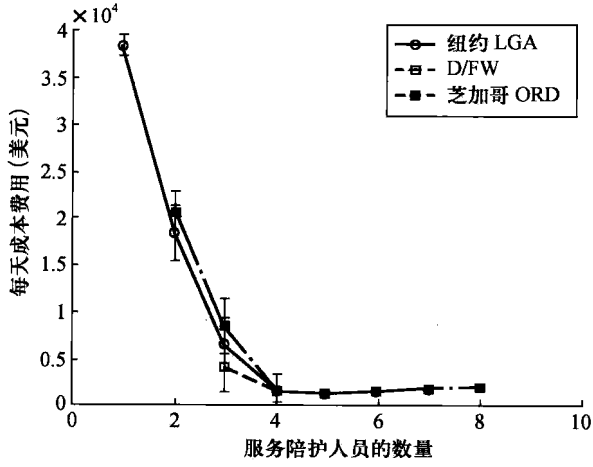


(b) 护送时间比例

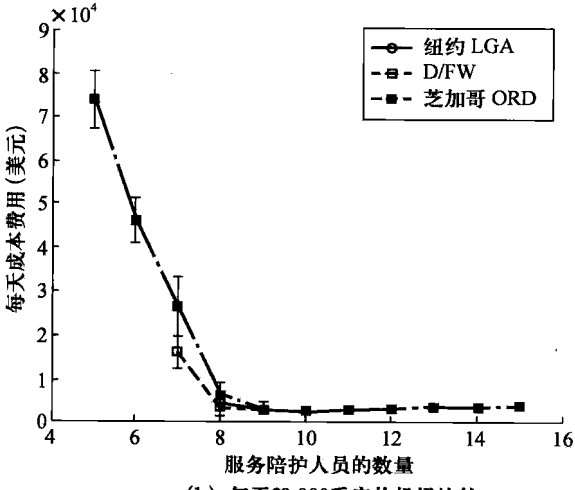
图 5-9 奥黑尔 3 号候机楼

从实际情况来看, D/FW 机场的效果是最差的, 因为该机场是一个圆形的航站楼, 乘客往往需要从航站楼的一边走到另一边, 差不多有 1 英里的距离, 需要的时间比较长. 芝加哥奥黑尔机场的效果是最好的, 因为该机场的设施布局很紧凑, 需要转机的乘客可以只穿过一个三角形交叉点状的候机厅就可以到达目的地.

图 5-10 的模拟结果与实际统计结果是一致的，表明了该模型算法的有效性. 虽然不同机场的几何形状和设施布局不同，每天乘客数量的变化范围也会很大，对每种情况都做了 10 次以上的模拟检验，检验结果保持着良好的稳定性. 同时注意到，在忽略登机口分布的条件下，对模型算法性能的影响也不大. 在多客流量情况下，由模拟结果可以发现机场设施布局的影响将很大.



(a) 每天10 000乘客的机场比较



(b) 每天30 000乘客的机场比较

图 5-10 比较不同机场的算法性能

4. 对人口老龄化影响的预测分析

随着社会的进步和发展，人们的生活水平不断地提高，人口的老龄化程度也在增加，乘飞机出行的乘客中需要轮椅和陪护帮助的数量将会增加. 如果总的客流量增加 10%，那么需要轮椅和陪护服务的乘客数量是否也会增加 10% 呢?

模型仿真结果表明：如果总的客流量增加 10%，那么需要轮椅和陪护服务的乘客数量会增加更多. 如果总的人口数量增加 10%，每天需要的航班数量会增多，而每个航班上对轮椅和陪护服务的需求量增加不会正好是 10%. 换句话说，由于乘客平均年龄的增加，将不仅使得轮椅和陪护服务的总需求量增多，而且也使得每个航班上乘客对轮椅和陪护服务需求量的增多，这显然是符合实际情况的.

对于一个具有两个候机厅的机场，我们利用这个模型做了两个方面的仿真模拟计算：

（1）将乘客的数量从 33 000 人增加到 42 000 人，航班的数量不变；

（2）乘客的数量保持在 30 000 人不变，将需要轮椅和陪护帮助需求的人数比例从 1.32% 增加到 1.68%.

针对这两种情况，模拟仿真结果说明，需要轮椅和陪护帮助请求的数量基本上不变，但平均最小成本费用会随着需要轮椅和陪护帮助需求人比例的增加而增大，如图 5-11 所示[1].

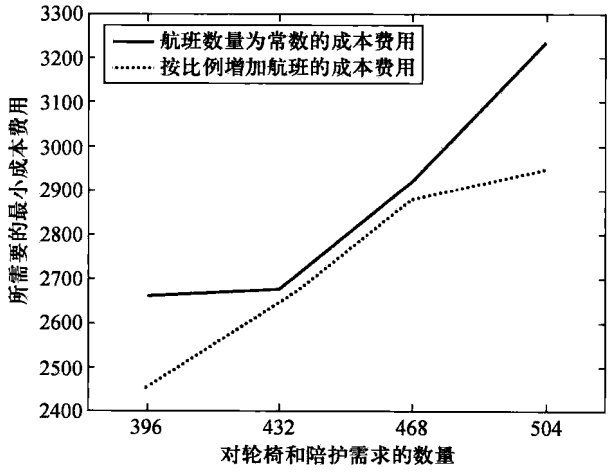


图 5-11 对轮椅和陪护帮助需求增加影响的比较

5.5.6 模型的结果分析

这个模型针对实际中不同的机场、不同数量的候机厅、不同数量的登机口和不同设施布局的情况，建立了基于最小成本费用的轮椅与服务陪护人员优化调度仿真模型，通过模拟计算，分别得到了相应的模拟结果，充分说明了模型算法的有效性和适用性. 并根据实际中可能的数据变化情况，对模型的灵敏度做了检验分析，也说明了模型算法的稳定性. 由于模型算法在各种环境下的有效性，则也能够说明该模型的算法优于其他算法，按照这个模型算法所得到的调度方案，能使得航空公司大大地减少成本费用，获得更多的利润.

但是，这个模型算法也有一个不可回避的不足之处，这就是模型没有对最小损失做特别的优化处理和研究. 比如，对于一个服务陪护人员在接到一个长时间（即长距离）的新的陪护任务时，可能在之前已经安排了5个短时间的陪护任务，为了保证这个长时间的新（重要的）陪护任务的完成，调度员可能不得不取消他先前安排的3个或4个陪护任务. 这样做可能使得被取消陪护任务的乘客在短时间内得不到陪护，甚至会导致误机. 在实际中这种情况应该是比较少有的，一般认为对模型的影响不大. 当然，如果这种情况出现的机会很多，我们可以不对服务陪护人员的数量进行优化，适当增加服务陪护人员，增加一些成本费用而减少损失，显然也是一种可行的方案.

在模型中假设没有因为其他原因使得航班延误到达或者延迟起飞，这显然也是一种理想的状态，但事实上，这些情况发生时势必会对服务陪护人员的正常调度产生一定的影响.

对于调度员来说，人工调度服务陪护人员是一件困难的事情，实际中这项工作可以由电脑系统自动完成，具体的细节问题在这里也没有考虑.

事实上，虽然这些不足之处没有单独考虑，但是，由于模型良好的稳定性和有针对性的改进措施，完全可以抵消这些问题的影响.

5.6 基于最小费用流的配置模型

这里将考虑如何确定利用最少的陪护人员，并且制定出陪护人员一天中最优的工作流程，以降低航空公司相应的管理费用和服务成本.

根据这个问题的实际情况和要求，将问题转化成一个网络流的优化问题. 将服务陪护人员和WP所在的位置视为网络图的节点；任意两个节点之间的成本费用视为网络图的边，它是时间的函数，并且与为WP提供陪护服务所延误的程度有关. 陪护人员总是沿着网络图的某条边移动，通常该网络流每时每刻都可能在流动中. 当一个陪护人员完成一项陪护任务，又被通知新的WP到达的时间，并要求其完成该项任务时，该网络将随之动态地发生变化. 利用网络流的理论，通过求解最小费用的网络流问题，就可以解决这个实际的问题.

5.6.1 基本术语与模型假设

1. 基本术语

（1）一次完整的陪护工作：一个陪护人员从在机场入口处接到一个WP开始，到护送该乘客至所要转机登机口的整个过程.

（2）理想的陪护开始时间：一个WP到达机场的时间，即乘客一来到机场即可接受服务.

（3）陪护结束时间：一个WP到达登机口并及时登机的时间. 如果在飞

机起飞之前不能完成护送任务，该 WP 将会误机.

(4) 价格、成本及有效成本：在这里不使用经济术语，价格与成本在实际意义有区别. 成本是由于未能完成护送工作而产生的费用，在实际中，通过良好陪护服务，期望这个成本有最小值. 而价格是通过运行 Dijkstra 算法计算出来的运行费用，并与模型有关.

2. 模型的假设与说明

(1) 在机场设施的网络图中，每条边上都有密度均匀的登机口. 实际中的机场不一定符合这种假设，但是该假设并不影响所建立数学模型的合理性，也不会影响实际问题的计算结果.

(2) WP 到达机场的登机口是随机的. 由于不同的 WP 可能去往不同的目的地，因此航空公司难以计算出所有出发口和到达口之间距离的精确值.

(3) 当一个 WP 即将到达机场的时间小于 1 小时时，飞机到达时间和机场进（出）口不会发生改变；距离飞机起飞时间小于 3 小时时，飞机的起飞时间与登机口不会发生改变. 当服务陪护人员正在陪护一个 WP 或正在前往转机登机口的路上时，一般不允许飞机出现延误和改变登机口的情况出现. 这种假设不一定符合实际，但它对最终计算结果的影响较小，并且使得模型更容易实现.

(4) 通常飞机中途经停的时间范围为 45 ~ 132 分钟，这里引入有效停留时间，简单地取为经停时间的较短的值. 飞机有效停留时间是考虑到 WP 一般是最先登机、最后下飞机，会耽误一些时间，而且这些时间难以基于经验数据确定.

(5) WP 到达机场是随机的，且是相互独立的. 一般不会有组团旅行的 WP，除非是残奥会或其他类似的特殊情况. 对于一般情况这种假设是合理的.

(6) 服务陪护人员都是通过电脑网络传递信息. 通过电脑网络为服务陪护人员派发任务和传递信息，速度快，准确度高，方便快捷.

(7) 服务陪护人员推着载有 WP 的轮椅的行进速度为推着空轮椅时的行进速度的一半. 二者之间的数量关系是人为设定的，因为缺乏相应的数据依据.

5.6.2 模型的建立

1. 网络流的定义及结论

为了研究问题的方便，先引入网络流的基本概念和结论.

(1) 网络流是一个有向图，其中每条边都用一个正数值表示其流量大小，其两个连接节点分别被标记为该流的源点和汇点.

(2) 网络流是用非负实数对每条边的一个流量分配，并使得每个数值等于相应边的最大流量值；每个节点或者是源点，或者是汇点，来自源点的各条边的流量之和等于流向汇点的各条边的流量之和. 一个网络流量的大小是指来自源点的各条边的流量之和减去流向汇点的各条边的流量之和的差值.

(3) 最大网络流是指一个给定的网络可能的最大网络流量值. 我们可以

为每条边定义相应的单位流量的成本，从而可以分析整个网络流的流通成本.

（4）一个网络流的成本是指所有边的单位成本与该边流量乘积的总和.

（5）最小费用最大流是指一个具有最小流通成本的最大网络流. 如果所有边的最大流量值是整数，则总有一个为每条边分配整数值流的解决方案. 这里建立一个数据结构来动态地搜索问题的解，即以最小成本并以最大流来优化航空公司轮椅的调度工作.

（6）对于一个网络（原始网络）和它的一个流，剩余网络是指另一个具有相同节点集的网络，剩余网络中边的最大流量等于减去原始网络的最大流量的流量值. 由于一些边的最大流量减小，剩余网络一些边的流向可能与原始网络中的流向相反. 更确切地说，在剩余网络中的边 (i, j) 的最大流量是原始网络的边 (i, j) 加上由节点 j 到节点 i 的流量，并减去由节点 i 到节点 j 的流量. 在剩余网络中，边 (i, j) 的成本是边 (j, i) 成本的相反数. 剩余网络是很有用的，因为在原始网络中求解最小费用最大流的问题就等价于在剩余网络中的相应计算.

（7）一个全通循环流（环流）是指流向所有点（包括源点和汇点）的流量之和等于相应的流出之和.

（8）一个网络中的最小费用流就是其最小费用的环流. 如果 f 是最大流，则剩余网络图不可以接受更多的流（即剩余网络的最大流量为0）. 因此，在这里需要解决的问题是要找到一个最小费用的环流.

（9）价格函数是将每个节点与价格成本相互联系的函数.

（10）网络流中边 (i, j) 的有效成本为

$$c_{ij}+p_i-p_j,$$

其中 c_{ij} 表示边 (i, j) 的成本，p_i 和 p_j 分别表示节点 i 和节点 j 的基本价格.

（11）以有效的费用成本来表示网络环流的流通成本，并且这个值是恒定的.

设 f_{ij} 表示通过网络环流 f 的边 (i, j) 的流量值，c_{ij} 表示边 (i, j) 的费用成本，p_i 表示节点 i 的价格成本，那么网络环流 f 的有效费用成本可表示为

$$\begin{aligned}\sum_{i,j} f_{ij}(c_{ij}+p_i-p_j) &= \sum_{i,j} f_{ij}c_{ij} + \sum_i p_i\Big(\sum_j f_{ij} - \sum_j f_{ji}\Big)\\ &= \sum_{i,j} f_{ij}c_{ij} + \sum_i p_i \cdot 0 = \sum_{i,j} f_{ij}c_{ij}.\end{aligned}$$

2. 成本费用的构成

考虑到机场的陪护任务负荷和陪护人员数量的限制，由此产生的 WP 不能被及时服务时的延误情况，在这里统计给出相关延误的基本数据（延误时间）与陪护人员数量的对应关系表，同时找出每增加一个服务陪护人员产生的最大边际效用.

主要的成本是因 WP 等待服务而产生的费用，根据这里所讨论问题，这个

费用不是一个直接的货币支出的形式，而是基于等待时间（分钟）来定义的相应单位成本. 实际中，由于 WP 必须等待轮椅和陪护人员被护送至登机口. 如果 WP 不能在飞机起飞前 15 分钟内到达相应的登机口，他们将延误登机，如果由此而导致飞机起飞的延迟，将产生 30 个单位成本的损失. 如果一个 WP 误机，则将产生 10 万个单位成本的损失.

3. 机场的布局模型

该模型依据机场布局结构网络图，一条边表示一条通道，顶点表示通道之间的交叉口. 每一条边定义了相应的成本，即指通过该通道所需要的时间（以分钟为单位，四舍五入），不区分通道终端的左右两侧. 一些边上带有乘客到达和离开的进出口. 服务陪护人员仅可以沿着部分边并在适当的时间内进行移动.

在机场网络图的边上，均匀地分布着一些位置点，即 WP 到达或离开机场的进（出）口，均匀地分布在机场候机厅内.

在这里，我们模拟几个机场的布局结构图：一是单一直线的小型机场（记为机场 1）；二是波士顿洛根候机楼 A（由一条地下通道连接的两个候机厅，为中型机场（记为机场 2））；三是芝加哥奥黑尔 3 号航站楼，包括候机厅 G，H，K 和 L，为大型机场（记为机场 3）.

采用 Floyd-Warshall 算法，预先计算每对顶点之间的最短路径. 运用该算法反复计算相应的距离 $d_k(i, j)(i, j=0, 1, \cdots, k)$，并有

$$d_k(i, j)=\min\{d_{k-1}(i, j), d_{k-1}(i, k)+d_{k-1}(k, j)\},$$

其中 i，j 表示机场网络图中的两个不同的节点，共有 $k+1$ 个节点.

对于一个有 n 个节点和 m 条边的机场网络图，该算法运行的复杂度为 $O(n^3)$，算法易于实现. 然后，采用 Dijkstra 算法，并基于不同的机场网络图求解相应的最短距离.

综上所述，求解最小费用环流问题等价于求解相应的有效费用成本最小的问题. 如果所有的费用成本都为非负，那么空的环流是最小费用的环流.

假设在网络图中，由某一节点 V 到所有的节点都是可达的，并且所有边的费用成本均为非负. 那么可以基于各边的成本权值来计算从节点 V 到 B 点的最短路径. 设 W′和 W 是任意的两个节点，引入价格函数为

$$p_w=d(\mathrm{V}, \mathrm{W}),$$

则节点 W 的价格成本可以表示为节点 V 和 W 之间的距离，并可得如下结论：

（1）由于 $d(\mathrm{V}, \mathrm{W})\leqslant d(\mathrm{V}, \mathrm{W}')+c_{\mathrm{W}'\mathrm{W}}$，那么从节点 W′到节点 W 的有效费用成本为

$$d(\mathrm{V}, \mathrm{W}')+c_{\mathrm{W}'\mathrm{W}}-d(\mathrm{V}, \mathrm{W}),$$

该价格成本为非负值.

（2）由于沿着节点 V 到节点 W 的路径，上面的不等式是严格成立的. 因

此对于所有的节点 W，一定存在一条由节点 V 到节点 W 的路径，并使得该路径的有效费用成本为 0.

5.6.3　模型的求解算法

针对一个机场的网络流图，对于每一个已经到达或事前预定需要轮椅陪护的 WP 的情况，这里采用 Dijkstra 算法进行相应的求解运算.

1. 模型求解的 Dijkstra 算法

对于各边权值均为正值，且具有单一源点的网络流图，采用 Dijkstra 算法可以求解其最短路径，即对某一节点 V，通过遍历所有其他节点 W 来计算 $d(V, W)$，该方法称为广度优先搜索. 我们采用 Dijkstra 算法而不采用 Floyd-Warshall 算法，这是因为对于求解单源最短路径的网络流问题，Dijkstra 算法更有效. 我们可以求解距离 $d(V, W)$ 和 $\min_{U}(d(V, U)+l_{UW})$ 的近似值，其中 l_{UW} 表示从点 U 到点 W 的距离长度，则可得到所要求的最短距离近似值.

我们注意到，如果没能在飞机起飞前的 15 分钟内护送 WP 登机，每一个 WP 将产生 30 个单位的成本费用. 如果在飞机起飞前没能将所有的 WP 护送至飞机上，则会导致某些 WP 误机，或使飞机延误，将使相应的有效费用成本达到最大值. 在任何时间，我们总是希望能够以最低的费用成本安排轮椅和陪护人员完成当前已知的陪护任务.

在这里，将该问题视为一个最小费用的最大流问题，如图 5-12 所示. 陪护人员的任务安排情况构成相应的网络流. 网络流的流入是在一天工作开始时服务陪护人员到达机场，网络流的流出是他们在一天工作结束时离开机场.

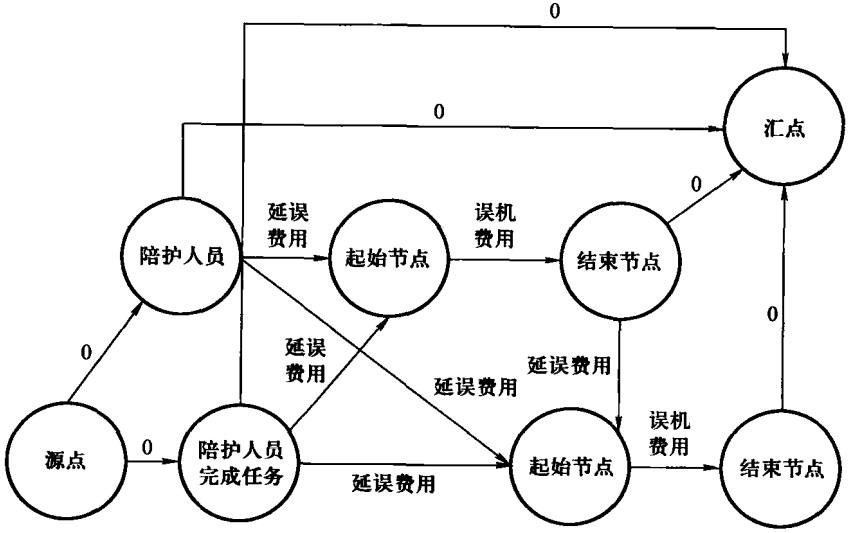


图 5-12　问题的最小费用最大流图

在图 5-12 中，节点表示陪护任务和陪护人员，边上的网络流表示陪护人员的移动情况. 护送延误和误机所产生的费用分别被标注在各条边上.

网络流的边（即护送路径）决定了相应的费用成本. 然而，对于一个陪护人员，如果刚开始接受护送任务，但还未沿边移动时，则可认为它不产生任何费用. 此时选取一个较大的负值（例如-100 000），将其设为一天工作开始时的负成本. 在这一天工作结束时，用这个负成本来核对有多少 WP 被准确地护送到了所乘的航班上.

如果预知每位 WP 的到达时间，那么可以简单地解决网络图中的最小费用最大流问题，并得出最优的轮椅和陪护人员的安排方案. 当已知 WP 即将到达时，将在网络图中增加相应的任务流；相反地，当一个护送任务结束或者被中止的，在网络图中减少相应边的任务流. 随着网络流量的改变，保持并不断更新相应的数据结构.

2. 求解算法的实现

（1）关于时间进程. 随着时间的推移，服务陪护人员或许不能够在规定的时间内完成护送任务，那么将要更新因该陪护人员的护送工作延误而产生的费用成本. 对于当前某一时刻的网络流，每条边分别对应于一个陪护人员正在进行的护送工作，并且该陪护人员总是在尽力准时护送 WP 到目的地，因此该边的费用成本不会变化. 对于一个正在进行中的护送工作，每流逝 1 分钟将产生相当于 1 分钟的费用成本，但同时也预示着护送工作更加接近其结束的时间，二者产生的效果可以相互抵消. 因此，此变化不会影响剩余网络流的有效成本，而是保持相应的值不变，如图 5-13 所示.

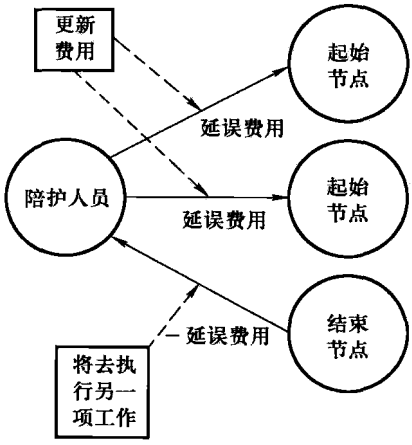


图 5-13 费用成本随时间进行的更新

在图 5-13 中，每条边的费用成本随着时间的推移而不断更新.

（2）护送工作的开始. 当一个陪护人员开始一项护送任务时，删除该陪

护人员的节点，并连接相应的源点和终点. 这样可以使一个剩余网络发生相应的变化，但不产生负的有效费用，如图 5-14 所示.

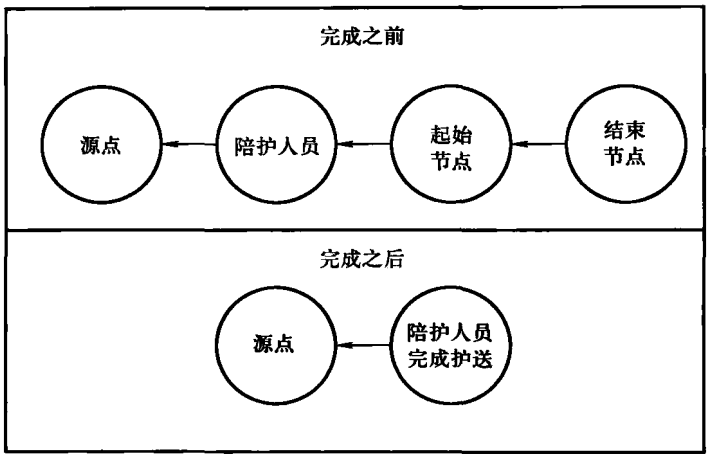


图 5-14 护送任务的分配

图 5-14 是为了说明避免将同一项护送工作安排给两个或两个以上的服务陪护人员去完成.

(3) 护送工作的结束. 当一项护送任务完成时，应该在图中重新标记对应于该项任务结束的节点，使其对应于完成该项任务的陪护人员所处的当前位置，如果之后分配给该陪护人员新的护送任务，则该陪护人员将移动至新的任务；若没有安排新的任务，则认为该陪护人员不发生任何的移动. 当一项护送任务结束后，将结束的节点标注为执行该任务的陪护人员的节点，以等待接受下一项护送任务.

(4) 新的护送工作开始的标志. 当有 WP 到达或者预知到达时间时，就在网络图中创建节点 U 和 V，分别对应该项护送任务的起始节点和结束节点. 然后，以每一项正在进行护送任务的结束节点和每一个等待护送任务的陪护人员节点为起始节点，分别建立到节点 U 的路径（即图的边）. 同时，分别以流出节点和可能与边 UV 相邻的下一项新护送任务的起始节点为终点，并以点 V 为起始节点建立相应的边. 最终可以建立从节点 U 到节点 V 的边 E.

基于当前时间，网络流中所有的边都会给出相应的费用成本. 赋予节点 V 一个足够高的价格，同时赋予节点 U 一个足够低的价格，则所有以节点 U 为终点（或以节点 V 为起始点）的边的有效费用成本均大于 0. 那么，在剩余网络图中（即不包括边 E），所有边的有效成本都是负数. 由节点 V 可达整个网络，既可以到达源点，又可以沿相应的路径到达每一个陪护人员的节点. 从源点，根据一个陪护人员的任务安排日程，就可以回溯至该陪护员的任一项工作的节点. 对于所有已经完成的工作源点（但不是预先设定的初始节点），可以

通过相邻的陪护人员的工作节点和路径到达.

应用上述方法，将从节点 V 到节点 W 的最小成本作为节点 W 的价格，并可以为所有的有效费用成本（包括在剩余网络图中由有效费用成本为 0 的边所构成的从节点 V 到节点 U 路径 P）设定一个负值. 如果边 E 的有效费用成本为非负，则求解结束；否则，通过环流 PE 来增加网络流量. 由此使得剩余网络图中各边的有效成本为负值，并改变了相应的流向，从而使所有的有效费用成本均不小于 0.

3. 寻求问题最优解的算法实现

首先注意到一个事实：实际中，一定存在一个能够按时将所有的 WP 护送至他们各自航班的调度方案. 下面考虑实际中更为普遍的问题.

设航空公司配有一定数量的服务陪护人员，他们每天按时到达各自的工作岗位（但实际中也不能保证某些人有特殊情况）. 同时要求陪护人员在执行一项护送任务中，在规定的时间到达规定的位置，完成此次护送任务. 因此，通过时间分配可以实现护送起始节点与结束节点的连通.

对于每一项护送任务，航空公司都会提前公布，并使得在候机厅任何位置的陪护人员都能按时到达护送工作的起始节点. 提前通知的时间下界等于通过候机厅最远两点所用的时间，例如在芝加哥奥黑尔机场需要 18 分钟，一般机场提前 30 ~ 60 分钟通知 WP 的到达时间就足够了. 在求解过程中做如下的约定：

（1）每个陪护人员都能从事相同的陪护工作.

（2）当一个陪护人员完成一项护送任务后，就相当于在确定的时间和地点增加了一个新的陪护人员.

（3）当一个陪护人员开始一项护送任务或离开工作区域时，就相当于在确定的时间和地点减少了一个陪护人员.

对于一项新的护送任务，求解算法将实现：

（1）要考虑所有正在执行护送工作的陪护人员和已到达机场开始上班（或已完成一项工作离开候机厅）的服务陪护人员，并包括所有已接受任务但还没有开始的护送任务的陪护人员，以确定哪些陪护人员可以完成这些新的护送任务.

（2）如果存在能够完成这些新的护送任务的陪护人员，则分别确定出哪名陪护人员去完成哪项护送任务，并通知相关的陪护人员去开始完成所分配的护送任务；如果不存在，则不能完成这些新的护送任务.

事实上，当且仅当在问题存在解决方案的情况下，该算法是有效的，并且运用该算法所得到的解决方案，则保证既不会出现延误，也不会无故中止任何一项护送任务.

5.6.4 模型的模拟检验

这里我们利用模型设计的求解算法和贪婪算法（即最邻近任务分配算法）分别对问题进行模拟检验. 贪婪算法主要是按照下面4个原则来分配护送任务：

（1）当有一项新的护送任务时，将其分配给最邻近的空闲的陪护人员.

（2）当一个陪护人员完成一项护送任务时，将其安排至下一项最邻近的护送任务.

（3）不能中止任何正在执行中的护送任务.

（4）除非是有新的护送任务，陪护人员只能原地待命.

对于原则（1）和（2），在一个WP到达机场之前的某一确定时刻，才开始对相应的陪护任务进行分配. 这个固定的时刻也就是相对于WP到达时刻的提前量，即可以取穿越机场候机厅中最远两个节点所需的时间.

1. 模拟检验的实现

在这里分别针对三种类型的机场[1]（如图5-15所示）进行模拟：

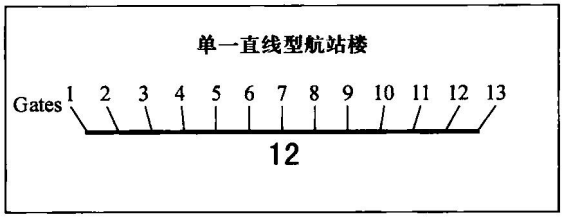
- 机场1：单一直线型航站楼（小型）.
- 机场2：波士顿洛根机场A航站楼（中型），由一条地下通道连接两个机场大厅.
- 机场3：芝加哥奥黑尔3号航站楼（大型），候机厅包括G，H，K和L.

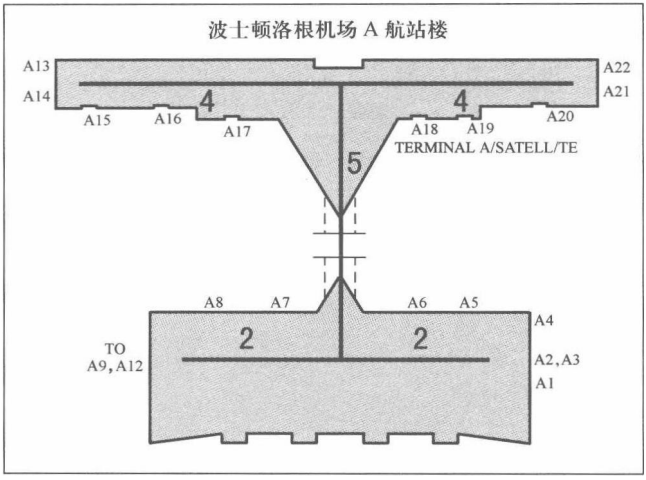
每一次模拟都是针对8个小时的工作时间，时间增量为1分钟. 对于每个机场和不同数量的陪护人员，在具有相同数量的WP及相同到达规律的情况下，分别进行了10次的模拟.

假设95%的WP都是提前1小时预约，其余5%的WP只提前5分钟预约，并将产生相应的累积损失. 我们分别针对护送任务繁忙和非繁忙两种情况对每一个机场进行模拟. 当护送任务繁忙时，在模拟连续工作的8小时中，3个机场分别有84 155和550个WP到达机场并需要陪护；当护送任务非繁忙时，不妨设各机场的WP人数分别是繁忙时的一半.

2. 模拟检验的结果分析

我们针对护送任务非繁忙和繁忙的两种情况进行了模拟，并用模型算法和贪婪算法进行了模拟比较，模拟结果如图5-16和图5-17所示[1]. 每个数据点表示陪护人员的数量，结果取10次模拟的平均值，平均等待时间等于总的等待时间除以WP的数量. 误机的WP不影响平均等待时间，即便他们的等待时间可能是一整天. 可以看出图中的曲线较为理想，但是左下角部分相对于x轴和原点存在一些偏差，这是因为一些特殊的WP被完全忽略所致.





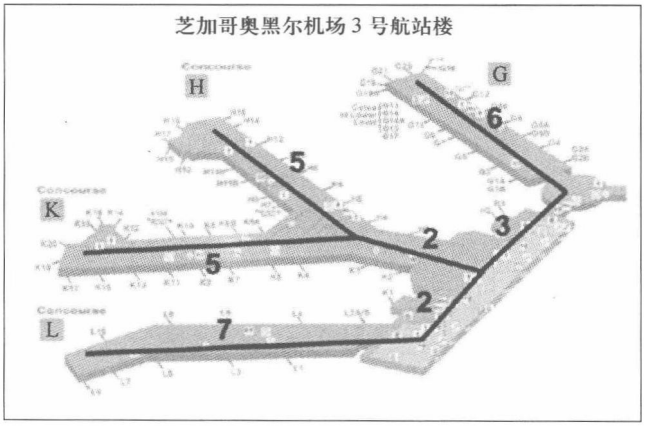
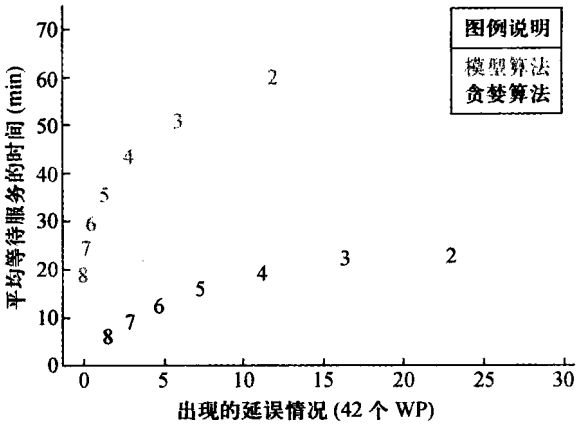


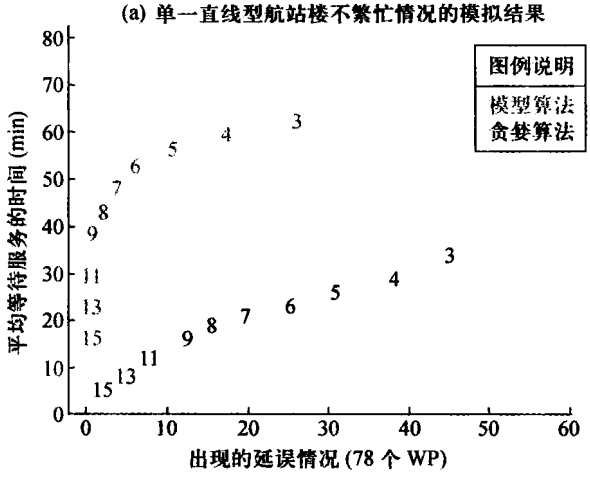
图 5-15 三种类型机场航站楼示意图

(图中每条边上的数字表示最远点的行走时间)

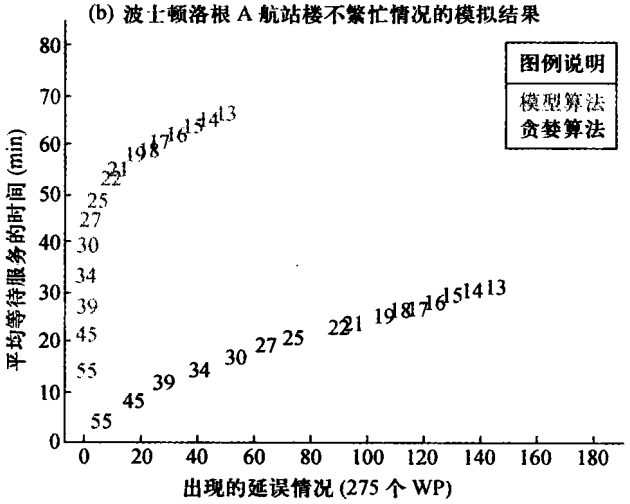
从模拟结果看出，采用该模型设计的算法求解得到的结果几乎接近于实际理论上的理想结果. 虽然基于最近邻任务分配的贪婪算法的平均等待时间较短，但是它将导致护送任务完成率较低，并且延误时间和任务未完成率大致成线性关系. 要想利用贪婪算法得到理想的结果，就需要更多的服务陪护人员来保证.



(a) 单一直线型航站楼不繁忙情况的模拟结果

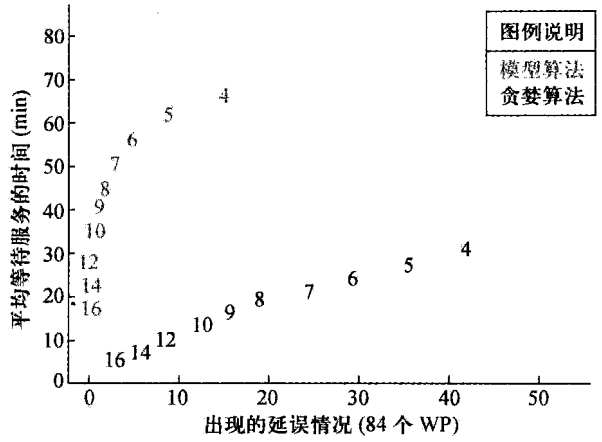


(b) 波士顿洛根 A 航站楼不繁忙情况的模拟结果

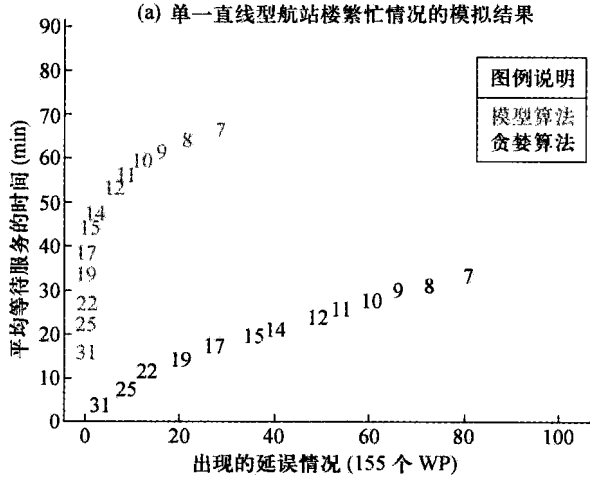


(c) 芝加哥奥黑尔机场 3 号航站楼非繁忙期的模拟结果

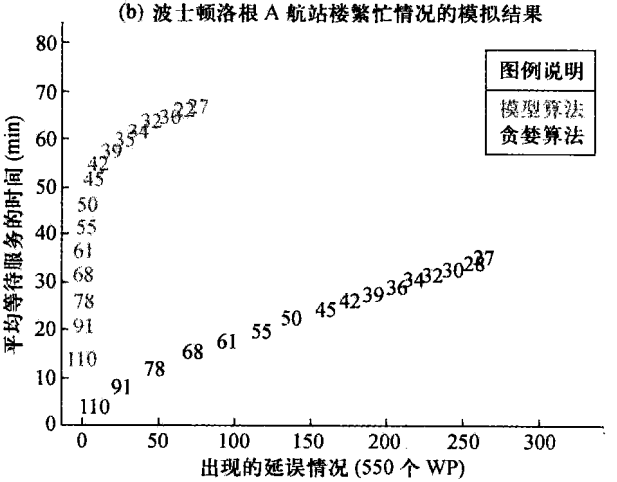
图 5-16 工作不繁忙时的仿真结果



(a) 单一直线型航站楼繁忙情况的模拟结果



(b) 波士顿洛根 A 航站楼繁忙情况的模拟结果



(c) 芝加哥奥黑尔机场 3 号航站楼繁忙期的模拟结果

图 5-17 工作繁忙时的仿真结果

同时，可以对陪护人员的服务等级做出评价. 较好服务等级是指在任何情况下，不能及时得到陪护服务的 WP 数量不超过 1；优良服务等级是指每个 WP 均会得到陪护服务，并且平均等待时间不超过 15 分钟. 表 5-3 给出了为满足不同服务等级所需要的陪护人员数量的最小值. 在同等条件下，要求提高服务质量等级，所需要增加服务陪护人员.

表 5-3 不同服务等级所需的陪护人员最小数量

机场	任务	WP 数量（人）	陪护人员数量（人）	
			较好服务等级	优良服务等级
机场 1（小型）	非繁忙	42	6	9
	繁忙	84	10	18
机场 2（中型）	非繁忙	78	9	17
	繁忙	155	15	31
机场 3（大型）	非繁忙	275	27	56
	繁忙	550	47	106

对于每一个机场，随着 WP 数量的增加，陪护人员会均匀地分散在候机楼的各个节点上，因此对于未事前预订的新的护送任务，总会有最邻近的陪护人员去完成.

3. 模型参数的灵敏度分析

（1）关于没有预订 WP 比例参数的灵敏度分析. 这里所说的没有预订临时通知的情况是指在 WP 到达机场前 5 分钟才通知相应的陪护人员. 针对机场 2 在护送任务繁忙的条件下，分别设定 WP 人数占乘客总人数的比例从 0% 到 100% 变化. 计算结果表明模型算法对这个参数的变化非常不敏感，这是由于 5 分钟就足以完成很多的护送任务.

（2）关于机场几何结构参数的灵敏度分析. 有 155 个或 550 个 WP，临时通知比例为 5%，对于具有不同几何结构的机场，该模型算法的仿真结果相当好. 事实上，机场的几何结构主要影响的是从完成一项护送工作到开始另一项护送工作的转移时间，而我们总是希望这个时间值尽量小. 这里仿真模拟机场的最大直线距离分别需时 11 分钟、12 分钟和 18 分钟（以时间为单位）.

（3）关于机场护送任务载荷参数的灵敏度分析. 机场护送任务载荷是指机场所能接受的需要护送服务 WP 的额定数量，它主要反映的是机场的繁忙程度，载荷参数越大，机场越繁忙. 运算结果表明机场载荷的敏感程度是一个很重要的参考指标，因为该数值直接影响到航空公司可服务 WP 合理的人数. 我们分别考虑机场载荷为标准载荷的 50% 和 150% 两种情况，基于机场 2 进行仿真. 由仿真结果可知，对于较好和优良两个服务等级要求，服务陪护人员的数量与实际载荷数基本成线性关系（参见表 5-3）. 为此，对繁忙期所需要的服

务陪护人员数 y 和 WP 的数量 x 的关系做线性拟合，有

要求优良等级：$y=0.19x+1.89$；

要求较好等级：$y=0.065x+4.79$.

拟合曲线如图 5-18 所示[1].

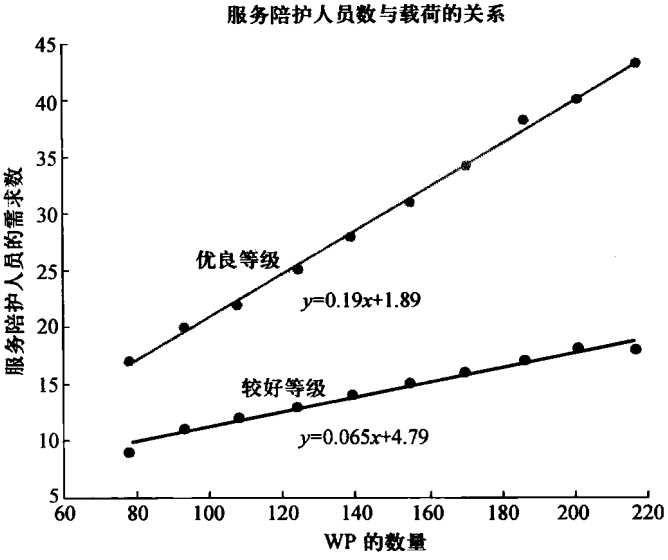


图 5-18 陪护人员数量与 WP 数量的变化关系

5.6.5 模型的评价

1. 模型的评价

该模型运用网络流理论，将机场轮椅和服务陪护人员的调度问题转化为最小费用流问题来解决，分别应用了价格函数与 Dijkstra 算法的相关知识，提出了可行有效的改进算法，经过多次有效的模拟仿真运行，得到了实际的结论，即为了寻求出无延迟服务的调度方案（如果存在的话），航空公司需要提前通知服务陪护人员 WP 到达的时间，并且需要提前的时间长度等于通过候机楼里距离最远的两个节点所需要的时间.

针对 3 种不同类型的机场，即单一线型机场候机楼（小型）、波士顿洛根机场候机楼（中型）和芝加哥奥黑尔 3 号机场候机数（大型）. 分别用模型算法进行了仿真，并与传统的贪婪算法（即基于最近邻任务分配算法）的计算结果进行了比较，结果表明，该模型的算法优于贪婪算法，并且只需要较少的服务陪护人员就可以保证所有 WP 都能得到服务.

基于模型所提出的算法，航空公司在提前 1 小时通知 WP 到达的时间情况下，WP 的平均等待时间是最少的，实际上这也是最优的调度方案. 此时，考虑航空公司仅提前 5 分钟通知 WP 到达时间的情况，也可以最小的延误来为 WP 提

供陪护服务. 我们对陪护服务的要求引入了两个服务等级：较好服务和优良服务. 模拟结果表明每个服务等级护送人员的数量与 WP 数目成线性关系.

通过对模型的模拟分析可知，航空公司提前 1 个小时得知 WP 到达时间是绰绰有余的. 相比于贪婪算法，该模型算法的可靠性更强、更实用.

对航空公司来说，如何确定最少的陪护人员，又如何制定他们一天最优的工作流程，以降低相应服务管理费用和成本. 那么，文中所提出的模型算法适用于实际中的各种情况，包括之前已知和未知 WP 抵达时间的情况. 该模型最主要的特点是将该问题转化为一个最小费用流的问题，将服务陪护人员每天的日程安排视为网络流.

2. 模型的求解算法评价

假设共有 m 个陪护人员，k 项护送工作，则该算法的运行时间为 $O(mk)$，开始或完成护送工作的时间为 $O(1)$. 那么求解一项新的护送工作时间的确定取决于 Dijkstra 算法的运行时间，并且有相应的一个界为 $O((m+k)^2\log(m+k))$.

（1）算法的优点

一是基于系统的网络流的理论知识，模型算法是在传统的 Dijkstra 算法的基础上改进而成. 二是对于临时通知 WP 到达的情况，该算法的运行结果也是比较好的，而对于提前 1 小时通知的情况，其运行结果就更为理想了. 即使是新增加一些未曾预知的 WP，该算法也能得到较好的结果. 三是假如具有足够多的提前通知 WP 到达的时间，运用该算法进行护送服务方案的安排，可以实现无延迟服务. 然而在实际中，当延误时间低于一定值的时候，航空公司就应考虑停止增加陪护人员，以有效地降低费用成本，而不是等到完全不存在延误的时候才考虑停止. 四是算法的运行效率很高，只需要占用较少的计算机资源.

（2）算法的不足之处

一是该算法必须需要计算机来处理，就可能相应地增加一些费用. 二是需要明确每一项护送任务的结束时间，并为等待任务的陪护人员进行未来护送任务的分配，即必须要设置一个护送工作的结束时间. 如果结束时间不确定，这个问题就变成了哈密顿路径问题. 三是该算法仅使用了现有的理论知识和方法，而没有基于实际数据的统计推算. 从某种角度来讲，对于非数学专业人员，可能难以理解该算法的描述.

5.7 问题的综合分析与进一步研究

5.7.1 问题的综合分析

1. 基于多候机厅的轮椅配备模型

该模型主要是根据有多个候机厅和每个候机厅包括 2 ~ 50 个登机口的不同

规模机场情况，利用1996—2005年10年间的WP占总乘客人数的比率和WP事先没有申请预约的统计概率，以及航空公司服务陪护人员和轮椅的运行成本等数据，建立了多候机厅的最小损失模型. 在模型中除了考虑航空公司的轮椅陪护服务成本以外，还考虑了因轮椅设置管理不合理所造成的飞机延误. 该模型通过多个航空公司的实际数据统计计算出航班延误所产生的效应，在尽量减少运行成本和延误损失效应的前提下，经过随机模拟找出了航空公司最佳的轮椅设置和三种不同方式的运行策略，并分别对不同策略适用情况做了比较分析. 该模型简单易行，考虑全面，依据实际数据的统计分析，注意到了各种不同情况、不同方案的比较分析，针对性、实用性强，可操作性强，这些解决处理问题的思想方法都值得借鉴和学习.

2. 基于航空公司间竞争的轮椅配置模型

该模型主要是考虑航空公司因为轮椅的配置和服务陪护质量所产生效应，而引起的航空公司客流的变化所导致的竞争问题. 因为航空公司的经济效益主要取决于客流量的变化. 实际中，航空公司的轮椅配置合理、陪护服务到位，顾客的满意度增加，客源就会增加. 相反的，客源不仅不会增加，甚至因为对航空公司的服务不满意，已有客源会流失，更会影响到潜在的客源. 为此，在该模型中分别研究了航空公司的占有率模型（分析了影响航空公司客源占有率的因素和关系）、航空公司客源的流失模型（分析了影响客源流失的因素和关系）以及航空公司的利润模型（分析了影响航空公司利润的因素和关系），最后给出了基于公司最大利润的轮椅配置模型和策略. 主要是把航空公司的不合理的轮椅配置和对乘客不满意的陪护服务，都视为对乘客做的无用功，将其转换为了一种乘客的“流失”，这关系到市场竞争中航空公司的服务质量. 航空公司在制定规划时，特别要重视这个问题，无论是对乘客的长期的影响因素，还是短期的影响因素，都会影响到航空公司未来市场占有率的变化.

3. 基于最小成本费用的轮椅配备和陪同服务模型

该模型主要是在充分分析航空公司实际运作情况的基础上，航空公司向WP提供轮椅及陪同人员的服务所产生的费用主要有以下几个方面：购置轮椅的费用、轮椅的维护保养费用、轮椅在机场的存储费用和陪同人员的薪酬等. 因此，航空公司在选择轮椅的配置策略时，应当充分考虑这些费用与WP需求的关系，在尽量满足WP需求的条件下，力求所需费用最小化. 该模型从实际出发，首先建立了相关单个因素的费用模型，包括航班延误费用、WP等待费用、服务陪护人员的费用、轮椅的成本费用和服务陪护人员兼顾其他工作所产生的效益等模型. 然后综合考虑航空公司在轮椅的配置和陪护服务策略执行过程中总费用模型，包括航班延误和乘客等待的总费用、一个班次陪护的总费用和航空公司一个年度的总成本费用模型. 分别就一个年度中轻度繁忙、中度繁忙和重度繁忙三个时段的情况计算出了总成本费用模型. 最后给出了模型求解

的算法和过程，并分别就约翰肯尼迪国际机场、芝加哥的奥黑尔国际机场和亚特兰大国际机场的具体情况，给出了仿真模拟求解和分析. 模型切合实际、考虑全面、算法简单、易于操作、结果合理、实用性强.

4. 基于成本效益的轮椅配备与陪护服务的仿真模型

该模型是在考虑航空公司的最小成本和最大效益情况下，对轮椅配置和陪护服务的模型进行了仿真. 模型主要针对两个问题进行仿真和分析，即航空公司在一个机场应该配置多少轮椅和服务陪护人员？这些轮椅和服务陪护人员如何调度？首先模拟了机场设施的布局，根据需求模拟计算出了合理的轮椅和服务陪护人员的配置数量，制订出 WP 需求计划表和服务陪护人员的调度计划表. 通过模拟仿真计算出航空公司最小的运行成本，从而确定出最优的配置和调度方案. 并分别就不同候机厅、不同机场的具体情况模拟检验了模型的性能和方案的最优性. 同时，针对未来的人口变化预测分析了 WP 需求情况，说明方案的调整变化对策. 该模型最突出的特点是切合问题的实际，有针对性地仿真计算得出问题的结果，并对多参数的灵敏度进行了分析检验，仿真结果充分说明了模型算法通用性很强，可以适用实际中的不同机场、不同候机厅、不同登机口、不同布局的各种情况.

5. 基于最小费用流的航空轮椅服务模型

该模型主要是根据问题的实际情况和要求，将问题转化成一个网络流的优化问题. 基于网络优化的最小费用理论来研究航空轮椅的配置和陪护服务问题，即将服务陪护人员和 WP 所在的位置视为网络图的节点，任意两个节点之间的成本费用视为网络图的边，则构成了一个赋权网络图. 模型在合理的假设之下，首先给出了该网络的最小费用最大流的有关概念界定，接着分析航空公司运行成本费用的构成和机场设施的布局模型. 然后，对于经典的 Dijkstra 算法进行了改进，用来解决这个最小费用最大流问题. 针对小型、中型和大型三种航站楼，分别对于陪护任务繁忙期与非繁忙期两种情况进行了模拟求解，根据服务要求的不同等级分别得到了航空公司应该配备的轮椅和服务陪护人员的数. 模型的求解结果表明，模型所给出的方案完全与实际相符，模型的参数的灵敏度有很好的稳定性，算法的可靠性和实用性都很强. 最后，对该模型的计算复杂度和复杂度做了分析和评价.

5.7.2 进一步研究的问题

在这里所介绍的几个模型都是在一定的简化假设下，基于一些实际的数据和不同的理论基础建立了相应的数学模型，多数都是针对有代表性的机场和候机厅具体布局情况进行模拟求解，得到了较为合理的轮椅配置和服务陪护人员的调度方案. 虽然这些模型都较好地解决了问题，但从实际考虑各有优缺点，特别是对问题的简化和求解算法的设计还有待于进一步研究和探讨.

(1) 多数的数学模型及相应的求解算法，都只考虑了一天的陪护人员流动的问题，应用系统理论算法，模拟得到了较好的计算结果. 为此，需要进一步改进和完善模型算法，以降低提前预知 WP 到达的时间界限. 对于每个航班来说，在乘客办理登机手续时，核实 WP 需求情况，并通知航空公司. 如果能够利用这个时间做到机场的安排陪护日程，则将会大大减少 WP 等待的时间，也能减少航班延误的可能性，可以有效降低航空公司的服务成本，从而提高效益，增加乘客的满意度.

(2) 如果能够更多地掌握航空公司过去一段时间内 WP 的精确数据，依据这些数据预测未来的相关数据，则可以更有针对性地制定出合理的轮椅配置和服务陪护人员的调度方案，使得航空公司的工作计划安排会更科学，更有计划性，也能够提高航空公司的服务质量，以满足乘客的需求.

(3) 多数的模型是以航空公司的最小运行成本为方案设计的主要目标，当然也考虑到了顾客等待的时间因素，但从实际出发还是不够的. 事实上，一个航空公司在一个机场或一个候机楼配置轮椅和服务陪护人员数据与调度方案的合理性，不应该取决于航空公司当时的运行成本最小，而更多地应关注乘客的满意度如何. 因为这直接影响着航空公司的服务质量，即潜在的客流量变化. 服务质量高，乘客的满意度就高，可以吸引更多的乘客，自然可以增加航空公司的效益；相反的就会出现客流量的流失，使航空公司的效益减少. 所以，如果更多地关注乘客满意度对轮椅配置和调度方案影响，则相应的模型会更有实际意义，实用性会更强.

(4) 航空公司在确定陪护人员的数量时应当考虑服务质量的因素，往往服务质量与陪护人员的数量有直接的关系. 如果服务陪护人员数量过少，虽然降低了航空公司的服务成本，但可能会出现服务质量问题，导致航空公司更大的经济损失；如果服务陪护人员数量多了，必然会增加航空公司的运行成本，但会提高服务质量，从某种程度上可以增加航空公司的效益. 因此，如何在二者之间找到一个最佳的方案，这是一个较复杂的多目标优化问题. 如果能够结合实际航空公司和机场的具体情况，找到一个切合实际的平衡方案，则将是最好的结果.

(5) 实际中的航班受很多不确定因素的影响，如天气因素、航空调度因素、各种突发事件因素等，所以在研究这个问题时应充分考虑实际工作中的不确定性和随机因素的影响，这将会产生一系列的随机问题，采用随机性的方法来处理应该有效的，但问题本身会变得复杂，也可以侧重于考虑实际中的一个方面来研究.

这里给出一种相对简单的解决问题的思路：首先建立航班的流量和登机口进出乘客的随机概率模型，用于确定乘客中需要轮椅和陪护服务的乘客数量，并在候机厅中将其护送到转机的登机口. 然后设计一套随机经济计量模型，用

来度量航空公司的短期和长期运行成本，成本主要包括服务陪护人员的薪酬、轮椅的维护与储存成本以及当航班延误时，后期陪护费用和支付乘客因航班延误所发生费用的总和. 接着用模拟退火启发式算法（SA）对所建立的数学模型进行模拟求解，模型所给出的方案是以降低陪护成本费用为目标，通过优化调度确定出航空公司服务陪护人员的数量. 试图在搜索空间中寻求问题的全局最优解，事实上，因为空间过大，难以找到全局最优解. 模拟求解结果表明，这种算法法是能够很好地逼近这个问题的全局最优解的. 针对不同的交通流量、不同的机场规模和不同的需要轮椅与陪护服务的乘客比例情况下，利用一个机场和航空公司的实际数据，通过模拟服务陪护人员在机场内动态护送服务的过程，分别计算出短期和长期两个方面的运行成本费用.

考虑实际中的任何一个随机因素的影响来研究这个问题都是有益的.

（6）国内能够提供轮椅服务目前仅限于几家较大的航空公司，而且也只在像北京、上海、广州等规模较大的机场内，但都必须提前预订申请方可提供服务. 从这个角度来说，与这个问题相关的研究课题还有很多，潜在的应用是很大的，结合一个航空公司和具体的机场设施有针对性地研究轮椅的合理配置、服务陪护人员的数量和合理调度等问题都有潜在的应用价值.

参考文献

[1] COMAP, Inc. The 2006 Mathemalical Contest in Modeling [J]. The UMAP Journal, 2006, 27 (3): 333-414.

[2] 吴德立. 关于完善机场轮椅服务的提案 [OL]. http://news.qq.com/a/20100211/001031.htm, 2010.

[3] Conway, Bernie. Development of a VR system for assessing wheelchair access [OL]. http://www.fp.rdg.ac.uk/equal/Launch_Posters/equalslidesconway1/sld001.htm,2001.

[4] Gross, Ralph, and Jianbo Shi. The CMU Motion of Body (MoBo) Database [M]. Pittsburgh, PA: Robotics Institute, Carnegie-Mellon University, 2001.

[5] Bureau of Labor Statistics. Occupational employment statistics, November [OL]. http://www.bls.gov/oes/current/oes396032.htm, 2004.

[6] Transport Wheelchairs. http://www.1-800-Wheelchair.com,2005.

[7] U.S. Census Bureau. 2005 Population projections [OL]. http://www.census.gov/.2005.

[8] Airline pricing data [OL]. http://www.expedia.com, 2005.

[9] Southwest Airlines. 2005 Flight Data: MAF, CMH, STL [OL]. http://www.southwest.com,2006.

[10] Backman, C., E. Di Mauro, L. Scheuring, and S. Wyss. Managing service quality: Taking care of handicapped passengers at Arlanda Airport [M]. Uppsala, Sweden: Uppsala University, 2004.

[11] Federal Aviation Administration, Bureau of Transportation Statistics. Aircarrier flight delays

and cancellations [OL]. http://www. oig. dot. gov/StreamFile? file =/data/pdfdocs/cr2000112. pdf, 2000.

[12] Mueller, Eric R., and Gano B. Chatterji. Analysis of aircraft arrival and departure delay characteristics, AIAA 2002 - 5866 [M]. American Institute of Aeronautics and Astronautics: AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical 1-3 October 2002, Los Angeles, California, 2002.

[13] Wikipedia, the Free Encyclopedia. Diminishing returns [OL]. http://en. wikipedia. org/wiki/ Diminishing_returns. Accessed5Feb2006, 2006.

[14] City of Chicago. Monthly operations, passengers, cargo summary by class: City of Chicago Airport activity statistics [OL]. http://www. flychicago. com/doa/stats/205SUMMARY. pdf, 2005.

[15] Schumacher, B. Proactive flight schedule evaluation at Delta Air Lines [M]. Winter Simulation Conference, INFORMS Simulation Society, 1999.

[16] Delta Airlines. Flight status and updates [OL]. http://www. delta. com/traveling_checkin/ flight_status_updates/index. jsp, 2006.

[17] Department of Health and Human Services, Administration on Aging. AoA statistics-Aging into the 21st century [OL]. http://www. aoa. dhhs. gov/prof/Statistics/future_growth/aging21/demography. asp, 2006.

[18] Haseltine Systems Corporation. Haseltine Systems Corporation [OL]. http://www. haseltine. com/data. html, 2006.

[19] City of Atlanta. Atlanta international airport flight information [OL]. http://www. atlanta-airport. com, 2006.

[20] Neufville, Richard De. Airport Systems Planning. Cambridge [M], MA: MIT Press, 1976.

[21] Friedman, S. Morgan. Inflation calculator [OL]. http//: www. westegg. com/inflation/, 2006.

[22] Avjobs. com. Aviation career salary ranges [OL]. http://www. avjobs. com/table/airsalry. asp, 2006.

第6章　艾滋病毒防控资源分配问题

HIV/AIDS（人体免疫缺损病毒/艾滋病）防控资源配置问题是2006年ICM赛题，研究的是各国如何合理配置各种HIV/AIDS防控资源，实现最大程度上控制HIV/AIDS的问题. 这也是生物医学、公共卫生和社会事业领域一个重要的研究课题.

6.1　问题的综述

6.1.1　问题的提出

HIV/AIDS大范围流行至今已经25年了，这种疾病导致的感染人数和死亡人数不断攀升. 人们在抗HIV/AIDS方面付出了巨大的努力，也取得了一定的成果，但目前国际社会对于如何有效地配置各种资源以更有效地防控HIV/AIDS，相关方面的研究仍然欠缺.

如果你们是联合国的一个专家组，请就怎样管理可利用的HIV/AIDS防控资源问题，向联合国提出合理建议. 你们需要对目前大家关注的几种防控方案进行建模，并用你们的模型就资金分配问题给出合理建议.

问题（1）：从各大洲（非洲、亚洲、欧洲、北美洲、澳洲和南美洲）中，分别选择一个你们认为HIV/AIDS流行程度最严重的国家. 建立数学模型粗略估计这些国家在没有任何有效干预措施时，2006—2050年HIV/AIDS感染人数的变化规律. 详细描述你们的模型并准确给出模型的基础假设. 另外，解释一下你们选择这些国家的原因.

问题附件中的电子表格提供了世界卫生组织（WHO）成员国家在相关领域2003年前的数据，你们的模型中可能要用到这些数据.

数据文件：list_WHO_member_states.xls，给出了WHO成员国家列表.

数据文件：hiv_aids_data.xls，给出了HIV/AIDS相关数据表.

（1）Global HIV-AIDS cases，1999，全球HIV/AIDS病例汇总（1999年）. 该数据来自联合国HIV/AIDS项目研究组（UNAIDS），给出了截止1999年底，

各国提供的0～49岁HIV阳性的大致人数.

(2) HIV-AIDS in Africa over time，非洲HIV/AIDS流行情况. 该数据来自美国政府，给出了非洲一些城市育龄妇女零碎的艾滋病发病率时间序列统计数据.

(3) HIV-AIDS subtypes，HIV/AIDS子类型表格. 该数据来自UNAIDS，表中按国家划分，给出了HIV-1子类型的地理分布.

附件还提供了这些国家的人口统计数据. 这些数据都来自联合国，包括了世界主要地区和国家1995—2050年的人口统计数据，其中1995—2005年的数据为统计估计数据，2006—2050年的数据为按照中等生育率假设下的预测数据.

相关的数据文件如下：

(1) fertility_data.xls，给出了生育率数据，指特定年龄妇女的生育率（每千名妇女生育子女的数目).

(2) population_data.xls，给出了人口统计表（分性别统计，单位：千人).

(3) age_data.xls，给出了年龄分布统计数据，每5年划分一个年龄组，统计各个年龄组的人数.

(4) birth_rate_data.xls，给出了出生率统计数据，单位：出生人数/千人·年.

(5) life_expectancy_0_data.xls，给出了人均寿命统计数据，单位：年.

HIV/AIDS防控资金主要以两种方式介入：预防介入和治疗介入. 预防介入主要包括：健康咨询、HIV检测服务、避孕套推广、以学校为基础的艾滋病教育、防止母-婴传染的药物等. 治疗介入包括：治疗其他性传染病、控制高致病感染等. 你们的工作要集中于两种最主要的介入方式：提供抗逆转录酶病毒（Antiretroviral，ARV）药物治疗和提供HIV/AIDS预防疫苗.

问题(2)：落后国家难以承担本国的AIDS防控资金，在UNAIDS的倡导下，发达国家建立了多项AIDS国际援助项目，用于帮助和支持落后国家开展AIDS防控工作.

对你们在问题(1)中选择的国家，估计这些国家在2006—2050年间期望得到多少国际援助资金用于AIDS防控.

用问题(1)中所建立的模型和这些资金数据，应用以下三种控制方案，分别估计所选国家在2006—2050年HIV/AIDS感染人数预期的变化率.

(1) ARV药物治疗.

(2) 预防性HIV/AIDS疫苗.

(3) 同时采用ARV治疗和预防性HIV/AIDS疫苗.

假设上述三种方案都不会有产生HIV抗药菌株的风险（问题3中需要考

察该问题). 请务必详细、准确地叙述你们模型的基础假设.

你们可以考虑是否这些方案适用于问题 (1) 选择的所有国家, 或者特定子类型的国家, 综合考虑国民收入水平和医疗负担等. 附件中提供了有关各国收入水平的表格, 或许可以为你的模型所用.

数据文件: income_data.xls, 给出了世界国民收入表.

该数据来自世界银行 (2002), 表中给出各地区国民人均生产总值 (GNP) 以及全球的收入分布. 你们可以在你们的模型中随意地使用这些数据.

ARV 药物治疗可以延长 HIV/AIDS 感染者的生命, 给 HIV/AIDS 感染者带来了极大的福音. 在发达国家, ARV 药物使用率很高, 延长了大量 HIV/AIDS 感染者的生命. 有关部门和国际组织一直致力于如何拓展落后国家的 HIV/AIDS 感染者获得 ARV 药物的渠道. 经济落后国家的卫生保健预算非常有限, 不可能仅仅依靠自己对大多数 HIV/AIDS 感染者开展 ARV 治疗. 附件 1 的数据来自 UNAIDS, 这些数据给出了当前正在使用 ARV 药物治疗的国家的相关数据.

ARV 的疗效在很大程度上依赖于感染者能否严格坚持治疗, 并且定期观察疗效. 公认有效的 ARV 治疗条件是: 全面的咨询和医生护理治疗计划、定期检查以观察病情的发展和感染人群的发病情况. 间断的治疗或者不恰当的治疗方法会导致严重的后果: 一是可能没有治疗效果; 二是可能会直接导致产生 HIV 抗药性菌株.

ARV 药物的制药厂商一直在努力降低成本从而降低价格, 虽然用于治疗 HIV/AIDS 早期感染药物的价格已经降到每个病人每年只需几百美元, 但如何分发这些药物, 并提供必要的辅助医疗保健措施以及进一步的治疗, 将考验政府的行政管理和财政能力. 经过比较, 短疗程治疗方案产生抗药菌株的可能性最小, 采用该方法的 ARV 药物成本可以控制在每人每年不超过 1100 美元.

针对预防性 HIV 疫苗问题, 请对以下参数给出你们认为合理的估计, 除这些情况以外, 还可以考虑其他情况.

(1) 预防性 HIV/AIDS 疫苗可以投入实用的时间;

(2) 疫苗接种普及速度:

① 如果你们希望使新生儿得到免疫, 那么假设该国新生儿 HIV/AIDS 疫苗的稳定接种水平和 WHO (2002) 报告的 DTP3 (白喉-百日咳-破伤风疫苗第三剂) 的接种率相同.

数据文件: vaccination_rate_data.xls, 给出了新生儿疫苗接种率数据.

② 如果你们希望使成人 (5 岁以上的人群) 得到免疫, 那么假设这部分人 HIV/AIDS 疫苗的稳定接种水平与 WHO (2002) 报告的 TT2 (破伤风类第二剂) 的免疫率相同.

数据文件：vaccination_rate_data.xls，给出了成人疫苗接种率数据.

（3）疫苗的功效及其有效期；

（4）是否存在来自疫苗接种的流行病学的外部偶然因素；

（5）假设疫苗共三剂，可以放进疫苗标准包，随 WHO 的免疫扩展计划（EPI）一起，以额外 0.75 美元的增加成本发放.

问题（3）：在如下关于 ARV 抗药性菌株的假设前提下，重新阐明你们在问题（2）中建立的 3 个模型.

当前的研究表明，采用 ARV 治疗的病人若坚持程度低于 90% ~95%，有产生抗药性菌株的“巨大风险”. 在你们的分析中可以采用假设：接受 ARV 治疗的病人如果坚持程度低于 90%，就有 5% 的可能性产生抗药性.

人们也可以使用二线和三线药物治疗. 但在你们的模型中应假设，除欧洲、日本和美国以外的国家要使用这些药物将会因为过于昂贵而无法负担.

问题（4）：向联合国相关部门写一份建议书，就如下几个方面给出你们的建议：

（1）在抗 HIV/AIDS 中，就 ARV 药物的供应和 HIV 疫苗的资源分配问题提出合理建议；

（2）相对于其他外交政策而言，HIV/AIDS 防控作为一项国际事务，应该赋予哪些优先权？

（3）针对如何合理配置 HIV/AIDS 国际援助资金问题给出建议.

对于第（1）个方面，假设从现在起到 2010 年，可利用的财政资源将专门划拨经费用以加速 HIV 疫苗的研制工作——或直接资助疫苗的研发，或通过其他机制来实现这个目的. 这些投入将促进问题（2）中疫苗研制成功日期的提前.

问题的原文如下：

As the HIV/AIDS pandemic enters its 25th year, both the number of infections and number of deaths due to the disease continue to rise. Despite an enormous amount of effort, our global society remains uncertain on how to most effectively allocate resources to fight this epidemic.

You are a team of analysts advising the United Nations (UN) on how to manage the available resources for addressing HIV/AIDS. Your job is to model several scenarios of interest and to use your models to recommend the allocation of financial resources. The narrative below provides some background information, and outlines specific tasks.

Task1: For each of the continents (Africa, Asia, Europe, North America, Australia, and South America), choose the country you believe to be most critical in terms of HIV/AIDS. Build a model to approximate the expected rate of change in

the number of HIV/AIDS infections for these countries from 2006 to 2050, in the absence of any additional interventions. Fully explain your model and the assumptions that underlie your model. In addition, explain how you selected the countries to model.

Use as a list of countries for inclusion in your analysis the countries included in the attached spreadsheet, which include all member states of the World Health Organization (WHO) as of 2003.

Data: "*list_WHO_member_states. xls*".

Reliable data on HIV prevalence rates by county are generally difficult to obtain. The attached spreadsheet includes several worksheets of data which you may use in your analysis.

Data: "*hiv_aids_data. xls*".

a. "Global HIV- AIDS cases, 1999": These data come from UNAIDS (the Joint United Nations Program on HIV/AIDS) and report the estimated number of HIV positive 0 to 49 year olds by country at the end of 1999.

b. "HIV - AIDS in Africa over time": These data come from the US government and give some piecemeal time series data on measured HIV prevalence rates among women of childbearing age, in urban areas, over time for some African countries.

c. "HIV - AIDS subtypes": These data come from UNAIDS and give the geographic distribution of HIV-1 subtypes by country.

Also attached, for your use, are some basic population and demographic data.

Data:

(1) "fertility_data. xls": These data come from the UN and give age-specific fertility rates by major area, region, and country, 1995—2050 (births per thousand women)

a. Estimates for 1995—2005.

b. Projections (under the assumption of medium fertility levels) for 2005—2050.

(2) "population_data. xls": These data come from the UN and give total population (both sexes combined) by major area, region, and country, annually for 1950—2050 (thousands).

a. Estimates for 1950—2005.

b. Projections (under the assumption of medium fertility levels) for 2006—2050.

(3) "age_data. xls": These data come from the UN and give population (for

both sexes, and by gender) by five-year age groups, major area, region, and country, 1950—2050 (thousands).

a. Estimates, 1950—2005.

b. Projections (under the assumption of medium fertility levels) for 2010—2050.

(4) "birth_rate_data.xls": These data come from the UN and give crude birth rates by major area, region, and country, 1950—2050 (births per thousand population)

a. Estimates, 1950—2005.

b. Projections (under the assumption of medium fertility levels) for 2005—2050.

(5) "life_expectancy_0_data.xls": These data come from the UN and give life expectancy at birth (by sex and both sexes combined) by major area, region, and country, 1950—2050 (years).

a. Estimates, 1950—2005.

b. Projections (under the assumption of medium fertility levels) for 2005—2050.

There are a number of interventions that HIV/AIDS funding could be directed towards—including prevention interventions (voluntary counseling and testing, condom social marketing, school-based AIDS education, medicines to prevent mother-to-child transmission, etc.) and care interventions (treating other untreated sexually transmitted diseases, treating opportunistic infections, etc.). You should focus on only two potential interventions: provision of antiretroviral (ARV) drug therapies, and provision of a hypothetical HIV/AIDS preventative vaccine.

Task2: First, estimate the level of financial resources from foreign aid donors that you realistically expect to be available to address HIV/AIDS, by year, from 2006 to 2050, for the countries you selected in Task 1. Then use the model you developed in Task#1 and these estimates of financial resources to estimate the expected rate of change in the number of HIV/AIDS infections for your selected countries from 2006 to 2050 under realistic assumptions for the following three scenarios:

(1) Antiretroviral (ARV) drug therapy.

(2) Apreventative HIV/AIDS vaccine.

(3) Both ARV provision and a preventative HIV/AIDS vaccine.

Assume in these scenarios that there is no risk of emergence of drug-resistant strains of HIV (you will examine this issue in Task 3).

Be sure to carefully describe the assumptions that underlie your model.

You can choose whether these scenarios should be implemented for all of the countries you selected in Task 1, or for certain subsets of countries based on income cut-offs, disease burden, etc. Available for use if you wish is a spreadsheet of country-level income data.

Data: "*income_data. xls*".

These data are from the World Bank (2002) and give per-capita gross national product (GNP) data as well as broad income classifications that you are free to use in your analysis if you wish.

ARV drug therapies can have tremendous benefits in terms of prolonging the lives of individuals infected with HIV/AIDS. ARVs are keeping a high proportion of HIV/AIDS-infected individuals in rich countries alive, and policy makers and international institutions are facing tremendous political pressure to increase access to ARVs for individuals in poor countries. Health budgets in low-income countries are very limited, and it seems unlikely that poor countries will be able to successfully expand these programs to the majority of their populations using their own resources. Appendix 1 presents country-specific data from UNAIDS on current access to ARVs for a number of countries.

The efficacy of ARVs depends in large part on adherence to the treatment regimen and to proper monitoring. The most favorable conditions for ARVs are structured programs with extensive counseling and physician care, as well as regular testing to monitor for disease progression and the onset of opportunistic infections. Non - adherence or inadequate treatment carries with it two very serious consequences. First, the treatment may not be effective for the individual undergoing treatment. Second, partial or inadequate treatments are thought to directly lead to the emergence of drug-resistant strains of HIV.

The price of the drugs initially used to treat patients has come down to several hundred dollars a year per patient, but delivering them and providing the necessary accompanying medical care and further treatment is the key administrative and financial challenge. It is estimated that purchasing and delivering antiretroviral using the clinically - recommended approach (DOTS, or directly observed short course treatments) which is intended to minimize the emergence of drug - resistant strains would cost less than $ 1,100 per person per year. (Adams, Gregor et al. [2001]. "Consensus Statement on Antiretroviral Treatment for AIDS in Poor Countries," http://www. hsph. harvard. edu/bioethics/pdf/consensus_aids_therapy. pdf)

For a preventative HIV vaccine, make assumptions you feel are reasonable

about the following (in addition to other factors you may choose to include in your model):

(1) The year in which an HIV/AIDS preventative vaccine might be available.

(2) How quickly vaccination rates might reach the following steady-state levels of vaccination:

a. If you wish to immunize new cohorts (infants), assume the steady-state level for new cohorts of the country-by-country immunization rates for the third dose of the diphtheria-pertussis-tetanus vaccine (DTP3), as reported by the WHO (2002).

Data: "*vaccination_rate_data. xls*".

b. If you wish to immunize adults (any group over age 5), assume the steady-state level for older cohorts is the second dose of the tetanus toxoid (TT2) rate, as reported by the WHO (2002).

Data: "*vaccination_rate_data. xls*".

(3) The efficacy and duration of protection of the vaccine.

(4) Whether there would be epidemiological externalities from vaccination.

(5) Assume the vaccine is a three-dose vaccine, and can be added to the standard package of vaccines delivered under the WHO's Expanded Programme on Immunization (EPI) at an incremental cost of addition of $ 0.75.

Task3: Re-formulate the three models developed in Task 2, taking into consideration the following assumptions about the development of ARV-resistant disease strains.

Current estimates suggest that patients falling below 90-95 percent adherence to ARV treatment are at a "substantial risk" of producing drug resistant strains. Use as an assumption for your analysis that a person receiving ARV treatment with adherence below 90 percent has a 5 percent chance of producing a strain of HIV/AIDS which is resistant to standard first-line drug treatments.

Second-and third-line ARV drug therapies are available, but assume for your analysis that these drugs are prohibitively expensive to implement in countries outside of Europe, Japan, and the United States.

Task4: Write a white paper to the United Nations providing your team's recommendations on the following:

(1) Your recommendations for allocation of the resources available for HIV/AIDS among ARV provision and a preventative HIV vaccine.

(2) Your argument for how to weigh the importance of HIV/AIDS as an international concern relative to other foreign policy priorities.

(3) Your recommendations for how to coordinate donor involvement for HIV/AIDS.

For (1): assume that between now and 2010 the available financial resources could be allocated so as the speed the development of a preventative HIV vaccine-through directly-financing vaccine research and development (R&D), or through other mechanisms. Any gains from such spending would move the date of development you assumed in Task 2 to some earlier date.

6.1.2 2006 年 ICM 竞赛的相关情况

2006 年 ICM (第 8 届 ICM) 总计有来自 6 个国家 122 所大学的 224 个小组参加. 该问题的命题人是哈佛大学经济学在读博士 Heidi Willianms, 该问题来源于他在全球发展中心 (Center for Global Development, CGD) 工作时的课题. CGD 是一个非营利性的智库机构, 其研究领域主要集中在如何在低收入国家加速疫苗的推广和使用等问题.

在评阅论文时, 阅卷专家强调: 阅卷人往往会根据文章的摘要和前面几段文字决定是否继续读下去, 不是每篇文章都仔细地读完. 因此文章的摘要和问题分析必须认真对待. 摘要中应包括你考虑的关键要素和关键资源的配置取舍方案, 针对你的假设, 给出明确的结果, 以及该结果对于假设的敏感性分析. ARV 疗法对于降低 HIV/AIDS 传染性风险的影响程度和有效疫苗的投入时间是本问题的核心假设.

另外就是论文的长度问题, 并不是越长越好, 表述方式才是最重要的. 只有清晰地表述出你们的模型, 评阅人才能看明白, 才能对你们的模型进行评价. 此外, 数学建模应该选择最重要的因素, 使用适当方法进行建模, 而不是模型越复杂越好.

经过评阅专家的精心评阅, 共评选出 4 篇特等奖论文, 分别是:

(1) 美国杜克大学

论文题目: The United Nations and the Quest for the Holy Grail (of AIDS)

论文作者: Aaron Wise, Arnav Mehta, Qianwei Li

指导教师: David Kraines

(2) 美国杜克大学

论文题目: Managing the HIV/AIDS Pandemic: 2006—2055

论文作者: Tyler Huffman, Barry Wright, Charles Staats

指导教师: David Kraines

(3) 美国哈维穆德大学

论文题目: AIDS: Modeling a Global Crisis (and Australia)

论文作者: Cris Cecka, Michael Martin, Tristan Sharp

指导教师：Lisette dePillis

（4）美国西点军校

论文题目：The Spreading HIV/AIDS Problem

论文作者：Adam Seybert，David Ryan，Nicholas Ross

指导老师：Randal Hickman

获奖数量及比例如表 6-1 所示. 本章主要是以这 4 篇优秀论文为基础，对该问题进行分析讨论.

表 6-1 2006 年 ICM 获奖情况

	Outstanding	Meritorious	Honorable Mention	Successful Participation	总计
队数	4	48	114	58	224
百分比	1.8%	15.9%	54.3%	28%	100%

6.1.3 问题的背景资料[1]

人类免疫缺陷病毒（Human Immunodeficiency Virus，HIV）是一种感染人类免疫系统细胞的慢病毒（Lentivirus），属反转录病毒的一种，如图 6-1 所示. 普遍认为，艾滋病是后天性细胞免疫功能出现缺陷而导致严重随机感染及/或继发肿瘤并致命的一种疾病. HIV 把人体免疫系统中最重要的 T4 淋巴细胞作为攻击目标，大量吞噬、破坏 T4 淋巴细胞，从而破坏人的免疫系统，最终使免疫系统崩溃，使人体因丧失对各种疾病的抵抗能力而死亡.

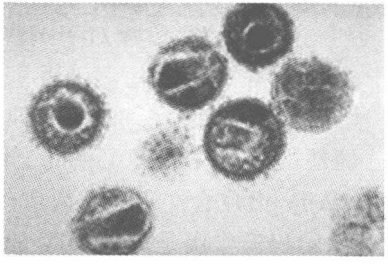


图 6-1 人类免疫缺陷病毒（HIV）

1. HIV 传播的特点

HIV 与常见传染病（例如流感、SARS 和霍乱等）的传播特性具有显著的不同.

首先，HIV 的传播途径具有特殊性，HIV 主要通过性传播、交叉使用毒品注射器传播和母婴传播，这三种传播方式决定了 HIV 的传播范围限定在一个“潜在传染人群”中.

其次，目前 HIV 是不可治愈的. 在研究感冒、麻疹等疾病时，除了易感

者和传染者以外，还存在治愈人群. 有些传染病的治愈人群具备病毒抗体(例如麻疹)，对该流行病产生免疫，不再可能感染该病毒. 而有些传染病的治愈人群无法形成病毒抗体，治愈后重新回归易感人群. 到目前为止，AIDS无法治愈，从感染到死亡期间都处于传染期.

再次，HIV传染期非常长. 常见传染病的传染期都在几天之内，例如流感一般在几天就能治愈，而后重新变为易感者. 而HIV的传染期长达十几年.

WHO给出了成年人感染HIV后的临床过程，主要包括3个阶段[2]，如图6-2所示.

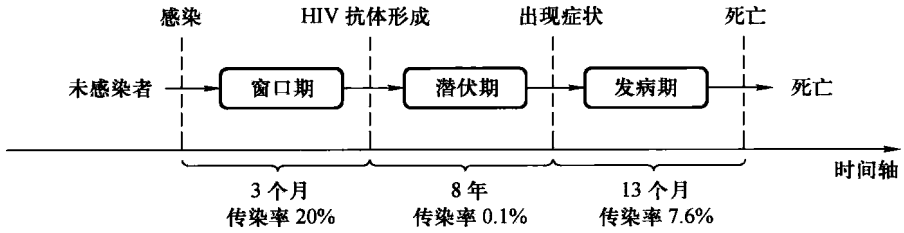


图6-2 AIDS患病过程

针对传染期较短（几天到几个月）的传染病，一般可以忽略人口出生和死亡对于传播规律的影响，而对于传染期超长的传染病，就必须考虑出生率和死亡率对传染规律的影响了.

最后，AIDS致死率极高. 到目前为止，罕有AIDS治愈的报告，一旦感染HIV后，不论是否采取治疗措施，一般在十几年内都将死亡.

2. ARV治疗对传染率影响

有关ARV是否能够降低HIV性传播的风险，学术上没有一致的结论[5-7]. 大多数人都认同的结论是ARV治疗对于HIV的传播具有抑制作用. 这种抑制作用不仅来自于ARV治疗的生物化学作用，也包含了HIV确诊所带来的社会作用. 若某人开始接受ARV治疗，则此时他已经知道自己携带HIV病毒，就会主动控制自己的性活动，从而很大程度上降低HIV的传染率. 然而，更近一步的研究表明，ARV治疗会产生竞争效应. Law等在2001年提出，ARV治疗对于控制HIV传播的积极作用是降低HIV的活性，降低传染率，降低AIDS感染者年平均性行为数量；而消极因素由于ARV治理延长了HIV感染者的寿命，同时有可能增加HIV携带者在传染期内总的性行为的数量，这样就有可能抵消ARV治疗降低传染率作用，反而使HIV更容易扩散. 在HIV传播中，积极因素和消极因素会相互竞争，可能导致各种可能的结果. 此外，接受ARV治疗的HIV携带者会因为症状的减轻而增加性活动的数量，也有可能减少. 例如，象牙海岸的调查表明，当HIV确诊后，HIV感染者的性活动的数量显著减少，即使开展ARV治疗后也没有提高.

部分论文假设ARV治疗后可以有效地控制传播，甚至假设接受ARV治疗

的患者不再具备传染性；部分论文假设 ARV 治疗并不能降低传染风险，只能有效地延长 AIDS 患者的生命.

对于有争议的结论，我们无法判定何种假设更合理. 只要论文能够在自己所设定的假设的基础上，针对 ARV 介入后对 HIV 传播的影响问题，建立合理的数学模型得到相关结果，并对结果进行分析即可.

更好的做法主要有：

(1) 针对两种假设分别建立模型，并比较两种假设下模型的结果的差异性.

(2) 假设 ARV 治疗对于传染风险会造成一定程度的影响，比较不同影响程度下 ARV 治疗对 HIV 传播的影响.

实时上第二种做法兼容了第一种做法. 因此后面的模型均假设 ARV 无法彻底地阻断 HIV 的传播，只能在一定程度上影响传染率和潜在传播次数. ARV 治疗的基本原理是增加肌体的 CD4 含量以对抗 HIV，提高肌体的免疫力，延长 AIDS 患者的生命，而不是阻止 HIV 的传播.

实际中，如果 HIV 感染者开始接受 ARV 治疗，说明患者已经知道自己感染了 HIV 病毒，会控制自己的传播行为. 同时与其密切交往的人群知晓后也会减少与之发生传播关系的接触. 从这个角度上看，ARV 治疗间接地降低了 AIDS 患者的传染风险.

3. ARV 的抗药性问题

使用 ARV 疗法的一大风险就是 HIV 患者会对 ARV 产生抗药性. 如果 HIV 患者没有完整地按照 ARV 的要求进行治疗，就会对 ARV 产生抗药性. 当一位 HIV 感染者对 ARV 产生抗药性后，即使继续进行 ARV 治疗，也已经没有任何作用了. 一个接受 ARV 治疗的 HIV 感染者一旦产生了抗药性，其死亡率、传染率、母婴传播风险等传播特征可能发生改变.

有些模型假设一旦接受 ARV 治疗的 HIV 感染者产生了抗药性，其传播特性就和未接受 ARV 治疗的感染者的传播特性一致，不同的是产生抗药性的个体以后再也无法获得 ARV 治疗. 也有些模型假设个体产生抗药性后，其传播特性与未接受治疗的不同，具备自己特殊的传播特性，此时需要将人群分为易感人群、感染 HIV 并接受 ARV 治疗并且未产生抗药性人群、感染 HIV 并接受 ARV 治疗已产生抗药性人群和感染 HIV 未接受治疗人群. 各类人群之间可能存在转化关系，各类人群的传播特征也不一致.

6.1.4 常见的传染病数学模型

常见的传染病数学模型的主要思路是：在一定的假设条件下，按照传播特点将人群分为几类，根据传染病的传播特性，研究这几类人之间的转化规律.

例如：SI 模型将人群分为易感人群 (Susceptibility) 和传染人群

(Infectant)，用传染率系数 λ 表示单位时间内社会中的一个传染源导致一个易感人被感染的可能性，通过传染人群的人数 $I(t)$、易感人群的人数 $S(t)$ 和传染率系数 λ 即可计算得到下一个时间间隔由易感人转化为传染人的人数为

$$-\Delta S(t)=\Delta I(t)=S(t)I(t)\lambda,$$

根据两类人群的初始人数，通过迭代或者微分方程求解即可计算任意时刻两类人群总数的变化规律.

在此基础上，还可以增加考虑出生率、死亡率、治愈率以及人口迁移等因素，从而得到更为复杂的模型.

在上述模型中，若一个传染者治愈后，他/她又重新进入易感人群中，可能再次感染. 例如病毒变异性比较大的流感就属于这一类传染病，流感治愈后可能再次患流感. 但有些传染病一旦患病并治愈后，机体就具有了病毒抗体，例如天花. 针对这类疾病，可以将人群分为三类：易感人群 (Susceptible)、传染人群 (Infectious) 和免疫人群 (Recovered compartment)，治愈后的人群不再加到易感人群中，而是加到免疫人群中，不再参与疾病的传播. 该模型即为 SIR 模型.

有些疾病被传染后可能先进入潜伏期 (Exposed)，在潜伏期该传染源不具有传染性，经过一定时间后演化为传染期. 此时在每个时间间隔内，易感人群以一定的比例进入潜伏期人群；潜伏期人群大部分进入传染人群，极小部分自然死亡；传染人群中一定比例的传染者被治愈成为具有免疫力的免疫人群，极小部分因疾病死亡或自然死亡. 此时模型即为 SEIR 传染病模型.

此外针对 SARS 这类没有疫苗、极易传染的特殊传染病，有部分学者采用小世界模型研究人际交往关系，再通过仿真的方法研究超级传染事件 (SSE).

在研究人群之间的转化时，可以通过微分方程的方法计算各类人群的连续转化关系，也可以用差分方程、数值仿真、元胞自动机等离散方法计算每个时间步长中各类人群的变化规律.

除上述方法之外，还可以使用马尔可夫随机过程模型和闭环控制模型等来解释各个状态的转移关系. 并用马尔可夫过程和控制论的相关理论和方法求解模型的相关指标参数. 这些只是在表述上具备一些新意，其本质还是微分方程模型.

建立模型的基本步骤如下：

(1) 在一定的假设前提下，简化问题；

(2) 将人群按照传染病的传播特点分为几类；

(3) 根据传染病的特点确定各类人群之间的转化关系.

6.1.5　对问题的分析

1. 问题 (1) 的分析

问题 (1) 要求参赛队从各个大洲中各选择一个你们认为 HIV/AIDS 疫情

最严重的国家，估计这些国家在没有有效干预措施时，2006—2050 年 HIV/AIDS 人数的变化规律.

该问题包括两方面的工作，其一是从各大洲中选出一个疫情最严重的国家；其二是研究这些国家没有采用有效干预措施时 HIV 的自然传播规律.

在选择各大洲疫情最严重的国家时，首先需要制定 HIV/AIDS 疫情评价标准，再参照该标准评价各个国家. 在制定疫情评价标准时，需要综合考虑这个国家的人口总数、HIV 感染者数量和年龄结构以及经济发展水平等因素. 在综合各国疫情时可以建立综合评价模型，也可以定性地给出评价结果.

问题（1）的第二部分是研究这些国家在没有有效干预措施的情况下，2006—2050 年 HIV/AIDS 人数的变化规律. 所谓没有有效的干预措施，可以认为是在这期间 HIV/AIDS 的传播规律与历史传播规律相同，即认为在未来的 45 年里，传染率等传播特性参数没有发生改变. 在解决该问题时，需要根据历史数据得到 HIV/AIDS 的传播特性参数，并使用该参数对未来 45 年里的疫情进行预测.

当传染率等传播特性参数为常量时，可以在传统传染病模型的基础上，考虑 HIV 传播的特殊性建立传播模型.

2. 问题（2）的分析

问题（2）包括两部分内容，第一部分是要求预测这些国家在 2006—2050 年可支配的 HIV/AIDS 防控资金总量；第二部分是根据第一部分预测出的可支配资金总量，比较不同防控措施下疫情的发展状态. 有效防控措施包括：

（1）全部资金用于开展 ARV 治疗项目；

（2）全部资金用于开展 HIV 疫苗推广项目；

（3）合理分配 ARV 治疗项目和 HIV 疫苗推广项目的资金比例.

在可支配资金总量已知且当前疫情确定的情况下，通过建立合理的传播数学模型，可以预测三种防控措施下的疫情.

当 AIDS 患者接受 ARV 治疗时，一方面延长了患者的生命，另一方面降低了传染的风险. 因此每年接受 ARV 治疗的数量会影响 HIV 疫情的发展. 若全部的资金用于开展 ARV 治疗项目，可以预测每年可资助 ARV 治疗的病例数，此时类似自由传播模型，根据每年 ARV 治疗病例数，适当修正传染率和死亡率等参数即可得到全部资金用于开展 ARV 治疗项目时的疫情.

若在将来的某一年出现了有效的 HIV 疫苗，将可以极大地降低易感人群的传染风险. 接种 HIV 疫苗后的传染风险可以降到很低的水平. 根据各个国家每年可支配的 HIV 防控资金可以计算出每年的疫苗推广数量，根据该数量和接种疫苗后的传染率即可获得全部资金用于开展 HIV 疫苗推广项目时的 HIV 疫情.

上述两种策略各有优缺点，ARV 治疗项目可以有效控制传染源的危害，

也体现出国家对AIDS患者的人道主义关怀，但资金利用率较低，平均每人每年的ARV治疗费用高达1000美元，在资金有限的情况下，疫情控制效果不佳. HIV疫苗推广项目可以有效隔离易感人群，疫苗成本也不高. 但是，当易感人群数量非常庞大时，HIV疫苗推广也需要庞大的资金支持. 例如南非的HIV感染者数量庞大，如果只开展ARV治疗项目，则需要大量的资金才能有效控制疫情，而如果将资金的一部分用于开展HIV疫苗推广，则只需要少量的资金即可实现全国易感者的疫苗接种，可以极大程度地控制新增感染者数量，在将剩余的资金用于开展ARV治疗. 而对于人口基数庞大，HIV感染者比例不高的印度，如果要实现全国HIV疫苗的普及，则需要庞大的资金支持，而开展ARV治疗项目的资金将远远小于疫苗推广项目. 各个国家应该根据本国的特点，合理配置ARV治疗项目和HIV疫苗推广项目的经费，以实现最大程度地控制HIV疫情.

在采用第三种防控措施，即ARV治疗项目和HIV疫苗推广项目同时开展时，不同的资金投入比例，可能导致不同的疫情，此时需要根据本国的特点和可支配HIV防控资金的数量，优化二者的资金比例，以实现最有效的疫情控制.

3. 问题（3）的分析

ARV治疗的抗药性问题一直是ARV治疗项目推广的一大障碍. 没有严格按照规定的时间和剂量用药就有很大的风险产生抗药性. 一旦产生了抗药性，则宣告治疗的失败，此时即使资金充裕，也无法降低AIDS的传染率和死亡率.

实际上，当资金严重不足时，是否考虑抗药性问题对疫情发展的影响不大. 若一位接受ARV治疗的AIDS患者产生了抗药性，则停止对其用药，他的ARV治疗经费将转移到一位新增的ARV治疗者上. 当资金充裕，所有HIV感染者都能够接受ARV治疗时（例如美国），抗药性病例的增多将导致有效ARV治疗比例的降低，此时抗药性问题就会严重影响疫情的发展.

实际上，问题（3）只需要在问题（2）的基础上，增加考虑ARV治疗的抗药性问题，重新评价三种防控措施的效果，即可得到考虑抗药性问题时的最佳资金分配策略.

4. 问题（4）的分析

问题（4）中需要给联合国提交一份有关HIV/AIDS防控资金配置的建议书. 建议书中应该体现前面各个模型的研究成果. 包括在抗HIV/AIDS中ARV药物的供应和预防HIV疫苗的资源的最佳配置策略，并给出该策略下的疫情发展趋势.

说明国际援助项目不但可以帮助欠发达地区抗艾滋病，也有利于减少本国的输入性病例数量，实现全世界协作抗HIV的良好氛围，最终实现全面消

灭 HIV.

通过对模型中各个参数进行灵敏度分析，说明在防控 HIV/AIDS 中哪些变量是防控的关键变量，为 UNAIDS 下一阶段的防控措施的制定提供建议.

5. 各问题之间的关系

通过前面对问题的分析，可以得到各个问题之间的关系如图 6-3 所示.

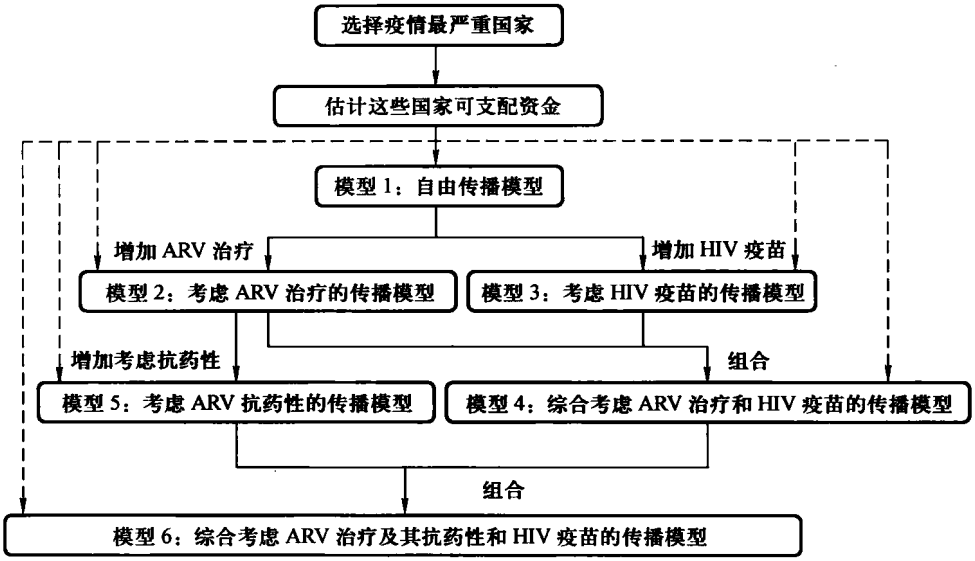


图 6-3 各个问题之间的关系

问题首先需要从诸多国家中按照一定规则选择出各大洲中疫情最严重的国家，此问题的目的是降低后面模型验证数据收集工作的复杂度.

之后需要估计这些国家在未来 45 年里每年可支配的 HIV/AIDS 防控资金. 该资金数量将会影响各种防控措施的力度，包括 ARV 治疗的普及程度、HIV 疫苗研发成功的时间和 HIV 疫苗的推广速度等，进而影响各国疫情防控效果.

为了使模型建立过程由浅入深，便于理解，问题中实际要求建立 6 个模型. 模型 1 是自由传播模型，此时不考虑 ARV 治疗、ARV 抗药性问题和 HIV 疫苗的作用. 该模型所需的参数（传染率、出生率和死亡率等）可以通过历史数据拟合得到. 该模型一方面为后面模型奠定基础，另一方面可以为 HIV 防控效果评价提供对照基础.

在模型 1 的基础上，增加考虑 ARV 治疗的作用就得到了模型 2. 模型 2 中的 ARV 治疗推广的力度受该国可支配 HIV/AIDS 防控资金额度的影响，进而影响疫情水平. 通过建立相关模型，最终可以得到若将所有防控资金全部用于开展 ARV 治疗项目时的疫情发展状况.

在模型 1 的基础上，增加考虑 HIV 疫苗的作用即可得到模型 3. 同理，疫苗推广项目的普及程度取决于该国可支配 HIV/AIDS 防控资金的影响. 在模型

1 的基础上建立考虑 HIV 疫苗的模型，最终可以得到如果全部防控资金全部用于开展 HIV 疫苗推广项目时的疫情发展状况.

组合模型 2 和模型 3，即可得到综合考虑 ARV 治疗和 HIV 疫苗时疫情的发展状况模型，即模型 4. 当同时存在 ARV 治疗和 HIV 疫苗时，就存在 ARV 治疗经费和 HIV 疫苗推广经费的分配问题. 不同的分配策略对于疫情的控制程度是不同的，此时需要最优化二者之间的比例，以实现最佳的疫情控制效果.

在模型 2 的基础上，增加考虑 ARV 治疗可能产生的抗药性问题，即可得到模型 5. 对某些国家来说，抗药性问题可能导致 ARV 治疗经费利用率降低，疫情恶化等不良后果.

组合模型 4 和模型 5 即可得到较接近实际情况的模型 6，此时综合考虑了 ARV 治疗、ARV 抗药性问题和 HIV 疫苗. 通过配置 ARV 治疗经费和 HIV 疫苗推广经费的比例，最终实现考虑抗药性问题的最佳资金配置方案.

在上述 6 个模型中，后面的模型一般都是在前面的模型基础上，增加所考虑因素，或者综合前面几个模型得到的. 这样做可以让模型建立过程层次更为清晰，便于说明. 下面将按照图 6-3 的顺序介绍各个问题的解决方案.

6.2 解决两个基本问题

6.2.1 基本问题（1）的解决方案

1. 方案 1：基于数据指标的国家选择方案

位于北美的海地的 HIV 感染率很高，达到了 2.6%，其邻国多米尼亚共和国的感染率也达到了 1.4%；其他发达国家，如加拿大的感染率为 0.15%，美国的感染率为 0.29%. 同时，海地还是世界上最贫困的几个国家之一，人均 GNP 仅有 510 美元. 若按照人均 ARV 治疗费用 1 000 美元计算，海地的 HIV 感染者无法承受如此高昂的治疗费用.

圭亚那是南美有代表性的国家，圭亚那的 HIV 感染率为 2%. 虽然巴西的总病例数是圭亚那的 36 倍，但是圭亚那的总人口要远小于巴西的人口数. 此外，巴西已花费 4.44 亿美元作为巴西境内所有 HIV 感染者的治疗费用. 圭亚那的 GNP 远小于巴西，没有足够的资金用于抗 AIDS.

非洲是遭受 AIDS 影响最严重的地区，其中南非又是非洲国家中最具代表性的国家. 南非全国总计有 420 万的 AIDS 患者，占总人口的 9.33%，不论是病例数还是感染比例都处于世界最高位置. 此外 29.5% 的孕妇被检测出 HIV 阳性，这意味着将进一步传染给她们的孩子[3].

1999 年，乌克兰共有 24 万的 AIDS 病例，而且这个国家的人口一直在

减少.

从印度近期的种种迹象表明，印度可能会成为下一个 HIV/AIDS 大国. 虽然印度的 HIV 感染率仅为 0.47%，远远低于海地、南非等 HIV 高发国家，但印度是世界第二人口大国，总数达到了 11 亿. 实际上，印度也是世界第二大的 AIDS 病例国，HIV 感染人数达到了 370 万. 预计印度将在 2010 年成为 HIV 阳性人数超过南非，成为 HIV 世界第一大国.

大洋洲没有更多的选择，只有选择澳大利亚. 大洋洲各个国家的 HIV 感染率都很低，澳大利亚是大洋洲 HIV 病例数最多的国家，也仅有 14000 例.

2. 方案 2：基于特征指标的国家选择方案

为了确定哪些国家的 HIV/AIDS 形势最为严峻，从众多指标中选用 5 项主要指标进行分析：患病率变化趋势；感染人群的人口特征；HIV/AIDS 教育与宣传水平；HIV 传播途径；当前和历史 HIV/AIDS 统计数据的可用性.

基于上述 5 项指标综合考虑，从五个大洲中分别选出如下具有代表性的国家：

非洲：南非. 南非的特点是有非常大的 HIV/AIDS 感染人群，但是很少使用毒品.

亚洲：印度. 印度的特点是有很庞大的人口基数，HIV/AIDS 感染比例并不高，但有快速增长的趋势.

欧洲：俄罗斯. 俄罗斯有庞大的 HIV 阳性人群以及注射毒品的群体.

美洲：美国. 得益于安全性行为的普及，美国的 HIV 感染率很低.

大洋洲：澳大利亚. 大洋洲的国家较少，仅有澳大利亚有少量的 HIV 感染者，其他国家几乎没有 HIV 携带者.

3. 方案 3：基于危急程度模型的国家选择方案

（1）模型的分析与假设

一个 HIV 防控处于危急状态的国家应该具备如下特征：

① 该国有很庞大的 HIV 携带人群，或者很大比例的 HIV 携带者；

② 一个防控方案在该国的实施能够带来防控效果的显著提升. 一些 ARV 治疗普及率比较高的国家已经无法为其 HIV/ADIS 患者做更多的事情，所以这些国家并不认为是处于危急状态；

③ 该国的 HIV/AIDS 在近期的危害很严重，迫切需要采取相关措施.

综上所述，模型用未来 5 年（2006—2010 年）的累计 AIDS 致死人数来定义这个国家 HIV/AIDS 的危急程度，以此量化在没有其他治疗方案介入的前提下，HIV/AIDS 的瞬即效应. 危急程度包括绝对危急程度和相对危急程度：

绝对危急程度：在 2006—2010 年期间，因为 HIV/AIDS 而减少的全国寿命之和；

相对危急程度：由于 HIV，导致人均减少的寿命.

为了建立危急程度的模型，给出下面的假设：

① 每个接受 ARV 治疗的病人都能坚持治疗，不考虑抗药性的问题. 一个病人要么接受 ARV 治疗，要么不接受治疗，不存在中间情况；

② 在未来的 5 年内没有有效的治疗方案的介入；

③ ARV 治疗的普及率是个常数；

④ 不考虑其他导致人群死亡的因素. 由于模型只进行短期的预测，一些可能导致人口显著变化的原因，例如天灾、战争和其他流行病的大规模爆发的可能性很小；

⑤ 不考虑未来 5 年该国人口的变化；

⑥ 定义绝对危急程度单位：人·年，表示一个人活了 1 年，称为 1 人·年.

（2）危急程度的基本模型

定义绝对危急程度 ζ_{abs} 和相对危机程度 ζ_{rel} 为

$$\zeta_{abs}=\alpha\gamma(1-\delta)+\beta\gamma\delta,$$
$$\zeta_{rel}=\frac{\zeta_{abs}}{P} \tag{6.1}$$

其中 α 为没有接受 ARV 治疗的 HIV 携带者寿命缩短的平均年数；β 为接受 ARV 治疗的 HIV 携带者寿命缩短的平均年数；P 为该国的人口总数；γ 为 2006—2010 年的累积新增感染人数；δ 为新增感染人员中，接受 ARV 治疗的比例.

在世界范围内，HIV 携带者被传染 HIV 病毒的平均年龄为 23 岁. 在发达国家，一位 HIV 感染者，若没有接受 ARV 治疗，从感染到死亡的时间平均为 $L=12$ 年，则这些 HIV 感染者的平均寿命为 $23+12=35$ 岁；假设 ARV 治疗能够让这些 HIV 感染者多活 8 年，那么这些 HIV 感染者的寿命就可以达到 43 岁. 如果该国家的平均寿命为 Γ 则 $\alpha=\Gamma-35$，$\beta=\Gamma-43$.

（3）平均寿命 Γ 的确定模型

随着医疗和生命科学技术的发展，人们的平均寿命 Γ 在不断延长，即不同年代出生的人的平均寿命是不一样的. 下面建立模型估计某国 2006—2010 年的平均寿命. 为此假设：

① 在未来的 5 年里，该国家人口的年龄结构没有发生显著的变化；

② 对全国人口年龄从 0～100 岁，按照每隔 5 年划分一个组，不考虑组内年龄的差异. 由于超过 100 岁的人口很少，可以将他们归到［95，100）的区间中；

③ 接受 ARV 治疗的 HIV 携带者都能够接受完整的治疗疗程；

④ 不考虑移民的影响.

到 2005 年为止，经过修正的全国平均寿命 Γ_0 为

$$\Gamma_0 = \frac{1}{P}\sum_{i=0}^{19} L_i P_i,$$

其中 L_i 是年龄处于 $[5i, 5i+5)$ 人口的平均寿命，P_i 是年龄处于 $[5i, 5i+5)$ 的人口数.

前面计算过的 Γ_0 是在当前人口统计数据下得到的，该数据包含了 HIV 对国民寿命的影响. 若这个国家没有 HIV 传播，则国民的平均寿命将有所提高，提高的程度越明显，则表明这个国家的 HIV 传播越危急，控制 HIV 传播的效果越明显.

在没有 HIV 传播时，若其国民的平均寿命无法统计得到，则可以用这个国家的非 HIV 感染人员的平均寿命 Γ_A 来估算. Γ_A，Γ_0 和 Γ_{HIV} 三者之间的关系为

$$\Gamma_0 P = \Gamma_A (P - P_{HIV}) + \Gamma_{HIV} P_{HIV},$$

进一步地有

$$\Gamma_A = \frac{\Gamma_0 P - \Gamma_{HIV} P_{HIV}}{P - P_{HIV}},$$

此时则有

$$\alpha = \Gamma_A - 35,\ \beta = \Gamma_A - 43. \tag{6.2}$$

(4) 新增感染人数 γ 的确定模型

下面采用 Logistic 人口增长模型，对未来 5 年的新增 HIV 感染病例的数量 γ 进行预测分析. 为此给出如下假设：

① 不考虑未来 5 年内全国人口数量的变动，即出生率和死亡率基本相等；

② 就如下 12 个区域研究其 HIV/AIDS 危急程度：非洲、东南亚和中亚、北亚和东亚、大洋洲、巴西、南美（巴西除外）、美国、墨西哥、中美洲、加勒比海国家和欧洲；

③ 在所研究的 12 个区域中，HIV/AIDS 的影响程度相对稳定；

④ HIV 感染者的数量占区域总人口的数量保持不变，记为 $H_{relative}$.

由上面的假设，各个区域的 HIV 感染者数量占区域总人口的比例 $H_{relative}$ 保持不变，因此，通过预测区域人口的变化即可预测出 HIV 感染数量的变化规律.

在实际中，Logistic 阻滞增长模型常常用来预测具有上限且可持续发展的问题，广泛地应用于人口的预测问题. Logistic 阻滞增长模型的基本形式为

$$\frac{\mathrm{d}P}{\mathrm{d}t} = r\left(1 - \frac{P}{K}\right)P,$$

其中 P 为该区域在 t 时刻的总人口数；r 为该区域人口的最大极限增长速度；K 为饱和人口数，即该区域资源所容许的最大人口数. 这是一个微分方程模型，其通解为

$$P(t) = \frac{r}{ce^{-rt} + r/K},$$

其中 c 是由问题的初值确定的一个常数.

下面用1985—2000年的实际数据来估计上式中的参数 r 和 K. 将 Logistic 模型改为

$$\frac{1}{P}\frac{\mathrm{d}P}{\mathrm{d}t}=a+bP,$$

其中 $a=r$，$b=-r/k$ 即 $P(t)$ 和 $\frac{1}{P(t)}\frac{\mathrm{d}P(t)}{\mathrm{d}t}$ 之间成线性关系. 用1985—2000年的实际数据代入 $P(t)$ 和 $\frac{\mathrm{d}P}{\mathrm{d}t}=P(t+1)-P(t)$ 中，利用最小二乘线性回归求解得到 a 和 b 的值，进一步计算得到参数 r 和 K，从而可以得到 $P(2010)$ 和 $P(2005)$ 的值. 此时即可得到2005—2010年的累计 HIV 感染人数为

$$\gamma=[P(2010)-P(2005)]H_{relative}. \tag{6.3}$$

将（6.2）式和（6.3）式代入（6.1）式，得到各个区域的绝对危急程度和相对危急程度.

（5）模型的结果分析

用108个国家的实际数据作统计计算，则可得到这些国家的 HIV 绝对危急程度和相对危急程度，相对危急程度最严重的国家为：博茨瓦纳、泰国、汤加、乌克兰、巴哈马群岛和圭亚那. 15个相对危急程度最严重的国家中，有14个在非洲. 绝对危急程度受该国人口基数的影响，那些相对危急程度中等的国家，如果人口基数大的话，则其绝对危急程度也会很大.

4. 模型的评价

针对问题1给出了3种方案，前两种方案都是通过定性分析方法直接给出了各大洲 HIV/AIDS 最严重的国家. 方案3则建立了较复杂的数学模型，该模型中定义了危急程度的概念，并用 Logisitic 阻滞模型通过预测人口总量的方法预测未来5年里一个国家 HIV 的传播规模. 从科学严谨的角度来看，通过定量的方法评价各个国家的危急程度的做法是值得肯定的，但从实际问题的角度也未必是完美的.

方案3通过简单的 Logisitic 阻滞增长模型估计未来5年 HIV 的传播规模，确定了传播最为危急的几个国家，而问题1的后半部分又采用其他方法重新确定这几个国家未来5年的 HIV 传播情况. 这样就可能出现用两种模型得到的传播结果不一致. 如果用两种模型对全世界各个国家重新评价 HIV 的危急程度，可能导致两种模型的结果是不同的，这似乎是解决问题的逻辑存在一定的问题.

通过建立数学模型的方法选择危急程度最严重的国家的方法，虽然存在一些逻辑上的问题，但只要适当改进模型即可消除. 例如：首先建立精确的疫情预测模型，预测未来50年各国的疫情，再对预测结果进行评价，从而选出各

大洲 HIV/AIDS 流行程度最严重的国家，并对这些国家的传播规律进行分析. 这样做就同时解决了问题 1 中的两个子问题. 但这样做的工作量会很大，需要对所有国家进行 50 年的疫情预测. 另一种可行的改进方法是只使用历史数据直接推断各个国家的疫情，不进行预测，找出当前疫情最严重的国家，再对这些国家的疫情发展进行预测.

从问题的开放性的角度来看，应该鼓励这种严谨的做法. 但是从审题的角度来看，这样做并不能给竞赛论文带来额外的加分，反而会消耗大量的时间和精力，并且导致竞赛论文的重点不突出.

6.2.2 基本问题（2）的解决方案

在这里依据一个国家的 GDP 增长和其他经济数据，来预测该国的 AIDS 防控资金情况.

1. 基于 GDP 增长预测模型

假设在未来 50 年里，世界经济格局不会发生显著改变；全球经济保持平稳增长；HIV 治疗费用的通胀可以通过世界经济的利率进行修正.

据统计数据，经济欠发达国家政府每年用于 HIV/AIDS 防控的资金约占 GDP 的 0.000 5% 左右. 在世界经济格局没有发生显著改变的前提下，这一比例保持不变.

设某国的 GDP 增长率为 ρ，则第 t 年政府拨付给 HIV 防控资金的额度为

$$G(t)=\mathrm{GDP}_{2005}\times(1+\rho)^{t-2005}$$

其中 GDP_{2005} 表示该国 2005 年的 GDP 总值.

对于大多数的经济欠发达国家而言，政府拨付的 HIV/AIDS 防控资金并不能满足有效防控的需要，缺口主要由国际援助来填补. 例如 2003 年，博茨瓦纳、汤加、巴哈马群岛以及圭亚那这五个国家的国际援助 AIDS 资金达到本国 AIDS 资金的 85% 以上. 而泰国和乌克兰相对发达，这两个国家的国际援助资金只占全部 HIV 防控资金的 25%[4]. 根据两类国家 2005 年的 GDP，可以计算上述两类国家获得的 HIV 国际援助资金分别为

$$A_s(2005)=0.0005\%\times\mathrm{GDP}_{2005}\times\frac{85}{15},$$

$$A_d(2005)=0.0005\%\times\mathrm{GDP}_{2005}\times\frac{25}{75},$$

其中 $A_s(2005)$ 是欠发达国家在 2005 年获得的国际援助额度，$A_d(2005)$ 是发展中国家在 2005 年获得的国际援助额度.

随着世界经济的发展，以及人们对 HIV 危害性认识的提高，HIV 的国际援助资金也在逐年增高. 假设世界范围内的 HIV 国际援助资金的增幅为 5%，则第 t 年这两类国家获得的国际援助额度为

$$A_s(t)=A_s(2005)\times 1.05^{t-2005},$$

$$A_d(t)=A_d(2005)\times 1.05^{t-2005},$$

这两类国家第 t 年可用于 HIV 防控的资金总额为

$$F_s(t)=G(t)+A_s(t),$$

$$F_d(t)=G(t)+A_d(t),$$

其中 $F_s(t)$ 和 $F_d(t)$ 分别表示欠发达地国家和发展中国家在第 t 年可用的 HIV 防控基金额度.

假设 HIV 感染者接受 ARV 治疗的成本为 1 000 美元/人年，假设治疗费用的通胀率为 2%，则第 t 年每位 AIDS 患者的 ARV 治疗费用为 $C_{ART}(t)=1.02^{t-2005}$.

第 t 年能够资助 HIV 感染者接受 ARV 治疗的数量为

$$N_s(t)=\frac{F_s(t)}{C_{ART}(t)}$$

$$N_d(t)=\frac{F_d(t)}{C_{ART}(t)}$$

其中 $N_s(t)$ 和 $N_d(t)$ 分别表示欠发达国家和发展中国家在第 t 年可资助 ARV 治疗的病例数量.

2. 基于统计数据的拟合模型

UNAIDS 在 2005 年上半年给出了世界 HIV 防控所需的资金总额和目前可用的资源总量，并制定了一项包括如下几个方面的防治计划：

（1）建立一个国际合作组织，致力于开展有关 HIV 预防和治疗的各个方面的研究；

（2）若国际援助能够持续，在 2010 年为 HIV 最危急的几个国家提供 75% 的经费援助；

（3）在低收入国家开展医务人员培训；

（4）在资金允许的前提下，2010 年前新建 2 700 个 HIV 康复中心.

根据 UNAIDS 统计，2004 年全世界总计有 61 亿美元的国际援助资金，预计 2005—2007 年将分别达到 73 亿、89 亿和 100 亿. 通过线性拟合，得到第 t 年全世界的 AIDS 国际援助资金数量模型为

$$W(t)=13.3t-2\ 659.$$

设某危急国家在 2005 年获得的国际援助基金为 $A(2005)$，假设 UNAIDS 对各个国家援助的比例关系保持不变，则第 t 年该国可获得的国际援助资金 $A(t)$ 为

$$A(t)=W(t)\frac{A(2005)}{W(2005)}$$

根据 UNAIDS 报告，对 AIDS 危急国家的国际援助比例将达到其全部费用的 75%，则某危急国家第 t 年的 HIV/AIDS 防控总费用为

$$F(t)=\frac{A(t)}{0.75}$$

一般来说，从AIDS病发到死亡的时间为3年，设第t年开始病发的AIDS感染者中，接受治疗的人数为$T(t)$，第t年可用的总经费为$F(t)$，则第t年可批复的申请数量为

$$T(t)=\max\left[\frac{F(t)}{(1+r)^{(t-2006)}}-T(t-1)-T(t-2),0\right],$$

其中r是ARV治疗药物的通胀率.

6.3 模型1：基于SI模型的HIV传播模型

模型1基于最简单的SI模型研究HIV的传播问题. 该模型不考虑出生和死亡对传播的影响，也不考虑传播途径的多样性问题. 分节分别从6个不同的角度来建模分析.

6.3.1 模型1.1：HIV自由传播模型

1. 模型的分析与假设

将人群分为两大类，一类是易感人群S，另一类是传染人群I，两类人群之间存在如图6-4所示的状态转移关系.

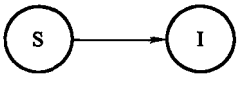


图6-4 SI模型状态转移关系图

在单位时间内有一定数量的易感者转化为传染者，其中传染率λ，用以表示转化速度. 根据上述转化关系可以建立HIV的SI传播模型.

假设在研究时间区间内，一个国家的人口总数N没有发生显著变化；在研究时间区间内，传染率λ保持恒定.

2. 自由传播模型的建立

设每位HIV携带者每年有效接触的人数为r_c，则一位HIV感染者在一年内的接触人数$N_{HIV}(t)$为

$$N_{HIV}(t)=\frac{S(t)}{N}r_c,$$

其中$S(t)$表示第t年易感人群的总数；N是第t年总人口数，是传染人群总数$I(t)$和易感人群总数$S(t)$之和.

设一位HIV感染者与一位HIV易感者在一年的接触中，将HIV传染给易感者的概率为λ，则每位HIV携带者每年感染的人数$Tran(t)$为

$$Tran(t)=N_{HIV}(t)\lambda,$$

因此，每年的新增感染人数为

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t}=\lambda S(t)I(t)\frac{r_c}{N},$$

其中 $I(t)$ 为第 t 年传染人群的总数.

在上式中，假设传染率 λ 是个常数. 但是，由于 HIV 的传播往往是与特定的行为密切相关，例如性行为、毒品静脉注射等，所以不同的 HIV 携带者其传染率有所不同. 也就是说，一些 HIV 携带者的传染率会很高，例如一些非正常的性行为的传染风险就要高于正常性行为的传播风险；此外，在一次危险性行为中，男性传染给女性的风险高于女性传染给男性的风险. 但是为了简化模型，这里仅仅使用平均传染率进行计算.

综合考虑易感人群的数量，建立如下微分方程组模型：

$$\begin{cases}\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=\lambda S(t)I(t)r_c,\\[2ex]\dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=-\lambda S(t)I(t)r_c,\\[2ex]I(t)+S(t)=N,\\I(0)=I_0,\\S(0)=S_0.\end{cases}$$

3. 模型的求解与结果分析

可以利用历史数据对模型中的 λr_c 值进行最小二乘拟合估计，对近期 HIV 传播情况进行预测. 具体做法如下：

假设 t_0，t_1，…，t_n 的数据为历史数据，需要预测 t_{n+1}，t_{n+2}，…. 设历史传染人群总数为 $I(t_0)$，$I(t_1)$，…，$I(t_n)$，历史易感人群总数为 $S(t_0)$，$S(t_1)$，…，$S(t_n)$. 则可得到下面的方程：

$$I(t_{i+1})-I(t_i)=\lambda r_c S(t_i)I(t_i)+\varepsilon_i,\ i=1,\ 2,\ \cdots,\ n,$$

其中 ε_i 为第 i 个方程的误差项，最小化所有误差即可得到参数 λr_c 的最佳估计：

$$\min\varepsilon=\sum_{i=1}^{n}\varepsilon_i^2$$

$$\text{s. t. }\varepsilon_i=\lambda r_c S(t_i)I(t_i)-[I(t_{i+1})-I(t_i)],\ i=1,\ 2,\ \cdots,\ n.$$

对上述模型代入典型参数后，求解得到传染人群总数和易感人群总数随时间变化的曲线，如图 6-5 所示.

从图 6-5 中可以看出，传染人群呈 S 型增长，初期传播速度较慢，这是因为此时的传染源数量较少；中期出现传播高峰，此时具有大量的传染源和易感者；到了后期传播速度再次变缓，这是由于此时的易感者人数逐步减少.

考虑未来较长时间的传播结果，来讨论上面微分方程组的平衡点，即令

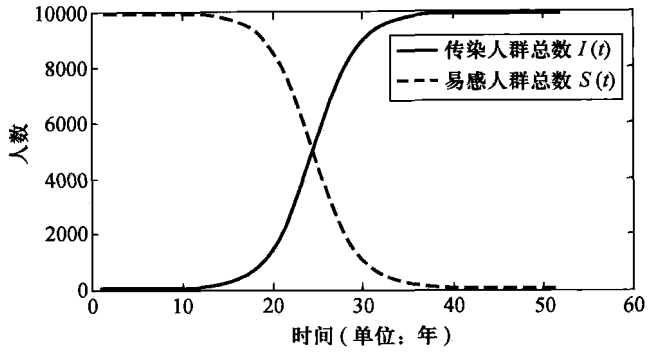


图 6-5 传染人群和易感人群总数随时间变化曲线

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t}=\lambda S(t)I(t)r_c=0,$$

则有

$$\Rightarrow\begin{cases}\lim\limits_{t\to\infty}S(t)=0,\ \lim\limits_{t\to\infty}I(t)=N, & I(0)>0\\ \lim\limits_{t\to\infty}I(t)=0,\ \lim\limits_{t\to\infty}S(t)=N, & I(0)=0\end{cases}$$

即该模型有两个平衡点，若初始时刻存在传染源，则最终会稳定在全部感染的状态；否则不会出现感染事件.

6.3.2 模型 1.2：考虑 ARV 治疗的传播模型

1. 问题的分析与假设

在未考虑 ARV 治疗时，所有感染者的传播特性都是一样的. 当考虑 ARV 治疗后，接受治疗的传染者的传染性有一定程度的降低. 按照传播特性来分，可以将人群分为易感人群、感染 HIV 未接受 ARV 治疗人群和感染 HIV 接受 ARV 治疗人群. 各类人群对于新增感染者数量的作用不一样，需要分别建立各类人群之间的转化关系，最终求得感染 HIV 人群总数随时间变化的规律.

假设接受 ARV 治疗后，设传染率为治疗前的 f 倍；不考虑出生和死亡对 HIV 传播的影响.

2. 传播模型的建立

根据当年可获得的 ARV 治疗资金，可以得到接受 ARV 治疗的人群占 HIV 感染者的比例为 $p_{arv}(t)$. 当 ARV 治疗介入后，将传染人群分为 2 类，一类是接受 ARV 治疗的人群，其数量为 $T(t)=p_{arv}(t)I(t)$，其传染率为 λf；另一类是未接受 ARV 治疗的人群，其数量为 $U(t)=[1-p_{arv}(t)]I(t)$，其传染率为 λ. 此时状态转移图如图 6-6 所示.

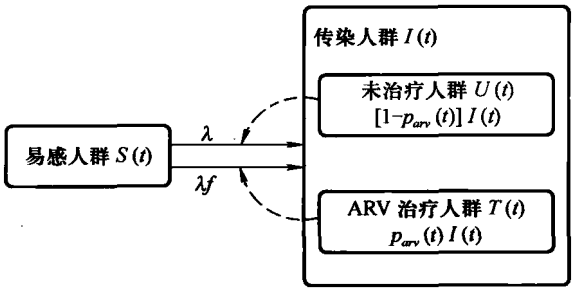


图 6-6　考虑 ARV 治疗的简单 SI 模型的转换关系图

在图 6-6 中，易感人群有两种被感染的方式，一种是与未接受治疗的传染源接触，另一种是与接受治疗的传染源接触，两类接触被感染的风险不一致. 易感者一旦被感染后就有一定的概率接受治疗，成为接受 ARV 治疗的传染源.

SI 模型考虑 ARV 治疗效果后修改为

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t}=S(t)r_c\lambda[U(t)+fT(t)],$$

用上式替换简单 SI 模型，即得到考虑 ARV 治疗的 SI 模型：

$$\begin{cases}\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=S(t)r_c\lambda[U(t)+fT(t)],\\ \dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=-S(t)r_c\lambda[U(t)+fT(t)],\\ T(t)=p_{arv}(t)I(t),\\ I(t)=U(t)+T(t),\\ I(t)+S(t)=N,\\ U(0)=U_0,\ T(0)=t_0,\ S(0)=S_0.\end{cases}$$

求解上述模型即可得到 HIV 病例数随时间的变化关系.

6.3.3　模型 1.3：存在 HIV 疫苗的传播模型

1. 问题的分析与假设

从可预见的 HIV 疫苗的研究进展来看，近几年难以研发出可用于临床实践的 HIV 疫苗. 预计 HIV 疫苗将在未来 10 年左右面世，该成果将成为 21 世纪最伟大的人类发明之一.

若存在 HIV 疫苗，则接种 HIV 疫苗者感染 HIV 的风险就可以降低到一个很小的程度，此时模型 1.2 中的易感人群可以划分为两类，一类是未接种疫苗的易感人群，另一类是已接种疫苗的免疫人群.

对于不发达国家而言，可用于开展 HIV 疫苗普及计划的经费十分有限. 这些国家无法实现短期内全面普及接种 HIV 疫苗，只能根据可支配的 AIDS 防

控基金逐步开展HIV疫苗普及计划. 因此，每年只有一部分的易感人群能够接种疫苗. 根据传播特点，建立各类人群之间的转化关系，即可得到存在疫苗时的HIV传播规律.

假设不考虑出生和死亡；疫苗将在未来10年后面世；疫苗并非百分之百有效，只是降低了传染风险. 设一个已接种疫苗的易感者与HIV接触后的感染率为 $\lambda_e=e\lambda$，这里取 $e=0.05$，即HIV疫苗的有效率为95%；不考虑疫苗的有效期，即一旦接种疫苗，则对HIV具有一定的抗体，有关疫苗有效期的问题将在后面模型中给出；不考虑ARV治疗，有关ARV治疗与疫苗的共同作用将在模型1.6中讨论.

2. 传播模型的建立

在疫苗出现之前，HIV一直按照模型1.1的规律进行传播. 当10年后出现并开始推广HIV疫苗时，各类人群的初值就是模型1.1在10年后的状态值.

根据各类人群之间的转换关系得到状态转移关系图，如图6-7所示.

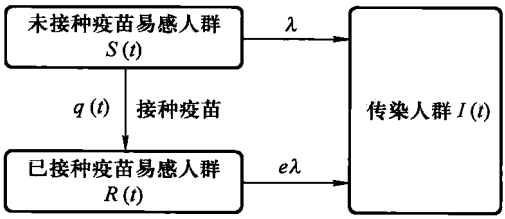


图6-7 考虑疫苗时的传染病模型转移关系图

在图6-7中，将人群被分为3类，一类是未接种疫苗的易感人群，其与传染人群接触被感染HIV的传染率为 λ. 根据可支配的HIV疫苗推广经费，可以得到第 t 年有可接种疫苗的数量为 $q(t)$，该值与当年可支配的HIV疫苗推广经费有关. 已接种疫苗的易感者与传染人群接触的传染率为 $e\lambda$. 根据各类人群的转化关系可以得到考虑疫苗时的HIV传播运动学方程：

$$\begin{cases}\dfrac{dI(t)}{dt}=\lambda S(t)I(t)r_c+e\lambda R(t)I(t)r_c,\\ \dfrac{dR(t)}{dt}=q(t)-e\lambda R(t)I(t)r_c,\\ \dfrac{dS(t)}{dt}=-q(t)-\lambda S(t)I(t)r_c,\\ I(t)+S(t)+R(t)=N,\\ I(10)=I_0,\\ S(10)=S_0,\\ R(10)=0.\end{cases}$$

其中 I_0 是模型1.1中在 $t=10$ 时传染人群的总数；S_0 是模型1.1中在 $t=10$ 时

易感人群的总数.

6.3.4　模型 1.4：考虑 ARV 治疗和 HIV 疫苗的传播模型

1. 问题的分析与假设

在模型 1.2 中，将传染人群分为接受 ARV 治疗的人群和未接受 ARV 治疗的人群. 在模型 1.3 中，将易感人群分为未接种疫苗的人群和接种疫苗的人群. 这里组合模型 1.2 和 1.3 即可得到同时考虑 ARV 治疗和 HIV 疫苗的传播模型.

假设疫苗不是百分之百有效，接种疫苗后能够将传染率降低到未接种疫苗时的 e 倍. 即接种疫苗者与未接受治疗的传染者接触的传染率为 $e\lambda$，与接受 ARV 治疗的传染者接触的传染率为 $e\lambda f$；不考虑出生和死亡；疫苗将在未来 10 年后面世；不考虑疫苗的有效期，即一旦接种疫苗，则对 HIV 具有一定的抗体；不考虑抗药性对疫情的影响.

2. 传播模型的建立

组合模型 1.2 和 1.3，得到同时考虑 ARV 治疗和 HIV 疫苗的状态转移关系如图 6-8 所示.

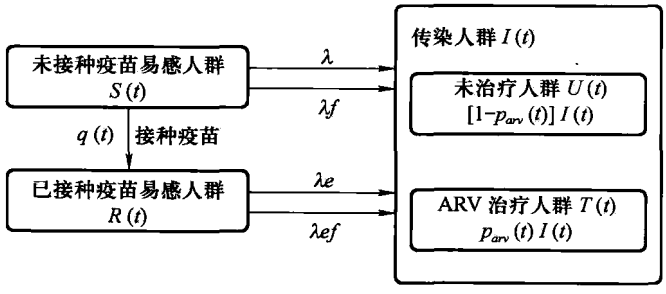


图 6-8　同时考虑 ARV 治疗和疫苗时的 HIV 传播模型的转移关系图

此时，将易感人群分为两类，传染人群也分为两类，不同类型的易感者与不同类型的传染者接触后，其传染率也有所不同. 未接种疫苗的易感者与未治疗的传染者接触的传染率为 λ，与接受 ARV 治疗的传染者接触的传染率为 λf；接种疫苗的易感者与为治疗的传染者接触的传染率为 λe，与接受 ARV 治疗的传染者接触的传染率为 λef. 依据上述传染率的描述，可以建立如下传播模型：

$$\begin{cases}\dfrac{dI(t)}{dt}=r_c\lambda[S(t)+R(t)e][U(t)+fT(t)],\\ \dfrac{dR(t)}{dt}=q(t)-r_c\lambda eR(t)I(t)[U(t)+fT(t)],\\ \dfrac{dS(t)}{dt}=-q(t)-r_c\lambda S(t)I(t)[U(t)+fT(t)],\\ T(t)=I(t)p_{arv},\\ T(t)+U(t)=I(t),\\ I(t)+S(t)+R(t)=N,\\ I(10)=I_0,\\ S(10)=S_0,\\ R(10)=0.\end{cases}$$

其中 I_0 是模型 1.2 中当 $t=10$ 时传染人群的总数；S_0 是模型 1.2 中当 $t=10$ 时易感人群的总数.

6.3.5 模型 1.5：考虑 ARV 治疗抗药性的传播模型

1. 问题的分析与假设

当接受 ARV 治疗的感染者对 ARV 治疗产生抗药性后，ARV 治疗不再发挥作用. 此时该人群的传染率将发生改变，需要将其设定为一个特殊的人群，并研究该人群与易感者接触后导致新增传染者数量的关系. 假设对 ARV 治疗产生抗药性感染者的传染率是未接受治疗的 g 倍，即产生抗药性人群的传染率为 λg；单位时间内有 q_W 的 ARV 治疗者产生抗药性；一旦 ARV 治疗者产生抗药性，即对其停药以提高 AIDS 基金效率；同样不考虑出生率和死亡率对传播的影响；不考虑疫苗.

2. 传播模型的建立

由于从 ARV 治疗人群向抗药性人群存在转移关系，因此需要将感染人群分为非抗药人群和抗药性人群两类. 各类之间的转移关系如图 6-9 所示.

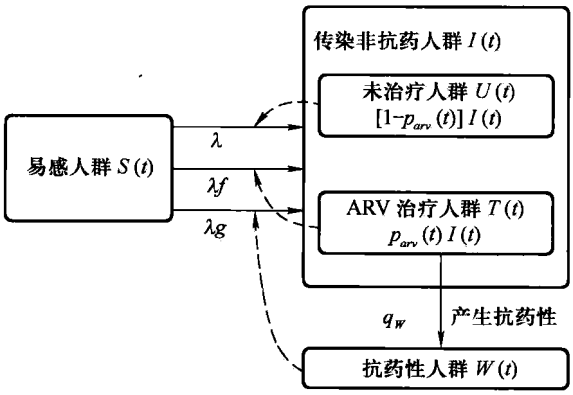


图 6-9 考虑抗药性的 SI 模型的转移关系图

在图6-9中，由于接受ARV治疗的感染者一旦产生了抗药性就对其停药，因此，接受ARV治疗的人群数量$p_{arv}(t)I(t)$可以由当年可用的AIDS基金总量决定，根据$I(t)$和当年可用基金总量可以计算出$p_{arv}(t)$.

易感者与未治疗人群、治疗人群和抗药人群接触后都有可能感染HIV，而且三类人群的传染率不同，一旦感染后，首先成为感染非抗药人群，之后按照一定的比例成为抗药性人群. 设单位时间有q_W的接受ARV治疗的感染者产生抗药性，具有抗药性的HIV感染人群的数量为$W(t)$，具有抗药性的人群的传染率为λg，则各类人群之间的转化模型为

$$\begin{cases}\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=S(t)r_c\lambda[U(t)+fT(t)+W(t)g]-q_WT(t),\\ \dfrac{\mathrm{d}W(t)}{\mathrm{d}t}=q_WT(t),\\ \dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=-S(t)r_c\lambda[U(t)+fT(t)+W(t)g],\\ U(t)=p_{arv}I(t),\\ U(t)+T(t)=I(t),\\ I(t)+W(t)+S(t)=N.\end{cases}$$

将各个状态变量的初值代入方程中，求解微分方程即可得到各个状态变量随时间变化的规律.

6.3.6 模型1.6：同时考虑ARV治疗抗药性和疫苗的传播模型

1. 问题的分析与假设

在模型1.4中同时考虑了ARV治疗和HIV疫苗对疫情的影响，在模型1.5中考虑了ARV治疗和抗药性对疫情的影响，组合模型1.4和模型1.5即可得到同时考虑ARV治疗和HIV疫苗时的传播模型.

模型假设同模型1.4及模型1.5，但需去除最后一条，即不考虑疫苗这一假设.

2. 传播模型的建立

组合模型1.4和模型1.5即可，得到同时考虑ARV治疗抗药性和疫苗的传播状态转移图，如图6-10所示.

相比模型1.4，该模型中增加了一类抗药性人群，该类人群具有传染性，其传染性为未接受治疗的传染者的g_r倍. 未接种疫苗的易感人群和接种疫苗的易感人群都有可能通过三种类型被感染，分别是与感染HIV未治疗人群、感染HIV接受治疗人群和接受治疗后产生抗药性人群.

各类人群之间的转化关系模型为

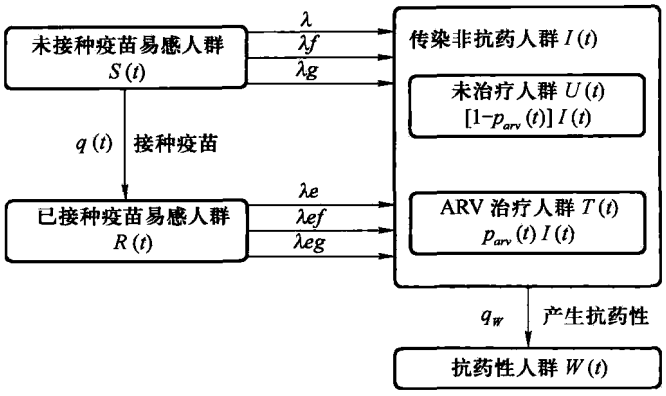


图 6-10 同时考虑 ARV 治疗抗药性和疫苗的 HIV 传播模型转移关系图

$$
\begin{cases}
\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=r_c\lambda[S(t)+R(t)e][U(t)+T(t)f+W(t)g]-T(t)q_W,\\
\dfrac{\mathrm{d}R(t)}{\mathrm{d}t}=q(t)-r_c\lambda R(t)e[U(t)+T(t)f+W(t)g],\\
\dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=-q(t)-r_c\lambda S(t)[U(t)+T(t)f+W(t)g],\\
\dfrac{\mathrm{d}W(t)}{\mathrm{d}t}=T(t)q_W,\\
U(t)+T(t)=I(t),\\
T(t)=I(t)p_{arv},\\
I(t)+S(t)+R(t)=N,\\
I(10)=I_0,\\
W(0)=W_0,\\
S(10)=S_0,\\
R(10)=0.
\end{cases}
$$

其中 I_0是模型 1.5 中当 $t=10$ 时传染非抗药人群的总数；W_0是模型 1.5 中当 $t=10$ 时传染抗药人群的总数；S_0 是模型 1.5 中当 $t=10$ 时易感人群的总数.

6.3.7 模型的分析与评价

上述几个模型都忽略了人口的出生、死亡以及迁移的影响，并且假设 HIV 传染率为恒定值. 模型思路清晰、简单，这是传统传染病 SI 模型的直接推广，对于短期的预测基本上能够反映实际情况，但对于长期的预测可能结果不理想. 在此理想的假设下，将得到所有易感者最终都将感染 HIV，显然这个结果是无法让人信服的.

究其原因，问题出在假设上. 如果是急性传染病，如流感，SARS 等几个月就可能蔓延全国的传染病，则作此假设可能是合理的. 但是 HIV 的传播速

度要远比流感慢，HIV 的潜伏期长达 10 年，在如此漫长的传播过程中，新生人口不断补充到人群中，同时也有感染者不断死亡．因此实际的传播速度要比模型中的更慢一些，一般是不会出现整个国家的人都感染 HIV 的情况．

此外，认为传染率 λ 为定值也过于简单，事实上 HIV 的传播途径多样，而且不同途径的传播特性差异性较大，笼统地将所有传播途径用一个传染率 λ 来量化会导致模型结果的不准确，或不符合实际的情况．这就需要进一步研究更精确的模型．

6.4 模型 2：基于 Logistic 模型的 HIV 传播模型

模型 2 是基于 Logistic 传播模型来研究 HIV 的传播问题．该模型在 SI 模型的基础上增加考虑传染率随疫情可变而得到的模型．由于政府的控制作用，传染率会随着疫情的加重而降低．

6.4.1 模型 2.1：HIV 自由传播模型

1. 问题的分析与假设

针对 HIV 非常危急的国家，政府部门不会放任疫情自由扩散．当疫情严重时，政府部门会采取宣传教育、加强监管、控制传播途径等手段降低传染率．因此，传染率 λ 会受到疫情的影响，当疫情严重（即 $I(t)$ 比较大）时，传染率就会减小，反之会增大．

该规律和人口增长的 Logistic 模型一致．在 Logisitic 人口模型中，人口的自然增长率会受到人口数量的限制，在这里传染者数量会受到国家容许上限的限制．

假设国家能够根据 HIV 疫情的严重程度适时制定有效措施控制病毒的扩散；不妨认为 HIV 扩散规律服从 Logistic 函数．

2. 传播模型的建立

事实上，在很多其他的领域都可以看到类似的模型．例如技术的传播、信息的传播、某些经济指标的增长、人口的增长等都是基于 Logistic 模型来研究的．在前面分析假设的基础上，建立如下 Logistic 增长模型：

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t}=\left[1-\frac{I(t)}{K_I}\right]\lambda_0 r_c S(t)I(t),$$

其中 K_I 表示一个国家可能容许的最大 HIV 感染者人数．该模型可以理解为在传染初期（$I\approx 0$），HIV 传染事件没有受到足够的重视，此时的传染率接近自然传染率 λ_0，随着疫情的恶化，政府部门开始重视控制 HIV，从而导致传染率随着疫情的严重而降低；疫情严重到一定程度后，接近该国容许上界时（$I\approx K_I$），政府的苛刻措施最终实现传染率 $\lambda=\left(1-\frac{I(t)}{K_I}\right)\lambda_0\approx 0$．对于上述模

型，状态转移图如图 6-11 所示.

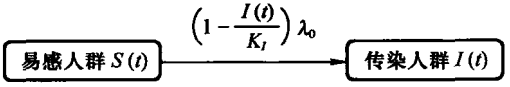


图 6-11 Logistic 模型状态转移关系图

在 Logistic 模型中，易感人群向传染人群转移的速度参数 λ 受传染人群规模的影响，传染人群规模越大，政府实施的控制措施就会越严厉，转移速度就会越小，最终不再增长. 综合考虑易感人群总数和人口变化情况，可得到如下微分方程模型：

$$\begin{cases}\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=\left[1-\dfrac{I(t)}{K_I}\right]\lambda_0 r_c S(t)I(t),\\ I(t)+S(t)=N(t),\\ I(0)=I_0,\\ S(0)=S_0.\end{cases}$$

其中 K_I 表示国家容许的 HIV 传染上限，其他变量同 SI 模型.

上述模型和模型 1.1 的不同之处在于该模型的传染率 λ 随传染人群总数的变化而变化，并且该模型中的人口总数 $N(t)$ 也随时间变化，$N(t)$ 的数值可以通过 UNAIDS 获得.

6.4.2 模型 2.2：考虑 ARV 治疗的 Logistic 传播模型

当 ARV 治疗介入后，Logistic 传播模型的状态转移图如图 6-12 所示.

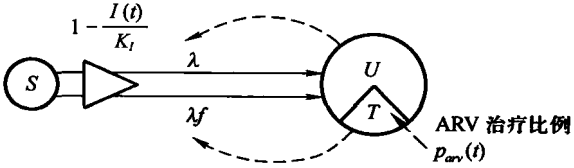


图 6-12 考虑 ARV 治疗的 SI 模型

如图 6-12 所示，与考虑 ARV 治疗的 SI 模型相比，Logistic 模型中易感人群向传染人群转移过程中始终受到传染人群总数 $I(t)$ 的限制，传播速度与简单 SI 相比相差一个上限抑制系数 $1-\frac{I(t)}{K_I}$. 此时传播模型可修正为

$$\begin{cases}\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=\left[1-\dfrac{I(t)}{K_I}\right]S(t)r_c\lambda[U(t)+fT(t)],\\ U(t)+T(t)=I(t),\\ I(t)+S(t)=N,\\ U(t)=p_{arv}I(t),\\ I(0)=I_0,\\ S(0)=S_0.\end{cases}$$

该模型仍然没有考虑 AIDS 死亡率高于非 AIDS 致死的死亡率，因此无法体现 ARV 治疗延长了 AIDS 寿命而对传染结果的影响.

6.4.3　模型 2.3：考虑 HIV 疫苗的 Logistic 传播模型

在这里与模型 1.3 的区别在于传染率 λ 会随感染人群比例 $I(t)$ 的变化而变化，即

$$\lambda(t)=\lambda_0\left[1-\frac{I(t)}{K_I}\right], \tag{6.4}$$

将（6.4）式代入模型 1.3 中，即得到考虑 HIV 疫苗的 Logistic 传播模型：

$$\begin{cases}\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=\lambda S(t)I(t)r_c+e\lambda R(t)I(t)r_c,\\ \dfrac{\mathrm{d}R(t)}{\mathrm{d}t}=q(t)-e\lambda R(t)I(t)r_c,\\ \dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=-q(t)-\lambda S(t)I(t)r_c,\\ I(t)+S(t)+R(t)=N(t),\\ \lambda(t)=\lambda_0\left[1-\dfrac{I(t)}{K_I}\right],\\ I(10)=I_0,\\ S(10)=S_0,\\ R(10)=0.\end{cases}$$

6.4.4　模型 2.4：考虑 ARV 治疗和 HIV 疫苗的 Logistic 传播模型

将（6.4）式代入模型 1.4 中，即得到考虑 ARV 治疗和 HIV 疫苗的 Logistic 传播模型：

$$\begin{cases}\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=r_c\lambda[S(t)+R(t)e][U(t)+fT(t)],\\ \dfrac{\mathrm{d}R(t)}{\mathrm{d}t}=q(t)-r_c\lambda eR(t)I(t)[U(t)+fT(t)],\\ \dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=-q(t)-r_c\lambda S(t)I(t)[U(t)+fT(t)],\\ \lambda(t)=\lambda_0\left[1-\dfrac{I(t)}{K_I}\right],\\ T(t)=I(t)p_{arv},\\ T(t)+U(t)=I(t),\\ I(t)+S(t)+R(t)=N(t),\\ I(10)=I_0,\\ S(10)=S_0,\\ R(10)=0.\end{cases}$$

6.4.5 模型2.5：考虑ARV治疗抗药性的Logistic传播模型

当考虑ARV抗药性时，传染人群分为两大类，一大类是非抗药性传染人群 $I(t)$，包括接受ARV治疗未产生抗药性的HIV传染人群和未接受治疗的传染人群；另一大类是产生抗药性的人群 $W(t)$. 此时的传染率同时受到两类人群总数的影响，即

$$\lambda(t)=\lambda_0\left[1-\frac{I(t)+W(t)}{K_I}\right], \tag{6.5}$$

将（6.5）式代入模型1.5中，即可得到考虑ARV治疗抗药性的Logistic传播模型：

$$\begin{cases}\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=S(t)r_c\lambda(t)[U(t)+fT(t)+W(t)g]-q_WT(t),\\ \dfrac{\mathrm{d}W(t)}{\mathrm{d}t}=q_WT(t),\\ \dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=-S(t)r_c\lambda(t)[U(t)+fT(t)+W(t)g],\\ \lambda(t)=\lambda_0\left[1-\dfrac{I(t)+W(t)}{K_I}\right],\\ U(t)=p_{aru}I(t),\\ U(t)+T(t)=I(t),\\ I(t)+W(t)+S(t)=N(t),\\ I(0)=I_0,\\ S(0)=S_0,\\ W(0)=W_0.\end{cases}$$

6.4.6 模型2.6：考虑HIV疫苗和ARV治疗抗药性的Logistic传播模型

将（6.5）式代入模型1.6中，即可得到考虑ARV治疗抗药性和HIV疫苗的Logistic传播模型：

$$
\begin{cases}
\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=r_c\lambda(t)[S(t)+R(t)e][U(t)+T(t)f+W(t)g]-T(t)q_W,\\
\dfrac{\mathrm{d}R(t)}{\mathrm{d}t}=q(t)-r_c\lambda(t)R(t)e[U(t)+T(t)f+W(t)g],\\
\dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=-q(t)-r_c\lambda(t)S(t)[U(t)+T(t)f+W(t)g],\\
\dfrac{\mathrm{d}W(t)}{\mathrm{d}t}=T(t)q_W,\\
\lambda(t)=\lambda_0\left[1-\dfrac{I(t)+W(t)}{K_I}\right],\\
U(t)+T(t)=I(t),\\
T(t)=I(t)p_{arv},\\
I(t)+S(t)+R(t)=N(t),\\
I(10)=I_0,\\
W(10)=W_0,\\
S(10)=S_0,\\
R(10)=0.
\end{cases}
$$

6.4.7　模型的分析与评价

模型 2 中的 6 个模型分别是对模型 1 中相应模型的传染率和人口总数进行了修正，这种做法在进行短期预测时具有一定的合理性，尤其是当 HIV 传播较为危急时，能够解释政府加大防控力度所产生影响的客观事实.

假设某国的人口总数比较稳定（例如南非），在典型参数下得到模型 2 的解如图 6-13 所示.

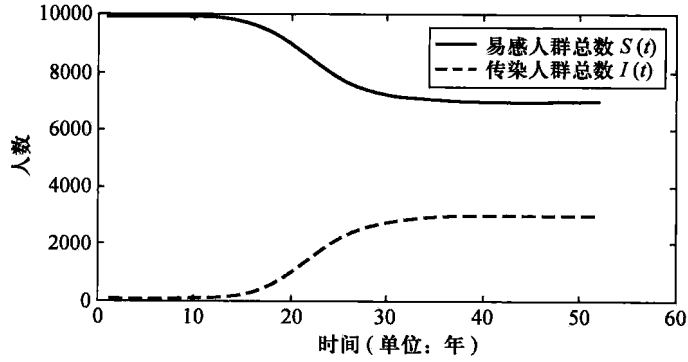


图 6-13　典型参数下两类人群总数变化曲线

从图 6-13 可以看出，若该国的人口数量稳定，最终的传染人群总数将稳定在 K_I 值上. 该模型无法得到下降的 $I(t)$ 变化规律，也就是说，无论政府部门怎

么控制，HIV传播永远也无法控制住，显然这个结果与实际是不太相符的.

6.5 模型3：考虑出生和死亡的HIV传播模型

模型3是在考虑出生和死亡的情况下，利用SI模型来研究HIV的传播问题. 该模型是在模型1的基础上增加考虑出生和死亡对疫情的影响而得到的.

6.5.1 模型3.1：考虑出生和死亡的HIV自由传播模型

1. 问题的分析与假设

在模型1和模型2中都没有考虑出生人口和死亡人口对HIV传播的影响是欠妥的. 由于HIV的传染期长达10年，如此长的传染期内出生的婴儿数量和死亡的人员数量将极大地影响HIV的传播规律. 此外，HIV可以通过母婴传播，只有考虑新出生和死亡才能得到通过母婴传播而导致的新增HIV感染者的数量真实值.

假设人口增长服从Logisitic模型；在研究时间区间内，人口寿命分布没有显著改变；传染率 λ 为常量.

图6-14给出了HIV传播时各个状态之间的转化关系.

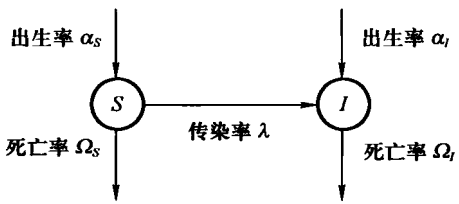


图6-14 考虑出生和死亡的HIV/AIDS状态转移关系图

在图6-14中，S 表示易感人群人群，I 表示传染人群. 新生婴儿既有感染的婴儿，也有未感染的婴儿，一个未感染的人有可能变为一个感染HIV的人，并且无法回到未感染状态. 无论是感染者还是未感染者，又有一定的比例进入死亡状态.

2. 出生率和死亡率的确定

设人群总数为 $N(t)=S(t)+I(t)$，其中 $S(t)$ 表示易感人群人数，$I(t)$ 表示感染HIV病毒但未接受治疗的人数. 设 γ 是人口的自然增长率，可以得到Logistic人口增长模型：

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t}=\gamma\left(1-\frac{N(t)}{K}\right)N(t),$$

其中 K 是人口最大容量上限. 模型中的参数 γ，K 可以通过对历史数据拟合得到. 由此可以计算出 t 时刻的总人口 $N(t)$ 以及增长率 $\gamma(t)=\gamma\left(1-\frac{N(t)}{K}\right)$.

由于人口的增长率为

$$\gamma(t)=\alpha(t)-\Omega(t), \tag{6.6}$$

其中 $\alpha(t)$ 是出生率，出生率 $\alpha(t)$ 定义为第 t 年新生婴儿的数量除以当年的总人口数 $N(t)$，$\Omega(t)$ 是死亡率，实际上死亡率 $\Omega(t)$ 由两部分组成，一部分是非 AIDS 致死 Ω_s，另一部分是 AIDS 致死 Ω_I，它们之间满足关系式：

$$\Omega(t)=\frac{I(t)\Omega_I+S(t)\Omega_S}{N(t)}, \tag{6.7}$$

事实上，Ω_S 由各国的医疗卫生条件、饮食习惯、体育锻炼情况等因素决定，Ω_I 由各国的 AIDS 治疗开展情况决定. Ω_S 和 Ω_I 可以通过历史数据统计得到，并且在相当长的一段时间内不发生改变. 根据 Ω_I 和 Ω_S 可以计算得到第 t 年 HIV 易感人群的死亡数量 $D_S(t)=S(t)\Omega_S$，HIV 传染人群的死亡数量为 $D_I(t)=I(t)\Omega_I$. 将 (6.7) 式代入 (6.6)，即可得到每年的出生率 $\alpha(t)$.

3. 新生婴儿感染数的确定

目前在世界范围内，母婴传播的 AIDS 病例大约占 HIV 感染者总数的 5% ~10%. UNAIDS 和 WHO 在 2005 年的报告中显示，在欠发达国家母婴传播率达到 35%，发达国家的母婴传播率为 1% ~5%.

当第 t 年的生育率为 $\alpha(t)$ 时，假设 HIV 感染不影响孕龄妇女的生育率，则 HIV 感染人群在第 t 年将生产出 $\alpha(t)I(t)$ 的新生婴儿，其中感染 HIV 的婴儿数量为 $B_I(t)=p_I\alpha(t)I(t)$，其中 p_I 是母婴传播率，当年新生 HIV 阴性的婴儿数量为

$$B_S(t)=(1-p_I)\alpha(t)I(t)+\alpha(t)S(t).$$

4. 传播模型与结果分析

当出生率和死亡率确定之后，根据传播过程的状态转移图，可以建立如下微分方程：

$$\begin{cases}\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=B_I(t)+\lambda I(t)S(t)-D_I(t),\\[2ex] \dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=B_S(t)-\lambda I(t)S(t)-D_S(t).\end{cases} \tag{6.8}$$

在一组确定的参数下，例如：$K=11\ 000$；$\gamma=0.01$；$\Omega_S=0.000\ 05$，$\Omega_I=0.002$，$\lambda=0.000\ 000\ 5$，$I(2\ 005)=1$，$S(2\ 005)=9\ 999$. 求解结果如图 6-15 所示.

从图 6-15 可以看出，在 Logistic 人口模型下，人口总数逐步增长，最后趋于一个定值 K. 初始阶段传染源数量较少，易感人群总数在初始阶段出现缓慢增长；随着传染人群总数的增长，越来越多的易感者感染 HIV，并最终稳定在一个平衡状态，此时每年的 HIV 新增数量等于 HIV 死亡数量，新生 HIV 阴性数量也等于非 AIDS 致死数量.

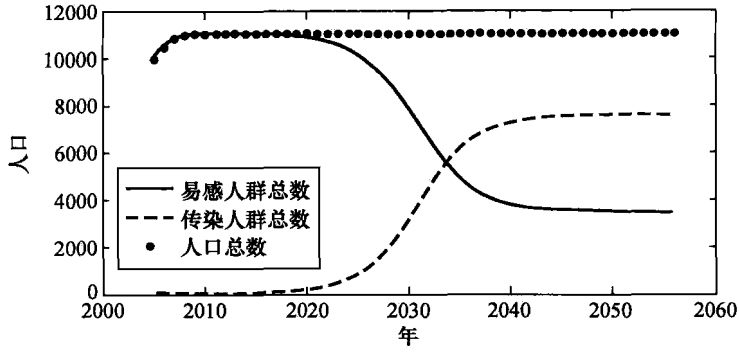


图 6-15 考虑出生率和死亡率的 HIV 传播模型的变化规律图

要考虑未来较长一段时间的传播结果，则可以通过求微分方程模型的平衡点来讨论. 在 Logistic 人口预测模型下有人口总数为 $\lim\limits_{t\to\infty}N(t)=K$，此时出生率等于死亡率，即 $\lim\limits_{t\to\infty}\alpha(t)=\alpha$，$\lim\limits_{t\to\infty}\Omega(t)=\Omega$，且 $\alpha=\Omega$，于是由（6.7）式得：

$$K\Omega=I(t)\Omega_I+S(t)\Omega_S,$$

当出生率 $\alpha(t)$、死亡率 $\Omega_S(t)$，$\Omega_I(t)$ 为常量时，则微分方程组的平衡点为

$$\begin{cases}\lim\limits_{t\to\infty}[\alpha(t)I(t)p_I+\lambda I(t)S(t)-I(t)\Omega_I]=0,\\ \lim\limits_{t\to\infty}[\alpha(t)S(t)+\alpha(t)I(t)(1-p_I)-\lambda I(t)S(t)-S(t)\Omega_S]=0,\\ K\lim\limits_{t\to\infty}\alpha(t)=\lim\limits_{t\to\infty}I(t)\Omega_I+\lim\limits_{t\to\infty}S(t)\Omega_S,\\ \lim\limits_{t\to\infty}S(t)+\lim\limits_{t\to\infty}I(t)=K.\end{cases}$$

$$\Rightarrow\begin{cases}\lim\limits_{t\to\infty}I(t)=0,\ \lim\limits_{t\to\infty}S(t)=K, & I(0)=0\\ \lim\limits_{t\to\infty}S(t)=\dfrac{\Omega_I-\alpha p_I}{\lambda},\ \lim\limits_{t\to\infty}I(t)=K-\dfrac{\Omega_I-\alpha p_I}{\lambda}. & I(0)>0\end{cases}$$

从平衡点的表达式可以看出，影响疫情的因素主要包括：传染率 λ、出生率 α 新生婴儿感染风险 p_I 以及 AIDS 致死的死亡率 Ω_I. 下面分析不同参数对疫情的影响.

首先研究传染率 λ 对疫情的影响，图 6-16 给出了不同传染率下疫情的发展趋势.

从图 6-16 可以看出，当传染率较小时，传染率对疫情的影响不大；当传染率增大时，健康人口急剧下降，并最终稳定在一个较大危急的状态.

新出生婴儿感染 HIV 的风险 p_I 对疫情的影响效果如图 6-17 所示. 可以看出，新生儿感染 HIV 风险越大，对疫情的影响越大.

HIV 感染者死亡率 Ω_I 对疫情的影响效果如图 6-18 所示. 可以看出，HIV 感染者死亡率 Ω_I 越低，疫情越严重. 即说明通过药物治疗降低 HIV 感染者死亡率将给 HIV 的防控带来更大的难度.

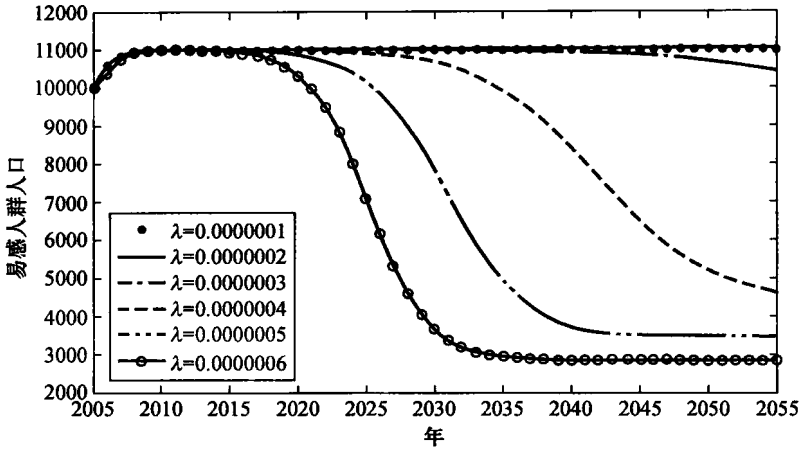


图 6-16　传染率 λ 对易感人群数量的影响

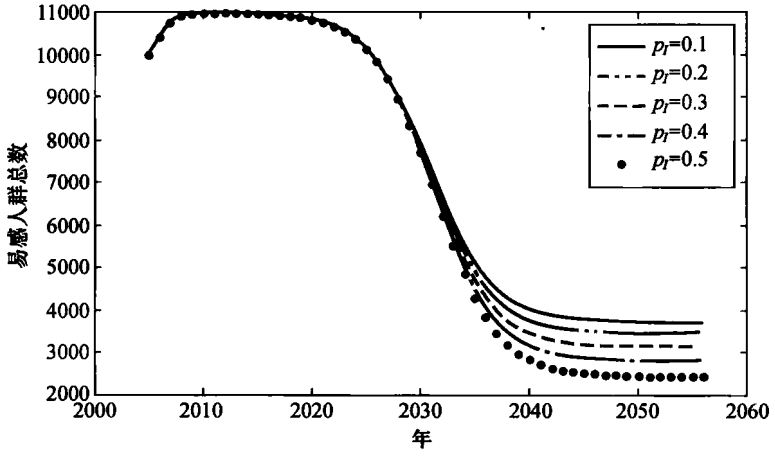


图 6-17　新生婴儿感染 HIV 的风险 P_I 对疫情的影响

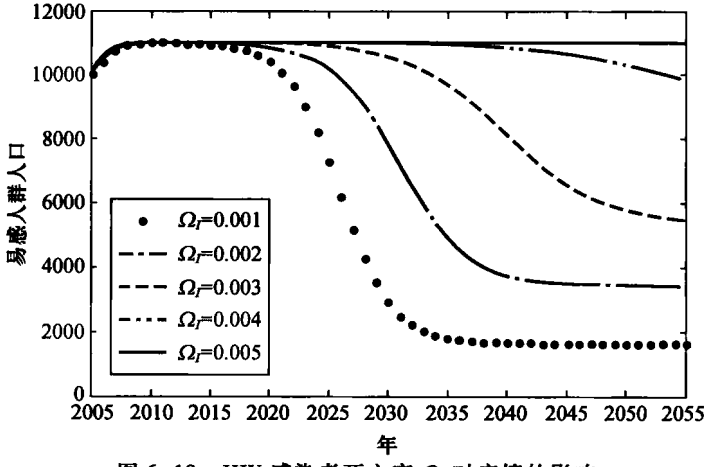


图 6-18　HIV 感染者死亡率 Ω_I 对疫情的影响

6.5.2 模型3.2：考虑ARV治疗的HIV传播模型

1. 问题的分析与假设

在人口增长服从Logistic模型的条件下，来研究引入ARV治疗的HIV传播规律. 假设ARV治疗能够在一定程度上降低传染率，易感者与接受ARV治疗的感染者接触后的传染风险由 λ 降低为 λf；ARV治疗能够降低死亡率，不妨设AIDS患者接受ARV治疗后，死亡率由 Ω_I 降低到 Ω_T；未接受治疗的母婴传播风险为 M，而接受ARV治疗后母婴传播风险为 M_T.

2. 传播模型的建立

当存在ARV治疗时，则将感染人群 I 分为两类，一类是感染未治疗人群 U，另一类是感染接受ARV治疗的人群 T. 各国政府在第 t 年可支配的ARV治疗经费决定了接受治疗的比例 $p_{arv}(t)$. 各状态转移关系如图6-19所示.

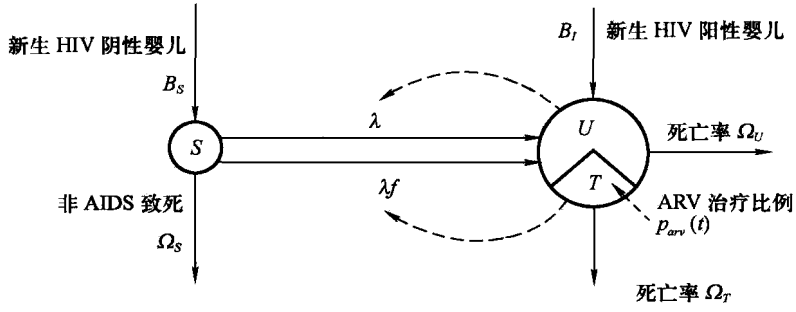


图6-19 考虑ARV治疗时HIV传播模型转移关系图

由此可得考虑ARV治疗和出生与死亡的HIV传播模型：

$$
\begin{cases}
\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=B_I(t)+\lambda U(t)S(t)+f\lambda T(t)S(t)-T(t)\Omega_T-U(t)\Omega_U,\\
\dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=B_S(t)-\lambda U(t)S(t)-f\lambda T(t)S(t)-S(t)\Omega_S,\\
I(t)=U(t)+T(t),\\
T(t)=p_{arv}(t)I(t),\\
\Omega(t)=\dfrac{T(t)\Omega_T(t)+U(t)\Omega_U(t)+S(t)\Omega_S(t)}{S(t)+U(t)+T(t)},\\
\alpha(t)=\gamma(t)+\Omega(t),\\
B_I(t)=\alpha(t)[U(t)M+T(t)M_T],\\
B_S(t)=\alpha(t)[U(t)(1-M)+T(t)(1-M_T)]+\alpha(t)S(t).
\end{cases}
$$

其中 p_I 表示母亲是HIV感染者，新生婴儿经过母婴感染的风险，$\gamma(t)$ 是通过Logistic人口模型得到的第 t 年人口增长率，需要根据历史数据预测得到.

6.5.3　模型3.3：考虑存在HIV疫苗的传播模型

1. 问题的分析与假设

在人口的增长服从Logistic模型的条件下，来研究存在疫苗时的HIV传播规律. 假设接种疫苗后能够将传染风险降低到接种前的 e 倍；单位时间可提供的疫苗数量为 $q(t)$，该值的大小由该国可支配的HIV疫苗资金决定；不考虑疫苗的有效期问题，即认为一旦接种疫苗，则可终身有效.

2. 传播模型的建立

在考虑存在HIV疫苗的情况下，易感人群被划分为未接种疫苗的人群 S 和已接种疫苗的人群 R，各人群之间的转化关系如图6-20所示.

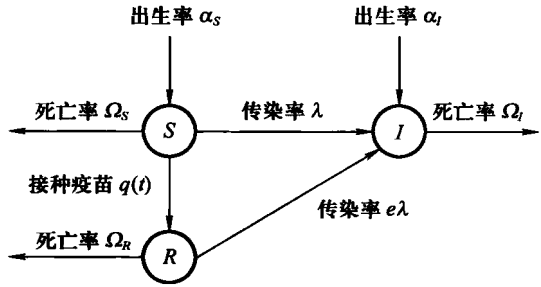


图6-20　考虑疫苗时的HIV传播模型转移关系图

由图6-20，可以得到考虑HIV疫苗时的传播模型：

$$\begin{cases}\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=B_I(t)+\lambda S(t)I(t)+e\lambda R(t)I(t)-I(t)\Omega_I,\\ \dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=B_S(t)-\lambda S(t)I(t)-q(t)-S(t)\Omega_S,\\ \dfrac{\mathrm{d}R(t)}{\mathrm{d}t}=q(t)-e\lambda R(t)I(t)-R(t)\Omega_R,\\ \Omega(t)=\dfrac{I(t)\Omega_I(t)+R(t)\Omega_R(t)+S(t)\Omega_S(t)}{I(t)+R(t)+S(t)},\\ \alpha(t)=\gamma(t)+\Omega(t),\\ B_I(t)=p_I\alpha(t)I(t),\\ B_S(t)=(1-p_I)\alpha(t)I(t)+\alpha(t)S(t).\end{cases}$$

这是一个较复杂的微分方程组模型，需要用数值方法求解.

6.5.4　模型3.4：考虑ARV治疗和存在HIV疫苗的传播模型

1. 问题的分析与假设

在人口增长服从Logistic模型的条件下，来研究加入ARV治疗和存在HIV疫苗的情况下HIV传播规律. 假设接种疫苗后能够将传染风险降低到接种前

的 e 倍；单位时间国家可提供的疫苗数量为 $q(t)$，该数值可由该国可支配的HIV疫苗资金决定；单位时间内接受ARV治疗的AIDS患者占全部患者的比例为 p_{arv}；不考虑疫苗的有效期问题，治疗前母婴传播风险为 M；ARV治疗后母婴传播风险为 M_T.

2. 传播模型的建立

在这里同时考虑ARV治疗和HIV疫苗情况，各状态之间的转移关系如图6-21所示.

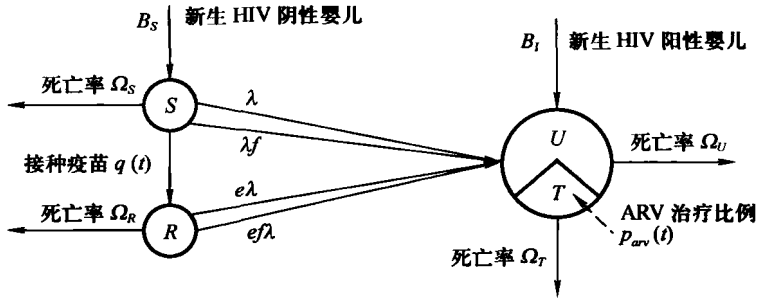


图6-21 考虑疫苗和治疗的HIV传播模型转移关系图

由图6-21，可建立如下的传播模型：

$$\begin{cases}\dfrac{dI(t)}{dt}=B_I(t)+\lambda[S(t)+eR(t)][U(t)+fT(t)]-u(t)\Omega_u-T(t)\Omega_T,\\ \dfrac{dS(t)}{dt}=B_S(t)-\lambda S(t)[U(t)+fT(t)]-S(t)\Omega_S-q(t),\\ \dfrac{dR(t)}{dt}=q(t)-e\lambda S(t)[U(t)+fT(t)]-R(t)\Omega_R,\\ \Omega(t)=\dfrac{U(t)\Omega_U(t)+T(t)\Omega_T(t)+R(t)\Omega_R(t)+S(t)\Omega_S(t)}{U(t)+T(t)+R(t)+S(t)},\\ \alpha(t)=\gamma(t)+\Omega(t),\\ B_I(t)=\alpha(t)[MU(t)+M_TT(t)],\\ B_S(t)=\alpha(t)[(1-M)U(t)+(1-M_T)T(t)]+\alpha(t)S(t).\end{cases}$$

这是一个较复杂的微分方程组模型，需要用数值方法求解.

6.5.5 模型3.5：考虑ARV治疗和抗药性的传播模型

1. 问题的分析与假设

在人口的增长服从Logistic模型的条件下，来研究加入ARV治疗和抗药性影响的情况下HIV传播规律. 假设单位时间内接受ARV治疗的AIDS患者占全部患者的比例为 p_{arv}；接受ARV治疗后，产生抗药性的比例为 q_W；未接受治疗的母婴传播风险为 M；接受ARV治疗后母婴传播风险降低到 M_T；产生抗药性后，母婴传播风险为 M_W；接受ARV治疗的患者一旦产生抗药性，其传播

风险将变为治疗前的 g 倍.

2. 传播模型的建立

若考虑抗药性对传播的影响，此时单位时间内有一定比例的接受 ARV 治疗的 HIV 感染者产生抗药性，成为抗药性人群 W. 此时易感人群有 3 种方式感染 HIV：与未接受治疗的 HIV 感染者接触、与接受治疗未产生抗药性的 HIV 传染者接触以及与产生抗药性的 HIV 感染者接触，各种方式的传染率不同. 各类人群之间的转化关系如图 6-22 所示.

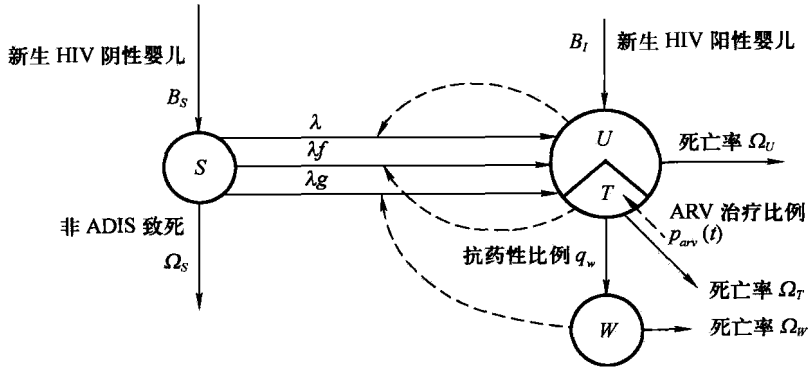


图 6-22　考虑 ARV 治疗及抗药性的 HIV 传播模型转移关系图

根据状态转移图 6-22，可以得到如下的传播模型：

$$
\begin{cases}
\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=B_I(t)+S(t)\lambda[U(t)+fT(t)+gW(t)]-T(t)\Omega_T-U(t)\Omega_U-T(t)q_W,\\
\dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=B_S(t)-S(t)\lambda[U(t)+fT(t)+gW(t)]-S(t)\Omega_S,\\
\dfrac{\mathrm{d}W(t)}{\mathrm{d}t}=T(t)q_W-W(t)\Omega_W\\
I(t)=U(t)+T(t),\\
T(t)=p_{arv}(t)I(t),\\
\Omega(t)=\dfrac{T(t)\Omega_T(t)+U(t)\Omega_U(t)+S(t)\Omega_S(t)+W(t)\Omega_W}{S(t)+U(t)+T(t)+W(t)},\\
\alpha(t)=\gamma(t)+\Omega(t),\\
B_I(t)=\alpha(t)[MU(t)+M_TT(t)+M_WW(t)],\\
B_S(t)=\alpha(t)[(1-M)U(t)+(1-M_TT(t))+(1-M_WW(t))]+\alpha(t)S(t).
\end{cases}
$$

6.5.6　模型 3.6：考虑 ARV 治疗抗药性和存在 HIV 疫苗的传播模型

1. 问题的分析与假设

在人口的增长服从 Logistic 模型的条件下，来研究具有 ARV 治疗抗药性和存在 HIV 疫苗的情况下 HIV 传播规律. 假设单位时间内接受 ARV 治疗的

AIDS患者占全部患者的比例为 p_{arv}；接受ARV治疗后，产生抗药性的比例为 q_W；不考虑ARV治疗对降低母婴传播风险的作用；受ARV治疗的患者一旦产生抗药性，其传播风险将变为治疗前的 g_r 倍；HIV疫苗可以将传染风险降低到未接种疫苗前的 e 倍（$e<1$）；不考虑疫苗的有效期.

2. 传播模型的建立

在同时考虑ARV抗药性和存在HIV疫苗的情况下，将易感者可以分为未接种疫苗的易感者 S 和接种疫苗的易感者 R；传染者包括未接受治疗的传染者 U、接受治疗未产生抗药性的传染者 T 和接受治疗产生抗药性的传染者 W. 这时两类易感者 S 和 R 各有3种方式被感染，分别是与 U、T 和 W 接触，各类接触的传染率不同. 各类人群之间的转化关系如图6-23所示.

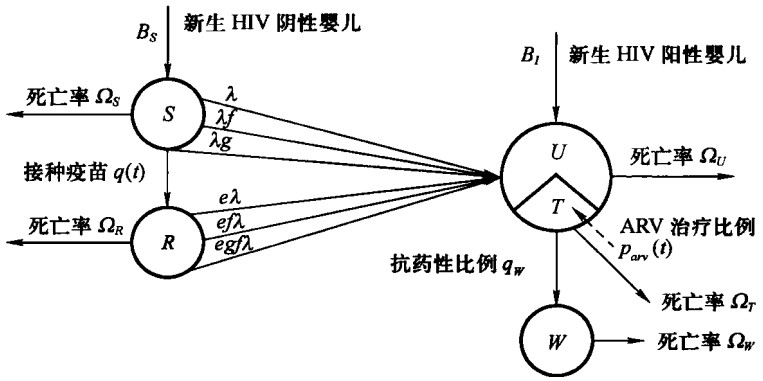


图6-23 同时考虑HIV疫苗、ARV治疗和抗药性的传播模型转移关系图

根据各类人群转移关系图6-23，则可得到如下的传播模型：

$$
\begin{cases}
\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=B_I(t)+\lambda[S(t)+eR(t)][U(t)+fT(t)+gW(t)]-T(t)(\Omega_T+q_W)-U(t)\Omega_U,\\
\dfrac{\mathrm{d}W(t)}{\mathrm{d}t}=T(t)q_W-W(t)\Omega_W,\\
\dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=B_S(t)-q(t)-\lambda S(t)[U(t)+fT(t)+gW(t)]-S(t)\Omega_S,\\
\dfrac{\mathrm{d}R(t)}{\mathrm{d}t}=q(t)-e\lambda S(t)[U(t)+fT(t)+gW(t)]-R(t)\Omega_R,\\
\Omega(t)=\dfrac{U(t)\Omega_U(t)+T(t)\Omega_T(t)+R(t)\Omega_R(t)+S(t)\Omega_S(t)+W(t)\Omega_W(t)}{U(t)+T(t)+R(t)+S(t)+W(t)},\\
\alpha(t)=\gamma(t)+\Omega(t),\\
B_I(t)=\alpha(t)[MU(t)+M_TT(t)+M_WW(t)],\\
B_S(t)=\alpha(t)[(1-M)U(t)+(1-M_TT(t))+(1-M_WW(t))]+\alpha(t)S(t).
\end{cases}
$$

6.5.7 模型的分析与评价

该模型考虑了出生和死亡对 HIV 传播的影响，共给出了 6 个不同的模型. 相比较而言，比模型 2 更精确、更符合实际. 此外，该模型针对各个参数的比较讨论可以直接应用于 AIDS 防控措施的制定当中.

但该模型对于 HIV 的传播的描述过于简单，只是用简单的传染率 λ 来解释传染过程，没有考虑不同的传播途径和传播规律的差异性，从这一角度还可以做进一步的完善，这正是下面要研究的问题.

6.6 模型 4：HIV 的多级传播模型

模型 4 是考虑不同传染阶段的 HIV 传播模型. 该模型将 SI 模型中的感染人群分解为多个传染阶段，并且不同感染阶段的传播特性不一致.

6.6.1 模型 4.1：HIV 多级自由传播模型

1. 问题的分析与假设

在模型 1.1 中，假设一个易感者一旦被感染 HIV，其传染率 λ 是一个定值，不随时间阶段的改变而改变. 在实际中，处于不同阶段的 HIV 感染者其传染性是不同的[8]. 其中窗口期（从感染到血清中检测出 HIV 阳性）的传染性最强；之后进入一个长达近 10 年的潜伏期，潜伏期内也具有传染性，但传染性较小；最后进入发病期，传染性较之潜伏期有所增强. 为了准确反映不同时期传染性的不同，在这里建立 HIV 的多级传染模型.

假设短期内出生率和死亡率是稳定的；单位时间内，从一级感染状态向下一级感染状态转移的比例是稳定的；不考虑人口迁移的情况.

2. 传播模型的建立

多级模型和 SI 模型比较相似，但是需要考虑更多的因素. 该模型将从 HIV 感染病毒到患 AIDS 的过程划分为 n 个阶段，建立分段马尔可夫过程模型，并假设每个传染阶段的传染率是不同的. 每个阶段为一个状态，各个状态之间的转移关系如图 6-24 所示.

（1）母婴传播模型

图 6-24 的下半部分解释了 HIV 的母婴传播过程. 易感者和各个阶段的感染者都可能生产婴儿. 易感妇女的新生婴儿仍然属于易感人群. 而 HIV 感染妇女的新生婴儿也有一定风险感染 HIV，而且处于不同传染阶段的妇女新生婴儿感染 HIV 的风险也不同. 假设 HIV 感染第 k 阶段的育龄妇女新生婴儿感染 HIV 的风险为 r_k.

由于 HIV 感染的育龄妇女仍然可能生产 HIV 阴性的婴儿，所以新生 HIV

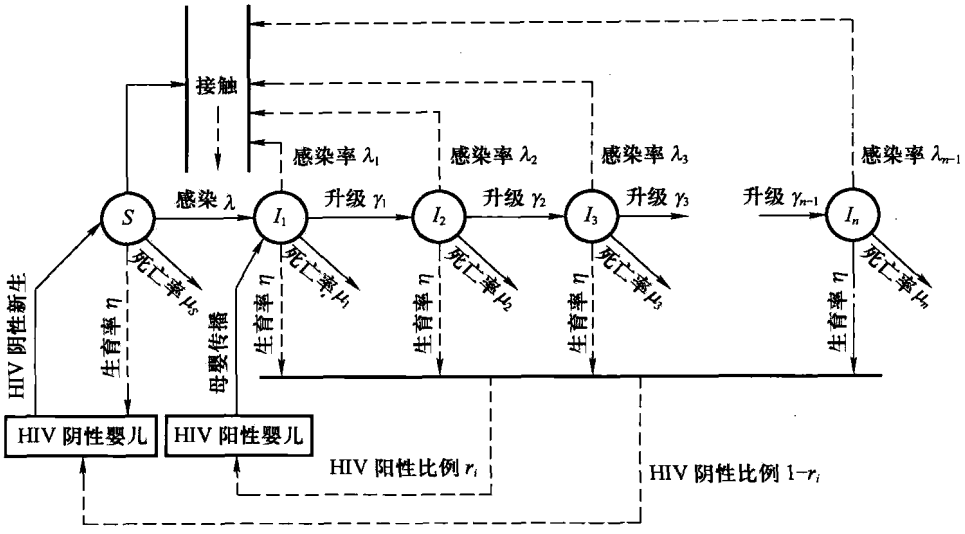


图6-24 多阶段HIV传播过程状态转移示意图

阴性婴儿的数量不仅取决于易感育龄妇女数量，还取决于各个阶段感染HIV的育龄妇女数量. 假设易感育龄妇女占易感人群 S 的比例为 C_S，育龄妇女在单位时间内的生育率为 η. 如果出生率系数为 $M_S=\eta C_S$，则单位时间内由易感人群导致的新增易感婴儿数量为 SM_S.

假设是否感染HIV不影响感染妇女生育率，也不影响育龄妇女的比例，则在单位时间内，传染阶段 k 的人群新生婴儿数量为 $I_kC_S\eta$，其中导致的新生婴儿HIV为阳性的数量为 $I_kC_S\eta r_k$，记 $C_S\eta r_k=M_k$，则传染阶段 k 的感染者导致的母婴传播病例数为 I_kM_k，新生婴儿HIV为阴性的数量为 $I_k(M_S-M_k)$. 于是在单位时间内新增婴儿HIV为阳性的数量为

$$B_I=\sum_{i=1}^{n} I_kM_k,$$

而在单位时间内新增HIV阴性婴儿的数量为

$$B_S=SM_S+\sum_{k=1}^{n} I_k(M_S-M_k),$$

（2）交叉传染模型

图6-24的上半部分解释了HIV交叉感染的过程. 一个易感者可能与各个传染阶段的传染者接触，在与各个阶段的传染者接触后被其感染的可能性不同. 假设易感者与第 k 个传染阶段的传染者接触的概率与第 k 个阶段的人数 $I_k(t)$ 成正比，其比例系数为 ρ. 一个易感者与第 k 个传染阶段的传染者接触后被感染的概率为 β_k，则一个易感者在单位时间内被感染的概率为

$$P=\sum_{i=1}^{n} I_k\rho\beta_i,$$

记 $\lambda_i=\rho\beta_i$ 表示感染率，t 时刻总计有 S 个易感者，则在单位时间内由易感者感染为 HIV 感染者的数量 ΔI_{S2I} 为

$$\Delta I_{S2I}=S\sum_{i=1}^{n}I_k\lambda_i,$$

(3) 多阶段 HIV 传播模型

导致易感人群 S 增加的因素包括易感者的新生易感婴儿和感染者的新生 HIV 阴性婴儿；导致易感人群减少的因素包括易感人群的非 AIDS 致死死亡及易感者被感染 HIV. 可以得到易感人群的传播微分方程为

$$\frac{\mathrm{d}S}{\mathrm{d}t}=B_S-\mu_S S-\Delta I_{S2I},$$

其中 μ_S 是非 AIDS 致死死亡率.

导致第一阶段传染人群 I_1 增加的因素包括：由易感者转化为第一传染阶段的传染者以及通过母婴传播而生产的 HIV 阳性婴儿；导致第一阶段的传染人群减少的因素包括：发病死亡或者感染转化到下一个传染阶段. 第一阶段传染人群的传播动力学方程为

$$\frac{\mathrm{d}I_1}{\mathrm{d}t}=\Delta I_{S2I}+B_I-(\mu_1+\gamma_1)I_1,$$

其中 μ_1 是第一阶段的死亡率，γ_1 是在单位时间内转化为第二阶段的比例.

导致第 k（$k=2$，$3\cdots$，$n-1$）阶段传染人群 I_k 增加的因素是从 $k-1$ 阶段转移到第 k 阶段的人群. 导致其减少的因素包括发病死亡，从 k 阶段发展转移到第 $k+1$ 阶段的人群. 第 k（$k=2$，$3\cdots$，$n-1$）阶段传染人群 I_k 的传播动力学方程为

$$\frac{\mathrm{d}I_K}{\mathrm{d}t}=\gamma_{k-1}I_{k-1}-(\mu_k+\gamma_k)I_k,$$

其中 $k=2$，3，$\cdots$，$n-1$，μ_i 是第 k 个阶段的死亡率，γ_k 是由第 k 个传染阶段转移到第 $k+1$ 个传染阶段的比例.

导致第 n 个阶段传染人群 I_n 增加的因素是从第 $n-1$ 个传染阶段转移到第 n 个传染阶段的人数；导致其减少的因素是第 n 个传染阶段的死亡率. 第 n 个传染阶段 I_n 的传播动力学方程为

$$\frac{\mathrm{d}I_n}{\mathrm{d}t}=\gamma_{n-1}I_{n-1}-\mu_n I_n,$$

其中 μ_n 是第 n 个阶段的死亡率，γ_{n-1} 是由第 $n-1$ 个传染阶段转移到第 n 个传染阶段的比例.

综上所述，得到多阶段 HIV 传播微分方程模型为

$$\begin{cases}\dfrac{dS}{dt}=B_S-\mu_S S-S\sum\limits_{i=1}^{n}I_i\lambda_i,\\ \dfrac{dI_1}{dt}=S\sum\limits_{i=1}^{n}I_i\lambda_i+B_I-(\mu_1+\gamma_1)I_1,\\ \dfrac{dI_k}{dt}=\gamma_{k-1}I_{k-1}-(\mu_k+\gamma_k)I_k,k=2,3,\cdots,n-1,\\ \dfrac{dI_n}{dt}=\gamma_{n-1}I_{n-1}-\mu_n I_n,\\ B_S=SM_S+\sum\limits_{k=1}^{n}I_k(M_S-M_k),\\ B_I=\sum\limits_{i=1}^{n}I_k M_k.\end{cases}$$

6.6.2 模型4.2：考虑ARV治疗的HIV多级传播模型

1. 问题的分析与假设

在建立考虑ARV治疗的HIV多级传播模型时，需要在模型4.1的基础上，将各个传染阶段的人群分为接受ARV治疗的人群和未接受ARV治疗的人群，再分别研究这两类人群的传播动力学模型.

假设ARV治疗会影响接受治疗人员的传染率、母婴传播风险和死亡率；第k阶段未接受治疗者的传染率为λ_k，从第k阶段开始接受治疗后，其传染率为治疗前的f_k倍，即此时传染率为$f_k\lambda_k$；第k阶段未接受治疗者的母婴传播率为$M_{k,U}$，从第k阶段开始接受治疗后，母婴传播率为$M_{k,T}$；不考虑ARV治疗的抗药性；传染者一旦开始接受ARV治疗，将能够一直享受政府资助的ARV治疗，不会出现中间断药的情况.

这里与模型1～模型3的情况有所不同，在模型1～模型3中，假设ARV治疗经费是逐年拨付的，已经开始接受ARV治疗的HIV感染者具有优先接受治疗的权利. 在经费逐年增长的情况下，可以保证所有ARV治疗者不会出现中途停药现象. 在该假设下，每年正在接受ARV治疗的比例与当年拨付经费有关. 这种情况主要适用于较落后的国家，国家无法一次性支付HIV感染者终生的治疗费用，只能逐年拨付. 该方案的好处是降低了资金压力，能够让更多的HIV感染者更早地接受ARV治疗，不足之处是抵御风险的能力不足，若中间出现资金链断裂，则可能导致大面积停药，给防治效果造成严重影响.

这里假设只要感染者申请到ARV治疗基金，就能够享受ARV治疗直到死亡，政府拨付的经费决定了每年可批复ARV治疗申请的数量. 该模型可以确保每个申请成功的HIV感染者能够得到充分的治疗，不足之处是资金利用率较低. 该方案比较适合于发达国家.

2. 多级传播模型的建立

考虑 ARV 治疗后，在模型 4.1 的基础上，增加第 k 阶段接受 ARV 治疗的人群 T_k，此时第 k 阶段的状态转移关系如图 6-25 所示.

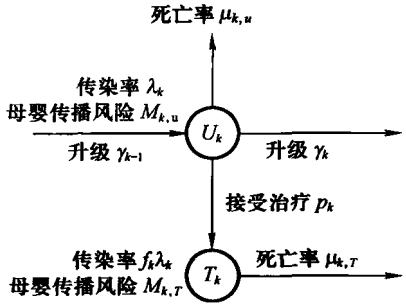


图 6-25　考虑 ARV 治疗后第 k 阶段的转移关系图

图 6-25 中，U_k 表示第 k 阶段未接受治疗的传染人群，T_k 表示第 k 个阶段开始接受 ARV 治疗的传染人群. 接受治疗之前，第 k 阶段的传染率为 $\lambda_{k,U}$，母婴传播风险为 $M_{k,U}$. 从第 k 个阶段开始接受治疗，此时传染率为 $\lambda_{k,T}$，母婴传播风险为 $M_{k,T}$. 在单位时间内，政府资金能够为第 k 阶段的 p_k 个传染者提供 ARV 治疗.

（1）新生婴儿数量确定模型

考虑 ARV 治疗后，未感染 HIV 的新生婴儿的母亲可能属于易感人群 S，也可能是处于第 k 个感染阶段未接受治疗的人群 U_k，也可能是从第 k 个感染阶段开始接受治疗的人群 T_k. 在单位时间内新生的 HIV 阴性婴儿的总量为

$$B_S = SC_S\eta + C_S\eta\sum_{k=1}^{n}[T_k(1 - M_{k,T}) + U_k(1 - M_{k,U})],$$

其中 C_S 是育龄妇女比例，η 是育龄妇女生育率. 在单位时间内新增 HIV 阳性婴儿的数量为

$$B_I = C_S\eta\sum_{k=1}^{n}(T_kM_{k,T} + U_kM_{k,U}).$$

（2）新增感染人群确定模型

每个时间周期内新增感染者的数量受到如图 6-26 所示因素的影响.

根据图 6-26，易感者人群 S 所满足的微分方程模型为

$$\frac{dS}{dt} = B_S - S\mu_S - S\sum_{k=1}^{n}(\lambda_{k,U}U_k + \lambda_{k,T}T_k),$$

初始感染人群 U_1 所满足的微分方程模型为

$$\frac{dU_1}{dt} = B_I + S\sum_{k=1}^{n}(\lambda_{k,U}U_k + \lambda_{k,T}T_k) - p_1 - U_1(\mu_{1,U} + \gamma_1)$$

初始治疗人群 T_1 所满足的微分方程模型为

$$\frac{dT_1}{dt} = p_1 - T_1\mu_{1,T}$$

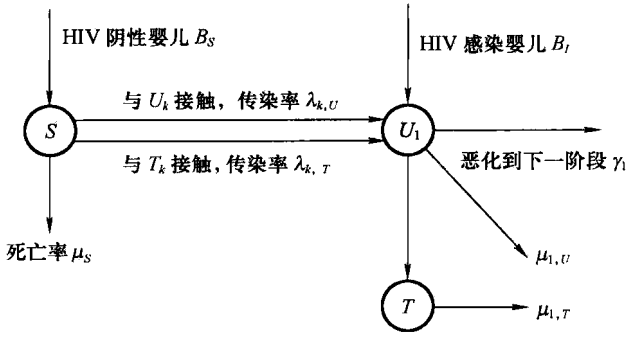


图 6-26 新增感染人群数量模型

（3）各感染期的状态转移关系

如图6-25所示，在每个时间周期内有 $U_{k-1}\gamma_{k-1}$ 的未治疗传染者进入第 k 个感染阶段，有 $U_k\gamma_k$ 的未治疗传染者离开第 k 个感染阶段，有 $U_k\mu_{k,U}$ 的第 k 阶段未治疗传染者死亡，有 U_kp_k 的第 k 阶段未治疗传染者开始接受治疗，有 $T_k\mu_{k,T}$ 的接受治疗者死亡.

此时，在单位时间内第 k 阶段（$k=2$，3，…，$n-1$）未接受治疗的HIV感染者 U_k 的改变量为

$$\frac{\mathrm{d}U_k}{\mathrm{d}t}=U_{k-1}\gamma_{k-1}-p_k-U_k(\gamma_k+\mu_{k,U}),$$

第 k 阶段开始接受治疗的人群 T_k 的改变量为

$$\frac{\mathrm{d}T_k}{\mathrm{d}t}=p_k-T_k\mu_{k,T},$$

（4）多级传播微分方程模型

综上所述，可得到考虑ARV治疗的HIV多级传播微分方程模型为

$$\begin{cases}\dfrac{\mathrm{d}S}{\mathrm{d}t}=B_S-S\mu_S-S\displaystyle\sum_{k=1}^{n}(\lambda_{k,U}U_k+\lambda_{k,T}T_k),\\ \dfrac{\mathrm{d}U_1}{\mathrm{d}t}=B_I+S\displaystyle\sum_{k=1}^{n}(\lambda_{k,U}U_k+\lambda_{k,T}T_k)-p_1-U_1(\mu_{1,U}+\gamma_1),\\ \dfrac{\mathrm{d}U_k}{\mathrm{d}t}=U_{k-1}\gamma_{k-1}-p_k-U_k(\gamma_k+\mu_{k,U}),k=2,3,\cdots,n-1,\\ \dfrac{\mathrm{d}U_n}{\mathrm{d}t}=U_{n-1}\gamma_{n-1}-p_n-U_n\mu_{n,U},\\ \dfrac{\mathrm{d}T_k}{\mathrm{d}t}=p_k-T_k\mu_{k,T},k=1,2,\cdots,n,\\ B_S=SC_S\eta+C_S\eta\displaystyle\sum_{k=1}^{n}[T_k(1-M_{k,T})+U_k(1-M_{k,U})],\\ B_I=C_S\eta\displaystyle\sum_{k=1}^{n}(T_kM_{k,T}+U_kM_{k,U}),\\ \gamma_n=0.\end{cases}$$

6.6.3　模型 4.3：考虑 HIV 疫苗的多级传播模型

1. 问题的分析与假设

针对实际中存在 HIV 疫苗的情况，来研究 HIV 多级传播问题. 假设不考虑疫苗的有效期，即对易感者一旦接种疫苗，则终生有效；HIV 疫苗能够有效降低 HIV 感染风险，设接种疫苗后能将传播风险降低到接种前的 e 倍；在单位时间内有 q 的易感人群接种疫苗.

2. 多级传播模型的建立

当实际中存在 HIV 疫苗时，并不影响新生婴儿是否感染，也不会影响各个感染阶段的转化，只能影响通过交叉感染的数量. 此时易感人群的转移关系如图 6-27 所示.

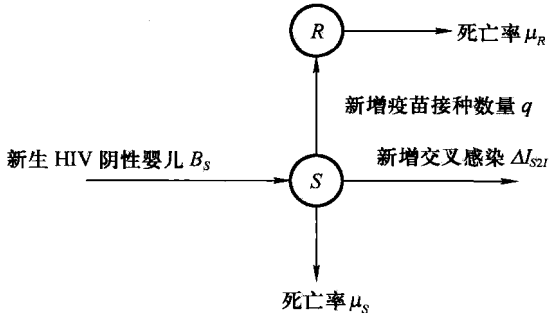


图 6-27　考虑疫苗时易感人群状态转移关系图

在图 6-27 中，单位时间内易感人群 S 的改变数量为

$$\frac{\mathrm{d}S}{\mathrm{d}t}=B_S-q-S\mu_S-\Delta I_{S2I}$$

在单位时间内新增交叉感染病例的来源包括 S 与 I_k 接触后感染和 R 与 I_k 接触后感染，交叉感染的过程如图 6-28 所示.

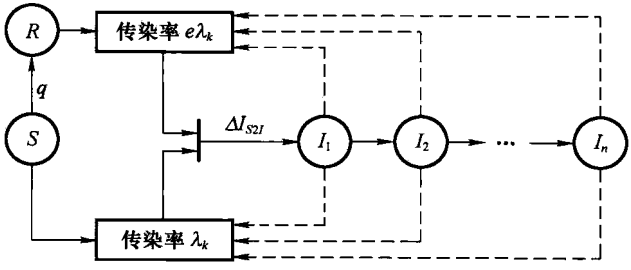


图 6-28　考虑 HIV 疫苗时单位时间新增交叉感染数量示意图

如图 6-28，两类人群的接触导致新增交叉感染病例数量 ΔI_{S2I} 为

$$\Delta I_{S2I}=(S+Re)\sum_{k=1}^{n}\lambda_k I_k\ ,$$

免疫人群 R 的改变数量为

$$\frac{dR}{dt}=q-R\mu_R,$$

第一个感染阶段 I_1 在单位时间内发生如图 6-29 所示的改变. 根据图6-29得到第一传染阶段的微分方程模型为

$$\frac{dI_1}{dt}=B_I-I_1(\mu_1+\gamma_1)+(S+Re)\sum_{k=1}^{n}\lambda_k I_k$$

第 k 个阶段 I_k 在单位时间内发生如图 6-30 所示的改变.

根据图 6-30 可得到第 k 传染阶段（$k=2, 3, \cdots, n$）的微分方程模型为

$$\frac{dI_k}{dt}=I_{k-1}\gamma_{k-1}-I_k(\mu_k+\gamma_k).$$

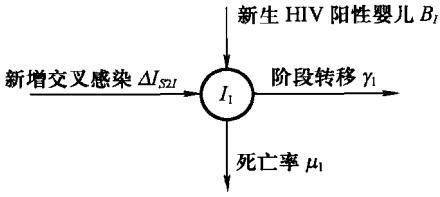


图 6-29 第一阶段人群转移图

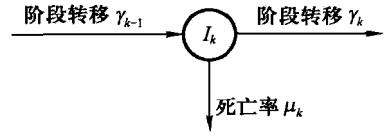


图 6-30 第 k 阶段人群转移关系图

综上所述，可以得到考虑 HIV 疫苗时的多级传播微分方程模型为

$$\begin{cases}\dfrac{dS}{dt}=B_S-q-S\mu_S-(S+Re)\sum\limits_{k=1}^{n}\lambda_k I_k,\\ \dfrac{dR}{dt}=q-R\mu_R,\\ \dfrac{dI_k}{dt}=I_{k-1}\gamma_{k-1}-I_k(\mu_k+\gamma_k),\ k=2,3,\cdots,n,\\ \dfrac{dI_1}{dt}=B_I-I_1(\mu_1+\gamma_1)+(S+Re)\sum\limits_{k=1}^{n}\lambda_k I_k,\\ S+R+\sum\limits_{k=1}^{n}I_k=N,\\ \gamma_n=0.\end{cases}$$

6.6.4 模型4.4：考虑ARV治疗和HIV疫苗的多级传播模型

1. 问题的分析与假设

组合考虑模型4.2和模型4.3，则可得到综合考虑ARV治疗和HIV疫苗的传播模型.

假设第 k 阶段未接受治疗者的传染率为 λ_k，从第 k 阶段开始接受治疗后，其传染率为治疗前的 f_k 倍，即此时传染率为 $f_k\lambda_k$；第 k 阶段未接受治疗者的

母婴传播率为 $M_{k,U}$，从第 k 阶段开始接受治疗后，母婴传播率为 $M_{k,T}$；一旦传染者开始接受 ARV 治疗，就能够一直治疗，不会出现中间断药的情况. 不考虑疫苗的有效期，即一旦接种疫苗，则终生有效；HIV 疫苗能够有效降低 HIV 感染风险，不妨设接种疫苗后能将传播风险降低到接种前的 e 倍；单位时间内有 q 的易感人群接种疫苗.

2. 多级传播模型的建立

（1）新生婴儿数量的确定

在实际中，HIV 疫苗的使用并不影响新生婴儿是否感染 HIV，因此新生婴儿数量的确定模型与模型 4.2 一致，即

$$B_S = SC_S\eta + C_S\eta\sum_{k=1}^{n}[T_k(1 - M_{k,T}) + U_k(1 - M_{k,U})],$$

$$B_I = C_S\eta\sum_{k=1}^{n}(T_kM_{k,T} + U_kM_{k,U}).$$

（2）新增感染人群数量的确定

在每个时间周期内新增感染者的数量都受到如图 6-31 所示因素的影响.

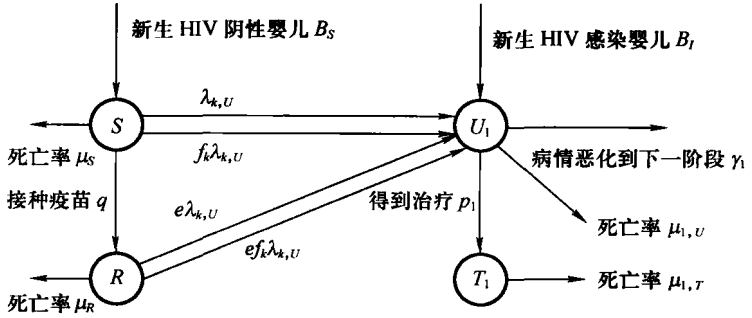


图 6-31　新增感染数量模型转移关系图

根据图 6-31，易感者人群 S 所满足的微分方程模型为

$$\frac{dS}{dt} = B_S - S\mu_S - S\sum_{k=1}^{n}\lambda_{k,U}(U_k + fT_k) - q,$$

接种疫苗人群 R 所满足的微分方程模型为

$$\frac{dR}{dt} = q - R\mu_R - Re\sum_{k=1}^{n}\lambda_{k,U}(U_k + fT_k),$$

初始感染人群 U_1 所满足的微分方程模型为

$$\frac{dU_1}{dt} = B_I + (S + Re)\sum_{k=1}^{n}\lambda_{k,U}(U_k + f_kT_k) - p_1 - U_1(\mu_{1,U} + \gamma_1),$$

初始治疗人群 T_1 所满足的微分方程模型为

$$\frac{dT_1}{dt} = p_1 - T_1\mu_{1,T},$$

（3）感染期状态转移方程

在感染期内的状态转移关系与模型4.2情况是一致的，即

$$\frac{\mathrm{d}U_k}{\mathrm{d}t}=U_{k-1}\gamma_{k-1}-p_k-U_k\ (\gamma_k+\mu_{k,U}),$$

第 k 阶段开始接受治疗的人群 T_k 的改变量为

$$\frac{\mathrm{d}T_k}{\mathrm{d}t}=p_k-T_k\mu_{k,T},$$

（4）多级传播微分方程模型

综上所述，可以得到考虑ARV治疗后的HIV多级传播的微分方程模型：

$$\begin{cases}
\dfrac{\mathrm{d}S}{\mathrm{d}t}=B_S-S\mu_S-S\sum\limits_{k=1}^{n}\lambda_{k,U}(U_k+fT_k)-q,\\
\dfrac{\mathrm{d}R}{\mathrm{d}t}=q-R\mu_R-Re\sum\limits_{k=1}^{n}\lambda_{k,U}(U_k+fT_k),\\
\dfrac{\mathrm{d}U_1}{\mathrm{d}t}=B_I+(S+Re)\sum\limits_{k=1}^{n}\lambda_{k,U}(U_k+f_kT_k)-p_1-U_1(\mu_{1,U}+\gamma_1),\\
\dfrac{\mathrm{d}U_k}{\mathrm{d}t}=U_{k-1}\gamma_{k-1}-p_k-U_k(\gamma_k+\mu_{k,U}),\ k=2,3,\cdots,n,\\
\dfrac{\mathrm{d}T_k}{\mathrm{d}t}=p_k-T_k\mu_{k,T},\ k=1,2,\cdots,n,\\
B_S=SC_S\eta+C_S\eta\sum\limits_{k=1}^{n}[T_k(1-M_{k,T})+U_k(1-M_{k,U})],\\
B_I=C_I\eta\sum\limits_{k=1}^{n}(T_kM_{k,T}+U_kM_{k,U}),\\
\gamma_n=0.
\end{cases}$$

6.6.5 模型4.5：考虑ARV治疗抗药性的多级HIV传播模型

1. 问题的分析与假设

接受ARV治疗的HIV感染者如果没有持续接受治疗或按照规定用药，就有很大风险产生抗药性．一旦产生了抗药性，其传染率、母婴传播风险和死亡率都会增加．此时将接受ARV治疗的人群分为接受治疗未产生抗药性人群和产生抗药性而停药的人群，并分别研究这两类人群的转化关系．

假设感染者一旦开始接受ARV治疗，就能够一直治疗，不会出现中间断药的情况；不考虑HIV疫苗的作用；第 k 阶段未接受治疗者的传染率为 λ_k，从第 k 阶段开始接受治疗后，其传染率为治疗前的 f_k 倍，即传染率为 $f_k\lambda_k$；从第 k 阶段开始产生抗药性后，其传染率为治疗前的 g_k 倍；在单位时间内处于 T_k 状态的HIV感染者产生抗药性的概率为 $q_{k,w}$．其他变量的符号说明如表6-2所示．其中 S 是易感人群，U_k 是处于感染的第 k 阶段但未接受治疗的人群，T_k

是从感染的第 k 个阶段开始接受治疗的人群，W_k 是 T_k 人群中产生抗药性的人群.

表 6-2　考虑 ARV 治疗的抗药性模型的符号说明

	S	U_k	T_k	W_k
传染率	0	$\lambda_{k,U}$	$\lambda_{k,T}=f_k\lambda_{k,U}$	$\lambda_{k,W}=g_k\lambda_{k,U}$
死亡率	μ_S	$\mu_{k,U}$	$\mu_{k,T}$	$\mu_{k,W}$
母婴传播风险	0	M	$M_{k,T}$	$M_{k,W}$

2. 多级传播模型的建立

（1）新生婴儿数量的确定模型

考虑 ARV 治疗及其抗药性问题后，未感染 HIV 的新生婴儿的母亲来自 4 种人群：S、U_k、T_k 以及 W_k. 因此，在单位时间内新增未感染 HIV 新生婴儿的数量为

$$B_S = SC_S\eta + C_S\eta\sum_{k=1}^{n}[T_k(1-M_{k,T}) + U_k(1-M_{k,U}) + W_k(1-M_{k,W})],$$

其中 C_S 是育龄妇女比例，η 是育龄妇女生育率.

在单位时间内新增 HIV 阳性婴儿的数量为

$$B_I = C_S\eta\sum_{k=1}^{n}(T_kM_{k,T} + U_kM_{k,U} + W_kM_{k,W}),$$

（2）新增感染人群数量的确定模型

在每个时间周期内新增感染者的数量会受到如图 6-32 所示各因素的影响.

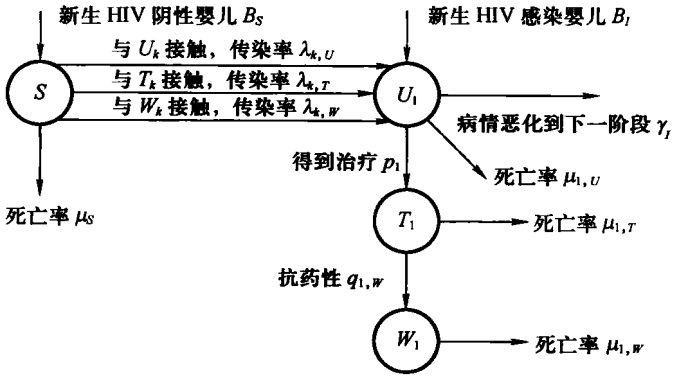


图 6-32　新增感染数量模型转移关系图

根据图 6-32，易感者人群 S 所满足的微分方程模型为

$$\frac{dS}{dt} = B_S - S\mu_S - S\sum_{k=1}^{n}\lambda_{k,U}(U_k + f_kT_k + g_kW_k),$$

初始感染人群 U_1 所满足的微分方程为

$$\frac{dU_1}{dt} = B_I + S\sum_{k=1}^{n}\lambda_{k,U}(U_k + f_kT_k + g_kW_k) - p_1 - U_1(\mu_{1,U} + \gamma_1),$$

初始治疗人群 T_1 所满足的微分方程为

$$\frac{dT_1}{dt} = p_1 - T_1(\mu_{1,T} + q_{1,W}),$$

初始抗药性人群 W_1 所满足的微分方程为

$$\frac{dW_1}{dt} = T_1q_{1,W} - W_1\mu_{1,W},$$

(3) 感染期内各状态转移关系

第 k 个感染阶段的状态转移关系如图 6-33 所示.

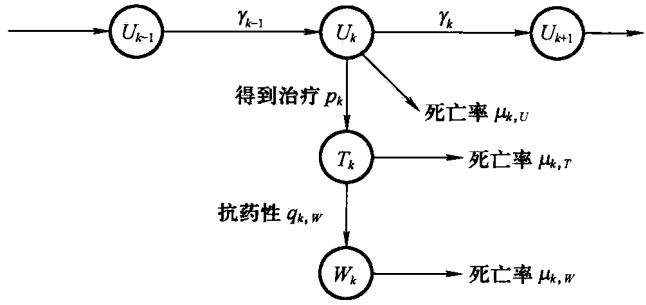


图 6-33 考虑 ARV 抗药性时第 k 个感染阶段状态转移图

如图 6-33 所示，在每个传染周期内，有 $U_{k-1}\gamma_{k-1}$ 的未治疗传染者进入第 k 个感染阶段；有 $U_k\gamma_k$ 的未治疗传染者离开第 k 个感染阶段；有 $U_k\mu_{k,U}$ 的第 k 阶段未治疗传染者死亡；有 U_kp_k 的第 k 阶段未治疗传染者开始接受治疗；有 $T_k\mu_{k,T}$ 的接受治疗者死亡.

此时，在单位时间内第 k 阶段 ($k=2, 3, \cdots, n-1$) 未接受治疗的 HIV 感染者 U_k 所满足的方程为

$$\frac{dU_k}{dt} = U_{k-1}\gamma_{k-1} - p_k - U_k(\gamma_k + \mu_{k,U}),$$

在单位时间内，第 k 阶段开始接受治疗的人群 T_k 所满足的方程为

$$\frac{dT_k}{dt} = p_k - T_k(\mu_{k,T} + q_{k,W}),$$

在单位时间内，T_k 人群中产生抗药性人群 W_k 的数量所满足方程为

$$\frac{dW_k}{dt} = T_kq_{k,W} - W_k\mu_{k,T},$$

(4) 多级传播微分方程模型

综上所述，可以得到考虑 ARV 治疗抗药性后的多级传播微分方程模型：

$$
\begin{cases}
\dfrac{\mathrm{d}S}{\mathrm{d}t} = B_S - S\mu_S - S\sum\limits_{k=1}^{n}\lambda_{k,U}(U_k + f_kT_k + g_kW_k), \\
\dfrac{\mathrm{d}U_1}{\mathrm{d}t} = B_I + S\sum\limits_{k=1}^{n}\lambda_{k,U}(U_k + f_kT_k + g_kW_k) - p_1 - U_1(\mu_{1,U} + \gamma_1), \\
\dfrac{\mathrm{d}U_k}{\mathrm{d}t} = U_{k-1}\gamma_{k-1} - p_k - U_k(\gamma_k + \mu_{k,U}),\ k = 2, 3, \cdots, n, \\
\dfrac{\mathrm{d}T_k}{\mathrm{d}t} = p_k - T_k(\mu_{k,T} + q_{k,W}),\ k = 1, 2, \cdots, n, \\
\dfrac{\mathrm{d}W_K}{\mathrm{d}t} - T_kq_{k,W} - W_k\mu_{k,T},\ k = 1, 2, \cdots, n, \\
B_S = SC_S\eta + C_S\eta\sum\limits_{k=1}^{n}[T_k(1 - M_{k,T}) + U_k(1 - M_{k,U}) + W_k(1 - M_{k,W})], \\
B_I = C_S\eta\sum\limits_{k=1}^{n}(T_kM_{k,T} + U_kM_{k,U} + W_kM_{k,W}), \\
\gamma_n = 0.
\end{cases}
$$

6.6.6　模型4.6：考虑ARV抗药性和HIV疫苗的多级HIV传播模型

组合模型4.4和模型4.5即可得到综合考虑ARV治疗抗药性及HIV疫苗的多级HIV传播模型. 同样，综合模型4.4和模型4.5的假设即可得到该模型.

首先，在实际中是否开展疫苗推广项目，对新生感染HIV婴儿的数量没有直接影响. 在这里，单位时间内新增未感染HIV新生婴儿数量和感染HIV新生婴儿的数量与模型4.5相同.

然后，考虑新增感染人群的数量变化. 在这里同时考虑ARV治疗抗药性和HIV疫苗的情况，初始阶段的状态转移关系如图6-34所示.

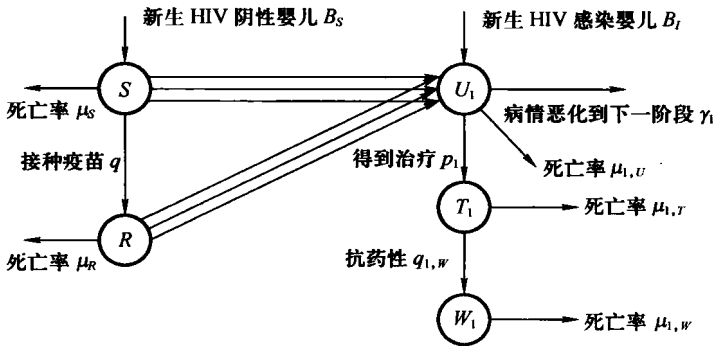


图6-34　新增感染数量模型的转移关系图

如图6-34，则可以得到易感者人群 S 所满足的微分方程模型为

$$
\frac{\mathrm{d}S}{\mathrm{d}t} = B_S - S\mu_S - q - S\sum_{k=1}^{n}\lambda_{k,U}(U_k + f_kT_k + g_kW_k).
$$

接种疫苗人群 R 所满足的微分方程模型为

$$\frac{\mathrm{d}R}{\mathrm{d}t} = q - R\mu_R - eR\sum_{k=1}^{n}\lambda_{k,U}(U_k + f_kT_k + g_kW_k).$$

初始感染人群 U_1 所满足的微分方程模型为

$$\frac{\mathrm{d}U_1}{\mathrm{d}t} = B_I + (S + eR)\sum_{k=1}^{n}\lambda_{k,U}(U_k + f_kT_k + g_kW_k) - p_1 - U_1(\mu_{1,U} + \gamma_1).$$

初始治疗人群 T_1 所满足的微分方程模型为

$$\frac{\mathrm{d}T_1}{\mathrm{d}t} = p_1 - T_1(\mu_{1,T} + q_{1,W}).$$

初始抗药性人群 W_1 所满足的微分方程模型为

$$\frac{\mathrm{d}W_1}{\mathrm{d}t} = T_1q_{1,W} - W_1\mu_{1,W}.$$

于是，在同时考虑ARV治疗抗药性和HIV疫苗的情况下，第 k 个阶段的状态转移关系模型与模型4.5是一致的.

综上所述，可以得到考虑ARV治疗抗药性和HIV疫苗的多级传播微分方程模型：

$$\begin{cases}
\dfrac{\mathrm{d}S}{\mathrm{d}t} = B_S - S\mu_S - q - S\sum\limits_{k=1}^{n}\lambda_{k,U}(U_k + f_kT_k + g_kW_k), \\
\dfrac{\mathrm{d}R}{\mathrm{d}t} = q - R\mu_R - eR\sum\limits_{k=1}^{n}\lambda_{k,U}(U_k + f_kT_k + g_kW_k), \\
\dfrac{\mathrm{d}U_1}{\mathrm{d}t} = B_I + (S + eR)\sum\limits_{k=1}^{n}\lambda_{k,U}(U_k + f_kT_k + g_kW_k) - p_1 - U_1(\mu_{1,U} + \gamma_1), \\
\dfrac{\mathrm{d}U_k}{\mathrm{d}t} = U_{k-1}\gamma_{k-1} - p_k - U_k(\gamma_k + \mu_{k,U}),\ k = 2,\ 3,\ \cdots,\ n, \\
\dfrac{\mathrm{d}T_k}{\mathrm{d}t} = p_k - T_k\ (\mu_{k,T} + q_{k,W}),\ k = 1,\ 2,\ \cdots,\ n, \\
\dfrac{\mathrm{d}W_k}{\mathrm{d}t} = T_kq_{k,W} - W_k\mu_{k,T},\ k = 1,\ 2,\ \cdots,\ n, \\
B_S = SC_S\eta + C_S\eta\sum\limits_{k=1}^{n}[T_k(1 - M_{k,T}) + U_k(1 - M_{k,U}) + W_k(1 - M_{k,W})], \\
B_I = C_S\eta\sum\limits_{k=1}^{n}(T_kM_{k,T} + U_kM_{k,U} + W_kM_{k,W}), \\
\gamma_n = 0.
\end{cases}$$

6.6.7 模型的分析与评价

多阶段HIV传播模型充分考虑了HIV感染者在不同感染阶段时，传染率、母婴传播风险和死亡率的差异性. 该模型可以认为是模型3的改进，将模型3

中的感染人群分为处于不同感染阶段的人群，并分别研究各个阶段人群数量、传播特征等对 HIV 传播的影响. 相比之下，该模型比模型 3 更接近实际情况.

但该模型仍然是笼统地用传染率 λ 来表示各种不同传播途径下的平均传播特性，并没有针对 HIV 的各个传播途径的特殊性进行分析，因此该模型不能精确地针对不同传播途径防控的资金分配问题进行优化分析. 下面的模型 5 和模型 6 中将针对不同的传播途径进行分析讨论.

6.7 模型 5：考虑不同传播途径的 HIV 传播模型

模型 5 是考虑不同传播途径的 SI 模型. 该模型是在 SI 模型的基础上将传播途径分为性传播、共用不洁注射器和母婴传播，并分别对不同的传播途径建立模型，从而得到 HIV 疫情变化情况.

6.7.1 模型 5.1：HIV 自由传播模型

1. 问题的分析与假设

事实上，HIV 的四种传播途径是有区别的. 母婴传播速度与育龄妇女感染 HIV 的比例及母婴传播风险有关；异性性行为传播速度与性伴侣数量、性行为频度、安全套使用比例、异性性行为传播风险等有关；同性性行为传播速度与男-男同性恋数量、同性性行为频度、安全套使用比例、同性性行为传播风险等有关；共用不洁注射器传播速度与感染 HIV 吸毒者的比例、共用注射器的范围、共用注射器频度、共用注射器传播风险等要素有关. 针对不同的国情和疫情，影响 HIV 传播的要素也不尽相同，简单地使用单一的传染率 λ 来解释传染过程无法精确地预测未来的疫情发展. 下面分别对四种传播途径建立相应的数学模型.

假设安全套使用率相对稳定，这只是最坏的情况，一般来说，随着性教育的推广与普及，安全套的使用率可能会有所提高.

假设安全套对于预防 HIV 感染是百分之百有效的，实际中安全套对于预防 HIV 感染的有效率达到 99% 以上.

假设所有的性行为中，被感染 HIV 的概率相同. 模型中不区分固定性伴侣和性滥交对于 HIV 感染率的影响.

假设所有的 HIV 感染者最终都死于 AIDS，不考虑其他原因导致 HIV 感染者死亡. 由于当前没有关于 HIV 携带者年龄和寿命分布的统计数据，因此简单地认为 HIV 感染者最终都死于 AIDS 病发. 这样做可能会略高估计 HIV 携带者的寿命，这个结果可以看做是 HIV 传播最坏的情况.

假设每个 HIV 携带者从感染到死亡的时间相同. HIV 感染者在平均“感染-死亡”时间后死亡.

假设不考虑人口迁移导致所研究国家人口的变化. 一般来说人口出入境数量相对平衡，且仅占总人口很小的一部分，对于总人口的影响很小.

2. 异性间的传播模型

据统计资料，美国成年人平均每人每周3次性行为，一年总计约160次性行为. 若包括一些非正常的高传染性行为在内，在美国平均每人每年有200次性行为[9].

资料显示，在一次无保护的性行为中，男性HIV病毒携带者传染给女性健康人概率要比女性HIV病毒携带者传染给男性健康人的概率大2倍. 考虑到不同性别之间的传播风险大小不同，分别建立传染模型.

首先考虑在一次女性HIV感染者与健康男性的性行为中，健康男性被感染的可能性.

在 t 时刻的一次性行为中，男性为健康者的概率为

$$P_M(t)=1-\frac{I^M(t)}{M(t)},$$

其中 $I^M(t)$ 是 t 时刻成年男性HIV病毒携带者总数，$M(t)$ 是 t 时刻成年男性总数.

在一次性行为中，女性为HIV携带者，健康男性被感染HIV的可能性为

$$P_{F2M}=P_M(t)(1-C)R_M,$$

其中 C 是性行为中安全套的使用比率，R_M 是女性HIV携带者在一次性行为中传染给男性的概率.

考虑在一年中，这位女性HIV感染者总计发生 $\rho=200$ 次性行为，则这位女性在第 t 年导致新增男性HIV感染者的数量为

$$N_{F2M}(t)=\rho P_{F2M}.$$

第 t 年总计有 $I^F(t)$ 位成年女性HIV病毒携带者，此时新增男性HIV感染者总数为

$$\Delta I_S^M(t)=I^F(t)N_{F2M}(t).$$

综上分析，可以得到第 t 年通过异性性行为传播而导致的新增男性HIV携带者的数量 $\Delta I_S^M(t)$ 为

$$\Delta I_S^M(t)=I^F(t)\rho\left[1-\frac{I^M(t)}{M(t)}\right](1-C)R_M. \tag{6.9}$$

同理可以得到新增女性HIV携带者的数量为

$$\Delta I_S^F(t)=I^M(t)\rho\left[1-\frac{I^F(t)}{F(t)}\right](1-C)R_F.$$

进而得到第 t 年通过异性性行为传播而导致的新增HIV携带者的数量为

$$\Delta I_S(t)=\Delta I_S^M(t)+\Delta I_S^F(t). \tag{6.10}$$

3. 同性恋间的传播模型

尽管在很多国家，传播HIV的主要途径是异性性交易传播，但在一些国

家，同性恋导致 HIV 传播的比例却很大. 假设同性恋者只进行同性性行为，并且在每次性行为中，同性性行为传播 HIV 的概率 R_H 要高于异性性行为[9]. 参照异性性行为传播模型来建立同性恋间的 HIV 传播模型：

$$\Delta I_H(t)=I_H(t)\rho_H\left[1-\frac{I_H(t)}{H(t)}\right]\ (1-C_H)\ R_H, \tag{6.11}$$

其中 $\Delta I_H(t)$ 为 t 时刻新增同性恋 HIV 感染者人数；$I_H(t)$ 是 t 时刻同性恋 HIV 感染者的人数；$H(t)$ 是 t 时刻同性恋人数；C_H 是同性性行为安全套的使用比例；ρ_H 是同性恋每年的同性性行为次数；R_H 是在一次同性性行为中，一个 HIV 携带者传染给一个易感者的概率.

4. 交叉使用毒品注射器传播模型

针对吸毒者交叉使用注射器的传播途径来建立传播模型. 假设不洁注射的次数占所有注射的次数比例是稳定的；对于每次不洁注射，感染 HIV 的概率是相同的.

在很多国家，交叉使用毒品注射器是导致 HIV 扩散的一个主要因素. 交叉使用毒品注射器导致的新增 HIV 感染的规律与同性性行为传播规律是类似的. 吸毒人群与同性恋人群对应；吸毒人员每年的注射次数 ρ_D 与同性恋每年的同性性行为次数 ρ_H 对应；不洁注射的比例 r_n 与同性性行为中未使用安全套的比例（$1-C$）对应；在一次不洁注射中感染 HIV 的概率 R_D 与在一次同性性行为中感染 HIV 的概率 R_H 对应. 类似同性性行为导致的新增 HIV 感染人数模型，则可得到 t 时刻由于不洁的毒品注射交叉感染的新增感染人数为

$$\Delta I_D(t)=I_D(t)\rho_D\left[1-\frac{I_D(t)}{d(t)}\right]r_n R_D, \tag{6.12}$$

其中 $d(t)$ 表示 t 时刻吸毒人数，在这里取 $r_n=0.35$，$R_D=0.8$.

5. 母婴传播导致新增 HIV 感染模型

针对母婴的传播途径，不考虑婴幼儿感染 HIV 导致的 HIV 的扩散. 新生感染 HIV 的婴儿一般不会传播 HIV，这是因为他们在有可能参与到各种传播活动（例如性行为、吸毒和生育）之前就已经发病死亡了. 育龄女性 HIV 携带者和正常女性一样生育. 在 HIV 携带者发病之前，女性 HIV 携带者并不知道自己携带 HIV，因此不影响其生育活动.

如果定义母婴传播率 P_I 为平均每位女性 HIV 感染者传染给自己孩子的比率. 在 t 时刻通过母婴传播导致的新增 HIV 感染者数量为

$$B_I(t)=I_S^F(t)\alpha_I(t)P_I, \tag{6.13}$$

其中 $\alpha_I(t)$ 是 t 时刻 HIV 感染人群的平均生育率，定义为 HIV 感染人群每年新生婴儿数与 HIV 感染人群总数的比值. 假设 HIV 感染人群的平均生育率与全国平均生育率 $\alpha(t)$ 相同，该数据可以通过人口统计数据获得.

6. AIDS致死亡数量模型

由于大多数的HIV感染者最终都死于AIDS病发，因此模型中不考虑其他原因导致HIV感染者死亡. 根据统计资料，从HIV感染到死亡的平均时间 $L=9$ 年. 在 $t+L$ 年AIDS的死亡人数等于第 t 年的新增HIV感染人数，其中包括第 t 年由易感者被传染为感染者的人数 $\Delta I_S(t)$ 和第 t 年的新生HIV携带者 $B_{HIV}(t)$. 即第 $t+L$ 年HIV致死人数为

$$D(t+L)=\Delta I_{S2I}(t)+B_I(t). \tag{6.14}$$

7. 多途径传播过程模型

这里建立一个基于迭代方法的确定性模型，该模型以一年为时间步长. 将人群分为两大类：HIV传染人群 I 和易感人群 S. 这两类之间存在如图6-35所示的转化关系.

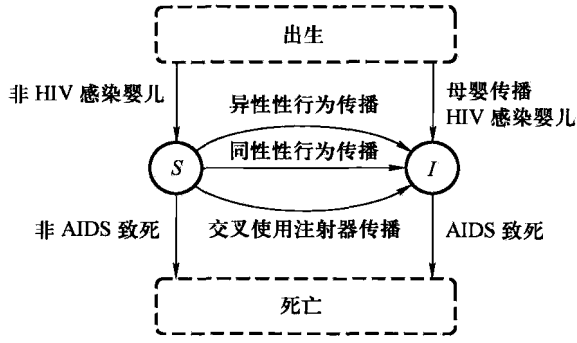


图6-35 HIV传染模型各类人群转化关系图

从图6-35可以看出，导致第 t 年易感人群数量 $S(t)$ 增加的因素是新生未感染HIV婴儿；导致 $S(t)$ 减少的因素包括通过三种传播途径而成为传染人群以及非AIDS致死的死亡.

导致感染人群数量 $I(t)$ 增加的原因包括：通过母婴传播而产生的HIV感染婴儿和通过三种传播途径而由易感人群变为传染人群；导致 $I(t)$ 减少的因素主要是AIDS致死. 在模型3中（6.8）式的基础上，建立如下微分方程模型：

$$\begin{cases}\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=B_I(t)+\Delta I_{S2I}(t)-D_I(t)\\[2ex]\dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=B_S(t)-\Delta I_{S2I}(t)-D_S(t)\end{cases} \tag{6.15}$$

其中 $I(t)$ 是 t 时刻HIV/AIDS人数，在模型中初始点 $t_0=2\,000$ 年；$S(t)$ 是 t 时刻易感人群人数；$\Delta I_{S2I}(t)$ 是 t 时刻通过三种传播途径导致由易感者变为感染者的新增感染人数；$B_S(t)$ 是 t 时刻新生未感染HIV的婴儿数量；$B_I(t)$ 是 t 时刻新生感染HIV的婴儿数量；$D_I(t)$ 是 t 时刻AIDS致死的死亡人数；$D_S(t)$ 是 t 时刻非AIDS致死的死亡人数；$B_I(t)$ 和 $D_I(t)$ 可以由模型3确定，也可

以通过该国的人口统计学数据得到.

下面讨论传染者的改变量 $\Delta I_{S2I}(t)$ 的确定模型.

因为 HIV/AIDS 主要有四种传播途径，异性性传播、同性性传播、交叉使用毒品注射器传播和母婴传播. 其中前三种传播途径可以将一个易感者传染为感染者，即 $I_{S2I}(t)$ 是由前三种传播途径决定的. 不同的传播途径导致的每天新增感染人数有所不同，每年新增转化感染人数 $I_{S2I}(t)$ 应为这三类传播途径导致的新增传染病人数之和，即

$$\Delta I_{S2I}(t)=\Delta I_S(t)+\Delta I_D(t)+\Delta I_H(t), \tag{6.16}$$

其中 $\Delta I_S(t)$ 表示异性性传播导致的每年新增 HIV 感染人数；$\Delta I_H(t)$ 表示同性性传播导致的每年新增 HIV 感染人数；$\Delta I_D(t)$ 表示交叉使用毒品注射器导致的每年新增 HIV 感染人数.

综上所述，将（6.9）式～（6.14）式代入（6.15）式和（6.16）式中，结合一个国家的人口统计数据建立人口模型，就可以预测第 t 年的新生婴儿数量 $B(t)$ 和死亡数量 $D(t)$，代入初始时刻（2000 年）各种人群数量的初值，通过迭代的方法即可得到这个国家在 2006—2050 年间任何一年的 HIV 携带者数量.

6.7.2 模型 5.2：考虑 ARV 治疗的 HIV 传播模型

如果实际中加入了 ARV 治疗方法，则通过 ARV 治疗会降低不同传播途径的传染风险，这个模型与模型 3 的相同之处是出生率（包括母婴传播途径）和死亡率相同；不同之处是新增感染人数不同. 下面针对不同点进行详细讨论.

在模型 5.1 假设的基础上，假设 ARV 治疗能够在一定程度上影响接受治疗者的传染率和寿命；这里不考虑 ARV 抗药性和 HIV 疫苗.

1. 考虑 ARV 治疗的异性性行为传播模型

图 6-36 给出了通过异性性行为传播 HIV 的状态转移关系示意图. 每年都有一定数量的 HIV 阴性男女进入成年阶段并成为易感人群，女性易感者 $S_S^F(t)$ 可能与两类传染人群接触，一类是接受 ARV 治疗的 HIV 感染者 $T_S^M(t)$，一类是未接受 ARV 治疗的 HIV 感染者 $U_S^M(t)$. 与两类人员接触的频率和传染风险也有所不同. 当该女性易感者感染 HIV 后，有一定比例感染者能够接受 ARV 治疗. 每个阶段都有一定概率的死亡. 同理可以得到男性易感者被感染的过程.

从图 6-36 可知，在单位时间内新增女性传染者数量 $\Delta U_S^F(t)$ 主要来自女性易感者与无治疗男性传染者以及进行 ARV 治疗的男性传染者接触. 其中女性易感人群 $S_S^F(t)$ 与无治疗男性传染人群 $U_S^M(t)$ 导致的新增女性传染人群数量 $\Delta U_S^{UF}(t)$ 为

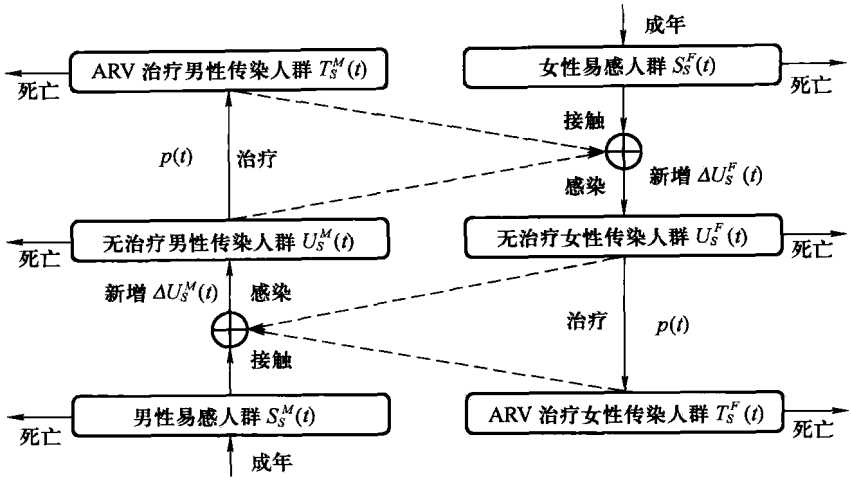


图6-36 考虑ARV治疗的异性性行为传播状态转移关系图

$$\Delta U_S^{UF}(t)=U_S^M(t)\rho_U\frac{S_S^F(t)}{F(t)}(1-C)R_{UF}. \tag{6.17}$$

其中ρ_U表示未接受治疗的传染者每年的性行为次数，$F(t)$是成年女性人数，C是安全套使用率，R_{UF}是HIV无治疗男性与易感女性发生一次性关系后易感女性感染HIV的概率.

女性易感人群$S_S^F(t)$与接受ARV治疗的男性传染人群$T_S^M(t)$接触导致的新增女性传染人群数量$\Delta U_S^{TF}(t)$为

$$\Delta U_S^{TF}(t)=T_S^M(t)\rho_T\frac{S_S^F(t)}{F(t)}(1-C)R_{TF}. \tag{6.18}$$

其中ρ_T表示接受ARV治疗的传染者每年的性行为次数，不妨取$\rho_T=j_{arv}\rho_U$；$M(t)$是成年男性人数，R_{TF}是接受ARV治疗的男性传染者与易感女性发生一次性关系后易感女性感染HIV的概率，一般来说$\rho_T\leqslant\rho_I$，$R_{TF}\leqslant R_{IF}$. 则单位时间内新增感染HIV女性数量$\Delta U_S^F(t)$为

$$\Delta U_S^F(t)=\Delta U_S^{TF}(t)+\Delta U_S^{UF}(t). \tag{6.19}$$

假设在实际中已实施了ARV资助计划，并且没有性别歧视，接受治疗的比例均为$p(t)$，此时

$$T_S^M=p(t)U_S^M(t).$$

由于$R_{TF}=fR_{UF}$，将（6.17）式和（6.18）式代入（6.19）式得

$$\Delta U_S^F(t)=\frac{S_S^F(t)}{F(t)}(1-C)U_S^M(t)R_{UF}\rho_U[1+p(t)fj_{arv}].$$

同理可得单位时间内新增感染HIV男性数量$\Delta U_S^M(t)$为

$$\Delta U_S^M(t)=\frac{S_S^M(t)}{M(t)}(1-C)U_S^F(t)R_{UM}\rho_U[1+p(t)fj_{arv}].$$

其中 $M(t)$ 是成年男性人数，R_{UM}是 HIV 无治疗女性与易感男性发生一次性关系后易感男性感染 HIV 的概率.

综上所述，在单位时间内，通过异性传播导致的新增 HIV 感染人数为

$$\Delta U_S(t)=\Delta U_S^F(t)+\Delta U_S^M(t). \tag{6.20}$$

2. 考虑 ARV 治疗的同性性行为传播模型

假设同性恋者不再是异性恋. 由于 HIV 主要是在男-男同性恋之间传播，因此不再考虑女-女同性恋的传播问题. 如果实际中存在 ARV 治疗方法，则同性性行为使 HIV 传播过程如图 6-37 所示.

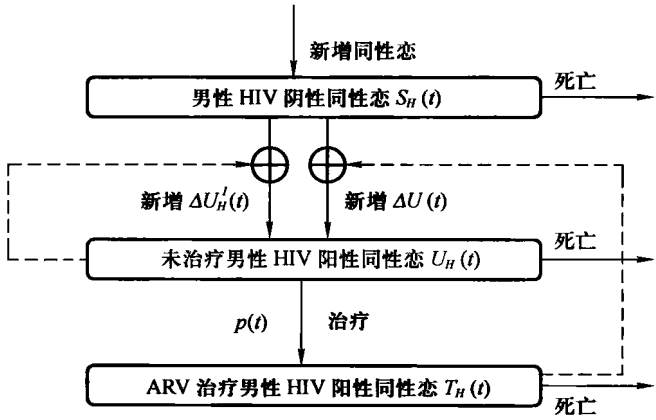


图 6-37　考虑 ARV 治疗的同性恋 HIV 传播过程图

如图 6-37，单位时间内新增的 HIV 阳性同性恋者包括两部分，一部分是易感者与未接受治疗的 HIV 同性恋发生同性性行为而传染的数量 $\Delta U_H^U(t)$：

$$\Delta U_H^U(t)=U_H(t)\rho_H\frac{S_H(t)}{H(t)}(1-C_H)R_{UH}.$$

其中 $U_H(t)$ 是 t 时刻未接受治疗的男性同性恋 HIV 感染者人数；$H(t)$ 是 t 时刻男性同性恋人数；C_H 是男性同性性行为安全套的使用比例；ρ_H 是男性同性恋每年的同性性行为次数；R_{UH}是在一次男同性性行为中，一个未治疗的 HIV 携带者传染给一个易感者的概率.

另一部分是易感者与接受 ARV 治疗的 HIV 同性恋发生同性性行为而传染的数量 $\Delta U_H^T(t)$：

$$\Delta U_H^T(t)=T_H(t)\rho_H\frac{S_H(t)}{H(t)}(1-C_H)R_{TH},$$

其中 $T_H(t)$ 是 t 时刻接受 ARV 治疗的男性同性恋数量，当接受治疗的比例为 $p(t)$时，则 $T_H(t)=\dfrac{p(t)}{1-p(t)}I_H(t)$；$R_{TH}$是在一次男性同性性行为中，一个接受 ARV 治疗的 HIV 携带者传染给一个易感者的概率.

综上所述，可得到在单位时间内新增同性恋 HIV 感染者的数量为

$$\Delta U_H(t)=\Delta U_H^U(t)+\Delta U_H^T(t). \tag{6.21}$$

3. 考虑ARV治疗的交叉使用毒品注射器传播模型

交叉使用毒品注射器的传播情况与同性恋者传播情况类似，通过交叉使用毒品注射器导致新增的HIV感染者包括：与未接受治疗的HIV感染共用注射器以及与已接受ARV治疗的HIV感染者共用注射器两种情况. 则有下面的模型：

$$\begin{cases}\Delta U_D^U(t)=U_D(t)\rho_{UD}\dfrac{S_D(t)}{d(t)}r_nR_{UD},\\ \Delta U_D^T(t)=T_D(t)\rho_{TD}\dfrac{S_D(t)}{d(t)}r_nR_{TD},\\ \Delta U_D(t)=\Delta U_D^U(t)+\Delta U_D^T(t).\end{cases} \tag{6.22}$$

其中$d(t)$是t时刻共用毒品注射器的人数，且$d(t)=S_D(t)+U_D(t)+T_D(t)$；$S_D(t)$是$t$时刻共用毒品注射器，但未感染HIV的人数；$U_D(t)$是$t$时刻共用毒品注射器，感染HIV且未治疗的人数；$T_D(t)$是$t$时刻共用毒品注射器，感染HIV且接受ARV治疗的人数；R_{UD}是易感者与未接受治疗的HIV感染者共用注射器被感染的概率；R_{TD}是易感者与接受ARV治疗的HIV感染者共用注射器被感染的概率，不妨取$R_{TD}=fR_{ID}$；ρ_{UD}是HIV感染未治疗者在单位时间内共用注射器的次数；ρ_{TD}是HIV感染接受ARV治疗者在单位时间内使用注射器的次数，取$\rho_{TD}=j_{arv}\rho_{UD}$；$r_n$是共用针头的比例.

4. 考虑ARV治疗的不同传播途径传播模型

根据前面3个模型中的（6.20）式、（6.21）式和（6.22）式，可以得到在单位时间内新增感染人数（除母婴传播外）的模型为

$$\Delta U_{S2I}(t)=\Delta U_S(t)+\Delta U_D(t)+\Delta U_H(t), \tag{6.23}$$

将其代入（6.9）式中得

$$\begin{cases}\dfrac{\mathrm{d}U(t)}{\mathrm{d}t}=B_U(t)+\Delta U_{S2I}(t)-p-D_U(t),\\ \dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=B_S(t)-\Delta U_{S2I}(t)-D_S(t),\\ \dfrac{\mathrm{d}T(t)}{\mathrm{d}t}=p-D_T(t).\end{cases}$$

其中$B_U(t)$是第t年新生感染HIV婴儿数量；$B_S(t)$是第t年新生HIV阴性婴儿数量；$D_U(t)$是第t年HIV感染未治疗死亡病例数；$D_S(t)$是第t年非AIDS致死人数；$D_T(t)$是第t年HIV感染后接受ARV治疗的死亡病例数.

模型中一些参量的计算方法与模型3中相同，此处不再赘述.

6.7.3 模型5.3：考虑HIV疫苗的多传播途径传播模型

针对实际中存在HIV疫苗的条件下，考虑多途径传播问题. 在模型5.1

假设的基础上，假设接种疫苗能够有效降低易感者感染 HIV 的可能性，能够实现将感染风险降低到接种前的 e 倍；不考虑疫苗的有效期问题，即一旦接种疫苗，则终生有效；政府每年可以资助 $q(t)$ 人接种疫苗.

对于有疫苗的情况，易感人群 S 的变化规律与模型 5.1 有所不同，即在单位时间内有 $q(t)$ 的易感者接种疫苗，从而进入免疫人群 R 中，即

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t}=B_S(t)-\Delta I_{S2I}(t)-D_S-q(t),$$

其中 D_S，$\Delta I_{S2I}(t)$ 的计算方法与模型 5.1 中（6.11）式相同.

由于接种疫苗的易感者与感染者接触后，被传染率降低到接种之前的 e 倍，而不是能够完全免疫，因此在单位时间内还会有 $\Delta R(t)$ 个接种疫苗者感染 HIV，即

$$\Delta R(t)=e\Delta I_{S2I}(t)\frac{R}{S},$$

所以，已接种疫苗的人群（免疫类人群）R 所满足的微分方程模型为

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t}=q(t)-\Delta R(t)-D_R(t),$$

其中 $D_R(t)$ 是在 t 时刻接种疫苗人群在单位时间内的死亡人数.

于是，新增感染人群可能来自于未接种疫苗的易感者，也可能是接种疫苗的易感者，即有

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t}=B_I(t)+\Delta I_{S2I}(t)\left(1+\frac{eR}{S}\right)-D_I,$$

综上所述，在实际中存在 HIV 疫苗的情况下，多种传播途径 HIV 传播模型为

$$\begin{cases}\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=B_I(t)+\Delta I_{S2I}(t)\left(1+\dfrac{eR}{S}\right)-D_I,\\ \dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=B_S(t)-\Delta I_{S2I}(t)-S(t)\Omega_S-q(t),\\ \dfrac{\mathrm{d}R(t)}{\mathrm{d}t}=q(t)-\Delta R(t)-R\Omega_R.\end{cases}$$

其中在单位时间内新生婴儿感染 HIV 数量为 $B_I(t)$，新生婴儿为 HIV 阴性数量为 $B_S(t)$，在单位时间内未接种疫苗人群中感染 HIV 的数量 $\Delta I_{S2I}(t)$ 的计算方法同模型 5.1.

6.7.4 模型 5.4：考虑 ARV 治疗和 HIV 疫苗的多途径传播模型

组合模型 5.2 和模型 5.4，即可得到综合考虑 ARV 治疗和 HIV 疫苗的多途径传播模型. 在这里结合模型 5.2 和模型 5.3 的假设来考虑问题.

（1）易感人群 S 在单位时间内的变化情况

用 B_S 表示新生婴儿HIV为阴性的人群；$q(t)$ 为易感者接种疫苗的人群；ΔU_{S2U}为易感者转化为感染者未治疗的人群；D_S 为易感者中死亡的人群. 则易感人群 S 所满足的微分方程模型为

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t}=B_S(t)-q(t)-\Delta U_{S2U}(t)-D_S(t),$$

其中 $\Delta U_{S2U}(t)$ 的确定方法同模型5.2中的（6.23）式.

（2）已接种疫苗人群 R 在单位时间内的变化情况

用 $q(t)$ 表示易感者接种疫苗后进入 R 人群；D_R 为已接种疫苗者中死亡的人群；ΔU_{R2U}为已接种疫苗者感染HIV的人群，即成为未接受治疗的HIV携带者.

由于已接种疫苗的人群与未接种疫苗人群的传染率之比为 e，而两类人群面对的传染源是一致的. 因此在单位时间内，R 人群感染HIV的数量为

$$\Delta U_{R2U}(t)=e\Delta U_{S2U}(t)\frac{R}{S},$$

则已接种疫苗人群 R 所满足的微分方程模型为

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t}=q(t)-\Delta U_{R2U}(t)-D_R(t).$$

（3）已感染HIV但未治疗的人群 U 的变化情况

用 B_I 表示新生婴儿HIV为阳性，即新加入感染未治疗的人群；D_U 为感染未治疗者中死亡的人群；$\Delta U_{S2U}+\Delta U_{R2U}$为HIV阴性人群通过交叉感染，变成为未接受治疗的HIV携带者；$P(t)$ 为HIV感染未治疗者开始接受治疗的人群. 则感染HIV未治疗人群 U 所满足的微分方程模型为

$$\frac{\mathrm{d}U(t)}{\mathrm{d}t}=B_I(t)+\Delta U_{S2U}(t)\left(1+\frac{eR}{S}\right)-P(t)-D_U.$$

（4）感染HIV参加治疗的人群 T 的变化情况

用 D_T 表示接受治疗的感染者中死亡的人群；$p(t)$ 为HIV感染未治疗者开始接受治疗的人群. 则感染HIV参加治疗的人群 T 所满足的微分方程模型为

$$\frac{\mathrm{d}T(t)}{\mathrm{d}t}=p(t)-D_T(t).$$

综上所述，则可得到同时考虑ARV治疗和HIV疫苗的传播微分方程模型：

$$\begin{cases}\dfrac{\mathrm{d}U(t)}{\mathrm{d}t}=B_I(t)+\Delta U_{S2U}(t)\left(1+\dfrac{eR}{S}\right)-P(t)-D_U,\\[2ex]\dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=B_S(t)-q(t)-\Delta U_{S2U}(t)-D_S(t),\\[2ex]\dfrac{\mathrm{d}R(t)}{\mathrm{d}t}=q(t)-\Delta U_{S2U}(t)\dfrac{eR}{S}-D_R(t),\\[2ex]\dfrac{\mathrm{d}T(t)}{\mathrm{d}t}=p(t)-D_T(t).\end{cases}$$

其中 $\Delta U_{S2U}(t)$ 的确定方法同模型 5.2.

6.7.5　模型 5.5：考虑 ARV 治疗抗药性的 HIV 多途径传播模型

设产生抗药性的人群为 W，单位时间内，每位接受 ARV 治疗的 HIV 感染者产生抗药性的概率为 q_W.

(1) 由异性性行为的传播模型

在实际中，一旦出现了抗药性人群，则易感人群中被感染来源多出于抗药性人群 W. 如果在单位时间内女性易感人群 S 与男性抗药性人群 W 通过性接触，则产生的新增感染女性人数为

$$\Delta U_S^{WF}(t)=W_S^M(t)\rho_W\frac{S_S^F(t)}{F(t)}(1-C)R_{WF},$$

其中 ρ_W 表示产生抗药性人群每年的性行为平均次数，不妨设 $\rho_W=j_W\rho_U$；$F(t)$ 是成年女性人数，R_{WF}是产生抗药性的男性与易感女性发生一次性关系后易感女性感染 HIV 的概率.

在单位时间内，通过异性性行为传播导致的新增女性 HIV 感染未治疗人群总数为

$$\Delta U_S^F(t)=\Delta U_S^{TF}(t)+\Delta U_S^{UF}(t)+\Delta U_S^{WF}(t),$$

其中 $\Delta U_S^{TF}(t)$，$\Delta U_S^{IF}(t)$ 分别由 (6.17) 式和 (6.18) 式确定.

同理，在单位时间内通过异性性行为传播导致的新增男性 HIV 感染未治疗人群总数为

$$\Delta U_S^M(t)=\Delta U_S^{TM}(t)+\Delta U_S^{UM}(t)+\Delta U_S^{WF}(t).$$

因此，在单位时间内通过异性性行为导致的新增 HIV 感染总人数为

$$\Delta U_S(t)=\Delta U_S^M(t)+\Delta U_S^F(t),\tag{6.24}$$

(2) 由同性性行为的传播模型

考虑在出现 ARV 抗药性后，在单位时间内新增的 HIV 阳性同性恋者包括：易感者与未接受治疗的 HIV 感染者发生同性性行为而感染的病例；易感者与接受治疗的 HIV 感染者发生同性性行为而感染的病例；易感者与接受治疗后产生抗药性的 HIV 感染者发生同性性行为而感染的病例.

模型 5.2 讨论了前两种新增病例数量的变化模型. 下面讨论易感者与接受治疗后产生抗药性的 HIV 感染者发生同性性行为而感染的病例数量，即

$$\Delta U_H^W(t)=U_H(t)\rho_H\frac{S_H(t)}{H(t)}(1-C_H)R_{WH},$$

其中 $U_H(t)$ 是 t 时刻未接受治疗的男性同性恋 HIV 感染者人数；$H(t)$ 是 t 时刻男性同性恋的人数；C_H 是男性同性性行为安全套的使用比例；ρ_H 是男性同性恋每年的同性性行为次数；R_{WH}是在一次男性同性性行为中一个产生抗药性的 HIV 携带者传染给另一个易感者的概率. 则在单位时间新增同性恋 HIV 感

染者的数量为

$$\Delta U_H(t)=\Delta U_H^U(t)+\Delta U_H^T(t)+\Delta U_H^W(t). \tag{6.25}$$

其中 $\Delta U_H^U(t)$，$\Delta U_H^T(t)$ 分别由模型5.2中有关同性性行为传播模型确定.

（3）由交叉使用毒品注射器的传播模型

交叉使用毒品注射器传播模型与同性恋传播模型类似，通过交叉使用毒品注射器导致新增的HIV感染者情况包括：与未接受治疗的HIV感染者共用注射器；与接受ARV治疗的HIV感染者共用注射器；与产生抗药性的HIV感染者共用注射器.

前两种情况模型5.2已做了讨论，这里只讨论易感者与产生抗药性的HIV感染者共用注射器导致的新增HIV感染病例的变化，即有

$$\Delta U_D^W(t)=U_D(t)\rho_{WD}\frac{S_D(t)}{d(t)}r_nR_{WD}$$

其中 $d(t)$ 表示 t 时刻共用毒品注射器的人数，且 $d(t)=S_D(t)+U_D(t)+T_D(t)$；$S_D(t)$ 是 t 时刻共用毒品注射器后未感染HIV的人数；$U_D(t)$ 是 t 时刻共用毒品注射器后感染了HIV，且未治疗的人数；R_{WD}是易感者与产生抗药性的HIV感染者共用注射器后被感染的概率；ρ_{WD}是产生抗药性的HIV感染者单位时间内注射的次数；r_n 是共用针头的比例.

在单位时间内，通过共用不洁注射器导致新增HIV感染者的数量为

$$\Delta U_D(t)=\Delta U_D^U(t)+\Delta U_D^T(t)+\Delta U_D^W(t) \tag{6.26}$$

其中 $\Delta U_D^U(t)$，$\Delta U_D^T(t)$ 的确定方法与模型5.2相同.

（4）考虑ARV治疗的不同途径传播模型

综合考虑上面的3个模型，由（6.24）式、（6.25）式和（6.26）式，则可得到单位时间内新增感染人数（除母婴传播外）为

$$\Delta U_{S2U}(t)=\Delta U_S(t)+\Delta U_D(t)+\Delta U_H(t). \tag{6.27}$$

将其代入（2.9）式，得

$$\begin{cases}\dfrac{\mathrm{d}U(t)}{\mathrm{d}t}=B_U(t)+\Delta U_{S2U}(t)-p-D_U(t),\\[2ex]\dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=B_S(t)-\Delta U_{S2U}(t)-D_S(t),\\[2ex]\dfrac{\mathrm{d}T(t)}{\mathrm{d}t}=p-D_T(t).\end{cases}$$

其中 $B_U(t)$ 是第 t 年新生婴儿感染HIV的数量；$B_S(t)$ 是第 t 年新生婴儿HIV为阴性数量；$D_U(t)$ 是第 t 年感染HIV未治疗死亡病例数；$D_S(t)$ 是第 t 年非AIDS致死人数；$D_T(t)$ 是第 t 年感染HIV后接受ARV治疗的死亡病例数.

6.7.6 模型5.6：考虑ARV治疗与抗药性和HIV疫苗的传播模型

组合考虑模型5.4和模型5.5，即可得到综合考虑ARV治疗与抗药性和存

在 HIV 疫苗的传播模型.

(1) 易感人群 S 的变化模型

易感人群 S 所满足的微分方程模型为

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t}=B_S(t)-q(t)-\Delta U_{S2U}(t)-D_S(t),$$

其中 $\Delta U_{S2U}(t)$ 的确定方法由模型 5.5 中的 (6.27) 式确定.

(2) 已接种疫苗人群 R 的变化模型

已接种疫苗的易感人群 R 的变化情况可以由模型 5.4 确定. 即在单位时间内，R 人群感染 HIV 的数量为

$$\Delta U_{R2U}(t)=e\Delta U_{S2U}(t)\frac{R}{S},$$

则已接种疫苗的易感人群 R 所满足的微分方程模型为

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t}=q(t)-\Delta U_{R2U}(t)-D_R(t),$$

(3) 感染 HIV 未治疗人群 U 的变化模型

在单位时间内，感染了 HIV 但未治疗的人群为 U；B_I 为新生婴儿 HIV 阳性，即新加入感染未治疗人群；D_U 为感染 HIV 未治疗死亡的人群；$\Delta U_{S2U}+\Delta U_{R2U}$ 为 HIV 阴性人员通过交叉感染，成为未接受治疗的 HIV 携带者；$P(t)$ 为感染 HIV 未治疗者开始接受治疗的人群.

于是有感染 HIV 未治疗人群 U 所满足的微分方程为

$$\frac{\mathrm{d}U(t)}{\mathrm{d}t}=B_I(t)+\Delta U_{S2U}(t)\left(1+\frac{eR}{S}\right)-P(t)-D_U.$$

(4) 感染 HIV 接受治疗人群 T 的变化模型

在单位时间内，已知感染 HIV 接受治疗的人群 T，D_T 表示接受治疗的感染者死亡数；$P(t)$ 为感染 HIV 未治疗者开始接受治疗的数量. 则有感染 HIV 接受治疗人群 T 所满足的微分方程模型为

$$\frac{\mathrm{d}T(t)}{\mathrm{d}t}=p(t)-D_T(t).$$

(5) 综合考虑 ARV 治疗和 HIV 疫苗的模型

综上所述，则可得综合考虑 ARV 治疗和 HIV 疫苗的微分方程模型：

$$\begin{cases}\dfrac{\mathrm{d}U(t)}{\mathrm{d}t}=B_I(t)+\Delta U_{S2U}(t)\left(1+\dfrac{eR}{S}\right)-P(t)-D_U,\\[2ex]\dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=B_S(t)-q(t)-\Delta U_{S2U}(t)-D_S(t),\\[2ex]\dfrac{\mathrm{d}R(t)}{\mathrm{d}t}=q(t)-\Delta U_{S2U}(t)\dfrac{eR}{S}-D_R(t),\\[2ex]\dfrac{\mathrm{d}T(t)}{\mathrm{d}t}=p(t)-D_T(t).\end{cases}$$

其中 $\Delta U_{S2U}(t)$ 的确定方法由模型5.5中的（6.27）式确定.

6.7.7 模型的分析与评价

模型5深入细致地分析了从HIV易感者转化为HIV传染者的四种传播途径，并针对每一种传播途径的特点给出了传播模型，比较准确地解释了HIV的传播规律. 该模型较前面的模型更为实际，能够更有针对性地解释不同国家的HIV传播特点和传播规律.

在这里还要注意一个基本事实，HIV的传播与一般传染病（例如流感、天花等）的传播特性具有明显的不同，因此在应用中，不能直接照搬一般传染病的数学模型，否则结果可能会出现较大的偏差.

我们知道，流感的传播途径主要是公共场所，一个流感患者每天传播流感的人数只与公共场所中易感者的比例有关，公共场所易感者的比例越高，可能传染的人就越多. 而且每天所接触的易感者是不确定的，可以假设不同时间内的感染是独立的.

而HIV的传播途径不同，一个HIV携带者在一段时间内性伴侣是相对稳定的，而且数量不会太大. 据有关资料统计，2001年瑞典人在一年中不同性伴侣的个数主要集中在3个以内，大多数性行为的对象都比较固定.

在研究流感的传播规律时，可以假设一个时间周期的新增感染人数只和当前的状态（感染人群总数和易感人群总数）有关，而与之前的感染过程无关，即流感的传播具备马尔可夫特性（即无后效性），因此可以直接建立简单的一阶马尔可夫状态转移方程来研究. 例如A是感冒病毒携带者，A在一天中接触了 n 个人，B是其中一个人，由于感冒病毒往往在公共场所传播，因此B就是一个采样样本. 此时B是健康者的概率与所有健康者的比例是一致的，即等于传染人群总数与总人口的比值. 在流行感冒传播模型中，A与B的接触不论是第一次接触，还是已经多次接触，发生新增传染源的概率都是相同的.

而在研究HIV的传播规律就不具备这一特性，A与B第一次发生性关系和多次发生性关系导致的新增传染源的概率是不同的. 例如，A是一个HIV感染者，A在一次性行为中只和B接触，且是第一次与B发生性关系，则可以假设B是易感者的概率与当时全民中HIV易感人群的比例相同，即 $P=\frac{S(t)}{N(t)}$. 即在这次性行为中，B被感染HIV的概率为 $P_1=\lambda\frac{S(t)}{N(t)}$，这一点与流感传染模型是一致的. 当A与B第二次发生性关系时（假设期间B没有感染HIV），B可能在第一次性行为中已经被感染HIV，此时B是HIV感染者的比例要比第一次的大得多，B在第二次性行为中是易感者的概率为 $P=(1-\lambda)$，在此次性行为中，B成为新增感染者的概率为 $P_2=(1-\lambda)\lambda$. 以此类推，B第 k

次与 A 发生性关系，成为新增 HIV 感染者的概率为 $P_k=(1-\lambda)^{k-1}\lambda$.

这就说明了 HIV 传播规律不具备马尔可夫特性，因此在 HIV 的异性传播途径中，往往会出现局部高危传染，而流感则会发生大范围传染.

为了解释 HIV 的上述特殊传播特性，若采用多阶马尔可夫过程的方法建立模型会使模型异常复杂. 为了降低模型复杂度，通常采用通过蒙特卡洛仿真的方法仿真复杂的传播过程. 蒙特卡洛仿真方法的基本思想是通过随机数的方法仿真每一次传播事件，最后统计传播规律. 基于蒙特卡洛仿真思想的算法从表述上来分包括：复杂动态网络模型、元胞自动机模型、基于 Agent 的传播仿真模型等.

6.8 模型 6：HIV 复杂动态网络传播模型

模型 6 是 HIV 仿真模型. 该模型将传播过程离散化，通过蒙特卡洛方法产生各种随机感染事件，可以更精确地仿真各种传播途径的传播特性.

6.8.1 模型 6.1：HIV 自由传播模型

近 10 年来，蓬勃发展的复杂网络理论研究已经成为复杂系统科学活动中最活跃的分支之一. 计算机病毒在互联网上的传播行为是复杂网络研究的主要内容之一，更是近年来人们在社会经济活动中最为关切的一个主题.

复杂网络是指具有复杂拓扑结构和动力学行为的大规模网络，它是由大量节点通过边的相互连接而构成的图. 例如 Internet、生物网络、无线通信网络、电力网络、流行病网络和谣言传播网络等都是复杂网络. 在复杂网络仿真模型中，每个个体的行为可以自由设定，可以处理各种复杂的因果关系，能够实现个体微观行为的仿真.

为了更好地对 HIV/AIDS 的传播过程进行长期预测分析，我们建立更符合实际的复杂动态网络模型来研究这一问题.

1. 问题的分析与假设

在上面的连续微分方程模型中，只能实现将人群进行粗粒度的分类，并且假设同一类人群的所有传染特征都相同，包括诸如传染率、死亡率、出生率等特征. 而在复杂动态网络模型中，可以自由设置每一个个体的统计学行为，可以实现每个个体的传染特征呈现特定的概率分布. 利用复杂网络模型可以实现更贴近实际的仿真.

但是，由于复杂动态网络仿真模型是针对每一个人的具体情况进行的，因此该模型可以更简便地处理各种复杂的人口统计数据. 不足之处是当一个国家的人口总数很大时，仿真的复杂度将达到一个无法接受的程度. 为了方便建模分析，给出如下的假设：

（1）为了降低计算复杂度，用1 000人组成的人群代替一个国家的人群，仿真HIV的扩散规律.

（2）HIV的传播途径包括异性性传播、男同性性传播、交叉使用毒品注射器传播和母婴传播.

（3）将人的HIV传播特征分为2个阶段，第一个阶段是未成年阶段，指从0岁到16岁；第二阶段是成人阶段，指16岁之后. 假设每个阶段内部的传播特征稳定.

（4）只有成年人（16岁以上）有性行为和共用毒品注射器的可能.

（5）未接受治疗的HIV感染者从感染到死亡的时间为10年.

2. 自由传播模型的建立

HIV自由传播仿真模型的基本思路：首先生成1 000个个体，并随机产生HIV携带者标记. 之后建立自由传播关系网络，再通过关系网络研究传播过程. 仿真模型流程图如图6-38所示.

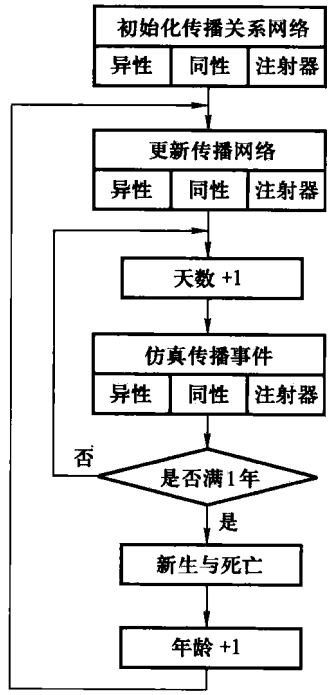


图6-38 仿真模型流程图

（1）初始潜在传播关系网的建立

仿真模型中，各种传播关系是通过“潜在传播人群”产生的. 一个人的“潜在传播人群”是指可能与这个人产生某种传播关系的人的集合. 由于传播关系网主要用于研究交叉传播，因此在这里传播关系网主要用于研究性传播途

径和共用不洁毒品注射器传播. 例如 i 的性传播关系的潜在传播人群是指可能与 i 发生性关系的人的集合. 为了说明方便，定义 l 为传播方式，$l=1$ 表示性传播，$l=2$ 表示共用毒品注射器传播.

仿真过程包括两重循环，内重循环用于模拟每天的各种交叉传染事件，其中的参数均以天为单位. 外重循环用以模拟出生、死亡和母婴传播. 在外重循环中，由于存在人口的出生和死亡，从而有了初始潜在传播网络，因此需要在每轮的外重循环中更新潜在传播网络. 下面具体介绍各个环节的仿真实现方法.

首先建立潜在异性传播关系网络. 这里以瑞典为例，图 6-39 是根据有关资料给出的最近 12 个月中不同性伴侣个数的分布.

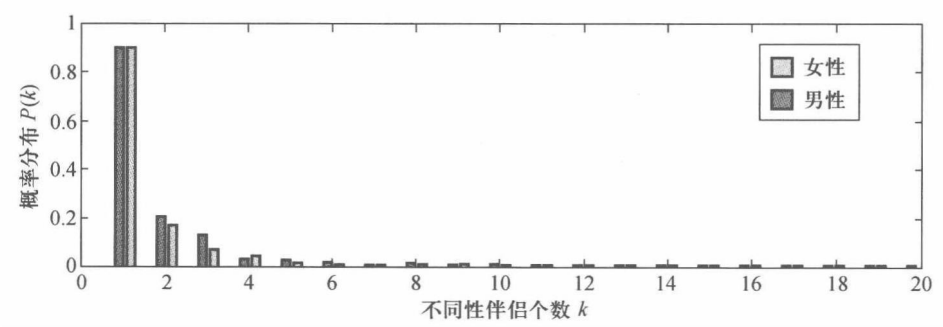


图 6-39　瑞典成年人 12 月不同性伴侣个数分布

设男性性伴侣个数为 k 的概率为 P_k^M，女性性伴侣个数为 k 的概率为 P_k^F，则异性传播关系网的生成算法如下：

Step1：随机生成 1 000 个人的年龄、性别、是否感染 HIV.

Step2：对于 16 岁以上的男性个体通过 P_k^M 产生随机数，确定该成年男性个体性伴侣的个数.

Step3：对于 16 岁以上的女性个体通过 P_k^F 产生随机数，确定该成年女性个体性伴侣的个数.

Step4：记迭代周期数 $t=0$，初始化为 M_i 分配的女性性伴侣个数为 $u_i(t)=0$，剩余未分配性伴侣个数记为 $v_i(t)=k_i$，其中 k_i 为男性 M_i 的性伴侣数量.

Step5：若所有男性成年人当前未分配性伴侣数 $v_i(t)=0$，则停止，此时已生成完整的潜在性关系网络. 否则，随机选择一名男性成年人 M_i，若当前未分配性伴侣数 $v_i(t)>0$，则从所有未分配性伴侣数大于 0 的女性成年人中随机地选择一位 F_j，即建立起 F_j 和 M_i 的潜在性关系.

Step6：重复 Step5，直至生成潜在异性性关系网络.

通过上述步骤即可生成社会中的潜在性关系网络，如图 6-40 所示.

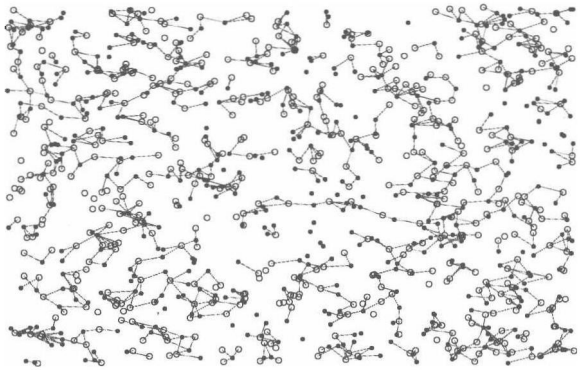


图6-40 潜在异性性关系网络图

图6-40中的节点表示一个个体，空心圈表示男性，实心点表示女性，男性与女性之间的连线表示存在潜在性关系. 从图中可以看出大多数节点的度为1，表示大多数人的性伴侣都固定，部分节点度为0表示该节点未成年，尚无性伴侣.

类似地，可以生成男同性恋潜在传播关系网及交叉使用毒品注射器潜在传播关系网.

(2) 考虑出生和死亡时传播关系网络的更新

该模型需要仿真未来45年的实际传播情况，关系网可能会随着出生和死亡的出现而改变，因此需要增加如下传播关系网的更新操作.

Step1：根据个体是否是HIV感染者，记录其寿命分布和年龄，判断他/她在该周期内是否会死亡.

若个体为HIV阴性，则他/她是否会死亡根据其年龄和该区域的寿命分布确定；若为HIV阳性，是否会死亡则根据其感染时间确定. 若感染时间已达到10年，则判定为死亡，否则按照HIV阴性者判断是否会死亡.

若一个个体死亡则清除所有与该个体有关的潜在传播关系，此时可能导致部分个体性伴侣个数不饱和，即出现 $v_i(t)>0$ 的情况.

Step2：对新进入的16岁个体随机生成性伴侣数量，并按照初始传播关系网的建立方法为16岁个体指定传播关系.

Step3：根据个体年龄、性别和生育率判断是否在该周期内有生育. 若有生育，根据母亲是否HIV感染者和母婴传播风险判断新生婴儿是否感染HIV.

Step4：所有个体年龄加1，回到Step1.

(3) 仿真传播感染事件过程

上面给出了HIV携带者潜在的传播网络，下面给出随机出现的交叉传播事件传播过程的算法.

Step1：寻求传播关系. 在实际中，并不是在每个时间周期里，个体 i 都会

主动寻求发生某种传播关系. 假设在一个周期里，个体 i 主动寻求第 l 种传播关系的概率为 $p_{i,l}$，通过轮盘赌的方法确定 i 是否有意愿发生第 l 种传播关系. 若 i 有意愿，则在 $A_{i,l}$ 中随机地选出一个人 j.

假设不考虑在一个时间周期内的连环传染事件，即不会发生 A 传染给 B，B 再传染给 C 的事件. 在这个时间周期内，只有与 HIV 携带者有直接传播关系的个体（即与 HIV 携带者有边连接的节点）才可能被传染. 但在这个时间周期内，与 HIV 感染者非直接连接的个体即使发生传播接触关系也不会传染 HIV. 为了降低仿真复杂度，不考虑这一部分传播接触. 算法中只有 HIV 携带者、与 HIV 携带者直接接触者才会发出传播关系的请求.

Step2：应答. 为了表示个体 j 是否愿意接受传播关系，建立应答机制. 个体 i 询问个体 j 是否愿意和其建立第 l 种传播关系，设其接受的概率为 $q_{i,j,l}$. 通过轮盘赌的方法随机确定是否接受这种关系，若个体 j 同意，则个体 i 和个体 j 就建立联系.

Step3：中断. 为了表示个体 i 和个体 j 是否愿意在下一个时间周期内继续保持这种传播关系，设立中断机制. 设个体 i 和个体 j 在一个时间周期里中断传播关系 l 的概率为 $u_{i,j,l}$，通过轮盘赌的方法确定已存在的关系在这个时间周期里是否继续维持.

Step4：判断感染事件. 若通过前面的步骤在个体 i 和个体 j 之间建立起了实际的传播关系 l，通过 $p_{i,j,l}$ 判断是否实际发生了感染事件.

类似地，按照上面的仿真步骤，仿真其他人群中各种传播感染事件的过程.

3. 仿真模型的参数确定

若母亲是 HIV 感染者，HIV 病毒通过生育传播给她孩子的概率为 15% ~ 30%. 此外，婴儿期 2 年内的母乳喂养又提高了 10% ~15% 的传播风险[10].

各个区域的人口统计数据主要来自 UNAIDS 的统计资料[11].

假设世界上各个区域的毒品静脉注射人群的比例相同，取自联合国毒品和犯罪办公室的统计数据. 迭代过程中的参数 $p_{i,3}$，$q_{i,j,3}$ 和 $u_{i,j,3}$ 的数据是在经过广泛调研的基础之上进行计算确定的[12].

有关性传播途径的相关参数来自相关的文献资料[13,14].

非 AIDS 致死的死亡率来自 WHO 对各国寿命分布的统计，AIDS 致死依据从感染到死亡的时间为 10 年设置.

6.8.2 模型 6.2：考虑 ARV 治疗的 HIV 传播仿真模型

1. 问题的分析与假设

模型 6.1 中仅给出了自由传播的仿真过程，没有考虑 ARV 治疗的情况，当实际中存在 ARV 治疗时，将会发生一些新的随机事件. 如降低 ARV 治疗者主动发起接触请求的概率；降低 ARV 治疗者母婴传播概率；当易感者向 ARV

治疗者发出接触请求时，降低应答概率；当易感者与ARV治疗者发生接触后，降低感染概率；假设一个感染者接受ARV治疗后，从感染到死亡的时间可以延长到12年.

2. 仿真模型的建立

（1）初始传播关系网的建立

考虑ARV治疗后，初始网络中每个个体的参数除了年龄、性别、是否感染HIV和潜在传播范围以外，还需要根据当时政府提供的ARV治疗经费计算出接受ARV治疗的人员数量，通过随机配置的方法为每个个体标记是否接受ARV治疗.

（2）关于死亡的判断

当存在ARV治疗后，人群的死亡依特征可分为三类：易感人群、感染未治疗人群和感染接受治疗人群. 前两类人群的死亡判断方法与模型6.1相同.

当个体i是接受ARV治疗的HIV感染个体时，根据个体i的年龄和寿命分布判断i是否会在该时间周期内发生非AIDS致死，如果判断个体i死亡，则结束；否则，根据个体i感染HIV至今的时间判断是否发生AIDS致其死亡，如果感染时间至今已达到12年，则判定为死亡，否则为存活.

（3）关于出生的判断

根据个体i的性别、年龄和生育率分布判断在该时间周期内是否会生育，若判断为是，则产生新个体j. 新生个体j是否感染HIV的判断方法如下：

- 若个体j的母亲i是HIV阴性，则j也是HIV阴性；
- 若个体j的母亲i是HIV阳性且未治疗，则根据未治疗情况下母婴传播概率判断j是否会感染HIV；
- 若个体j的母亲i是HIV阳性且接受ARV治疗，则根据ARV治疗情况下母婴传播概率判断j是否会感染HIV.

（4）传播关系网的更新与随机感染事件

ARV治疗只是会降低接触概率，从而降低个体之间的接触数量，而不影响关系网络，所以传播网络的更新方法与模型6.1相同.

当实际中存在ARV治疗时，接受ARV治疗的感染者将降低发起接触请求的概率，同时也降低应答其他接触请求的概率，还可能加大中断现有请求的概率.

若在某次计算中，仿真得到易感者j与ARV治疗者i接触，此时j被感染的概率较之与未治疗的感染者k接触的概率要小.

6.8.3 模型6.3：考虑HIV疫苗的HIV传播仿真模型

1. 问题的分析与假设

当实际中存在HIV疫苗时，则在每个时间周期内会有固定数量$E(t)$的易

感者开始接种疫苗；接种疫苗的易感者与 HIV 感染者接触后，接触感染的概率为 p^{v}_{ijl}；疫苗的有效期为 10 年，超过 10 年后重新成为普通易感者.

2. 仿真模型的建立

该模型的仿真过程与模型 6.1 相似，不同之处在于以下几个方面：

（1）为每个个体增加是否具有 HIV 抗体的标记，是标记为“1”，否标记为“0”.

（2）在初始时刻所有易感者都不具备 HIV 抗体，即对所有个体该字段都标记为“0”.

（3）为每个个体增加接种疫苗时间的字段 t，将所有个体疫苗接种时间字段初始化为“NULL”.

（4）在每个时间周期内随机选择 $E(t)$ 名易感者接种疫苗，即将接种疫苗的个体 HIV 抗体字段的 0 改为 1，并将疫苗接种时间修改为当前时间. 在不同的疫苗普及策略下，$E(t)$ 的取值有所不同. 在仿真计算中，通过修改不同的疫苗使用策略 $E(t)$，以实现对不同资金投入下的 HIV 防控效果进行评价.

（5）当具有 HIV 抗体的易感者与 HIV 感染者建立接触关系后，判断是否会感染 HIV 的概率为 p^{v}_{ijl}，一般认为接种疫苗后可以将 HIV 感染概率降低到接种前的 5%，即 $p^{v}_{ijl}=0.05p_{ijl}$.

（6）在每个时间周期开始时刻，根据当前时间和每位具有 HIV 抗体的个体接种疫苗时间，确定疫苗是否失效. 若从接种时间到当前时间超过 10 年，则让该个体的 HIV 抗体消失，即将抗体字段的“1”改为“0”，并且将接种时间字段重新修改为“NULL”.

6.8.4 模型 6.4：考虑 ARV 治疗和 HIV 疫苗的 HIV 传播仿真模型

因为模型 6.2 和模型 6.3 之间不存在冲突关系，因此只需要将模型 6.2 和模型 6.3 的仿真操作过程结合起来，即可得到综合考虑 ARV 治疗和 HIV 疫苗的 HIV 传播模型.

6.8.5 模型 6.5：考虑 ARV 抗药性的 HIV 传播仿真模型

1. 问题的分析与假设

模型 6.2 不考虑 ARV 抗药性的情况，但当实际中存在 ARV 抗药性时，将会发生如下的几种变化：

（1）在每个时间周期内，有一定比例的接受 ARV 治疗的感染者会产生抗药性，一般认为占每年治疗者的 5% 左右.

（2）产生抗药性的 HIV 感染者与易感者接触后，接触感染的概率为 p^{w}_{ijl}.

（3）产生抗药性的个体永远无法再次接受 ARV 治疗.

（4）假设产生抗药性后从 HIV 感染到死亡的时间为 9 年.

2. 仿真模型的建立

仿真过程与模型6.2相似，不同之处在于如下几个方面：

（1）为每个个体增加是否会产生抗药性标记字段，产生抗药性标记为“1”，否则标记为“0”. 初始化时所有的个体该字段都标记为“0”.

（2）每个周期中分配新增接受ARV治疗的个体时，只从未产生抗药性的人群中选择.

（3）在每个时间周期内按照5%的比例，随机从所有接受ARV治疗的人群中选择产生抗药性的个体，将这些个体“是否接受ARV治疗”字段修改为“0”，将“是否产生抗药性字段”修改为“1”，表示此后该个体不再接受治疗，并且无法再次接受治疗.

（4）产生抗药性的HIV感染者与易感者接触后，接触并感染的概率为 p_{ijl}^{w}.

（5）产生抗药性个体的死亡判定方法：若从感染时间到当前时间等于或者超过9年，则判断为死亡，否则按照正常死亡率判定是否死亡.

6.8.6 模型6.6：考虑ARV抗药性和HIV疫苗的HIV传播仿真模型

由于模型6.2、模型6.3和模型6.5之间都不存在冲突关系，组合3个模型的相关仿真操作过程，即可实现同时考虑ARV治疗、治疗抗药性和HIV疫苗的HIV传播模型的仿真过程.

6.8.7 模型的分析与评价

复杂动态网络仿真模型能够模拟每一个个体的传播行为，可以实现复杂行为和具有统计学传播特征的仿真过程，尤其是对于HIV传播的多阶马尔可夫特性的处理.

在此模型的基础上，还可以很方便地加入更多的因素，以更真实地研究实际的传播规律. 例如不同传染阶段传染性的不同、年龄与性伴侣数量的关系、吸毒人员的年龄分布、吸毒人员的HIV感染比例、同性恋的年龄分布、无保护性行为的概率、ARV治疗效果、疫苗效果、性工作者数量等情况的影响.

在使用复杂动态网络模型进行仿真时需要遵循如下步骤：

（1）统计确定参数. 通过历史数据统计所需要考虑因素的概率分布以及各因素之间的条件概率分布等.

（2）生成随机样本. 随机样本包括随机特征和随机网络，随机特征是指某次仿真中，各个个体的样本观测值，例如年龄、性别、是否感染HIV、是否接受ARV治疗、在一次危险接触中是否感染等. 随机网络是指各类关系的一次样本观测值，例如性关系随机网络、吸毒人员共用注射网络等. 在生成随机特征和随机网络时，需要按照步骤（1）中的概率和条件概率随机生成.

(3) 根据各种逻辑关系，对下一阶段的样本特征和网络状况进行更新.

基于复杂网络的多种途径 HIV 传播的仿真模型可以处理各种复杂问题，与实际情况较为接近. 但仿真方法的缺点是计算复杂度高，计算精度取决于仿真次数. 因此一般来说仿真的方法往往适用于对结果精度要求较高，能够承担起相应计算量的场合.

6.9 HIV 传播模型仿真求解

在前面几节中，我们从不同角度建立了 HIV 的传播模型，模型 1 至模型 5 都是基于群体演化的微分方程模型，模型的变量是各个群体的人数. 对于微分方程模型的求解可以使用多种不同的方法，其基本原理都是通过差分迭代的方法对微分方程组进行数值求解. 下面以模型 3 为例说明如何使用 Simulink 进行仿真求解微分方程模型.

6.9.1 自由传播模型的 Simulink 仿真求解

1. Logistic 人口模型仿真

由 Logistic 人口模型：

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t}=\gamma_0\left(1-\frac{N(t)}{K}\right)N(t),$$

该微分方程模型的 Simulink 模型如图 6-41 所示.

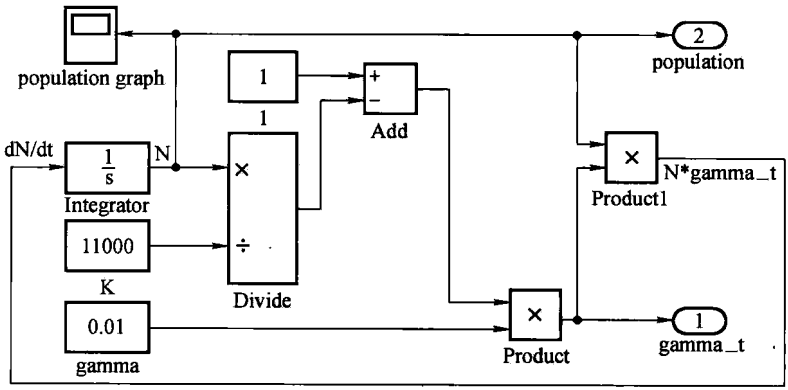
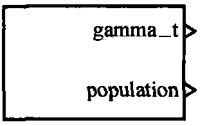


图 6-41　Logistic 人口模型仿真模块

该模型中设定的人口上限值为 $K=11\ 000$，原始出生率 $\gamma_0=0.01$. 在仿真过程中可以根据需要修改，即不断地产生各个时刻的人口总数 $N(t)$ 和人口净增长率 $\gamma(t)$.

该模型有 2 个输出端口，一个是时刻 t 的人口净增长率 gamma_t，另一个是时刻 t 的人口总数 population. 将 Logistic 人口模块封装为子模块，得到人口

模型子模块结构如图 6-42 所示. 该模块为信号源类型模块，不断地生成各个时刻的人口总量数据 population 和人口净增长率数据 gamma_t.



Logistic population module

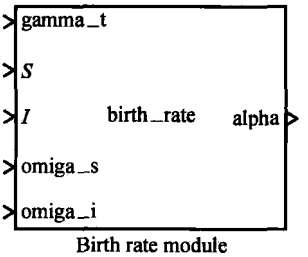
图 6-42 Logistic 人口仿真模块封装端口图

2. 出生率的仿真计算

根据当前的易感人群总数 $S(t)$、感染人群总数 $I(t)$ 和人口净增长率 $\gamma(t)$ 计算出生率 $\alpha(t)$. 时刻 t 的出生率模型为

$$\Omega(t)=\frac{I(t)\Omega_I+S(t)\Omega_S}{N(t)},\quad \alpha(t)=r(t)+\Omega(t).$$

该模型可以使用 Simulink 的 Embedded MATLAB function 模块自定义函数，根据上式编制出生率模块函数，该模块端口结构如图 6-43 所示.



Birth rate module

图 6-43 出生率仿真计算端口图

该模型包括 5 个输入，分别是当前易感人群总数 S，感染人群总数 I，当前人口净增长率gamma_t，易感人群死亡率 omiga_s，感染人群死亡率 omiga_i. 该模型的输出为当前的人口出生率 alpha.

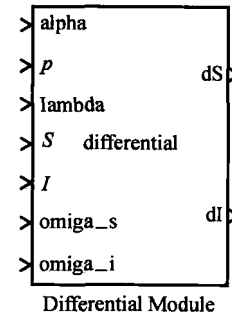
3. 微分方程组的仿真计算

根据出生率 $\alpha(t)$、易感人群总数 $S(t)$、感染人群总数 $I(t)$ 仿真计算易感人群和感染人群的增长速率 $\frac{\mathrm{d}S(t)}{\mathrm{d}t}$ 和 $\frac{\mathrm{d}I(t)}{\mathrm{d}t}$. 即

$$\begin{cases}\dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=B_I(t)+\lambda I(t)S(t)-D_I(t),\\[2ex]\dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=B_S(t)-\lambda I(t)S(t)-D_S(t).\end{cases}$$

其中 $B_S(t)=(1-p_I)\alpha(t)I(t)+\alpha(t)S(t)$，$B_I(t)=p_I\alpha(t)I(t)$，$D_S(t)=S(t)\Omega_S$，$D_I(t)=I(t)\Omega_I$.

使用 Embedded MATLAB function 编制该模块函数，模块的输入输出端口如图 6-44 所示.



Differential Module

图 6-44 微分方程组仿真计算端口图

该模块的输入包括出生率 alpha，易感人群总数 S，感染人群总数 I，母婴传播风险 p，传染率lambda，易感人群死亡率 omiga_s，感染人群死亡率omiga_i. 输出为易感人群的增长率 dS，感染人群的增长率 dI.

4. HIV 自由传播的 Simulink 模型

在上述三个仿真模块的基础上，根据模型 3，则可得到 HIV 自由传播的 Simulink 仿真模型，如图 6-45 所示.

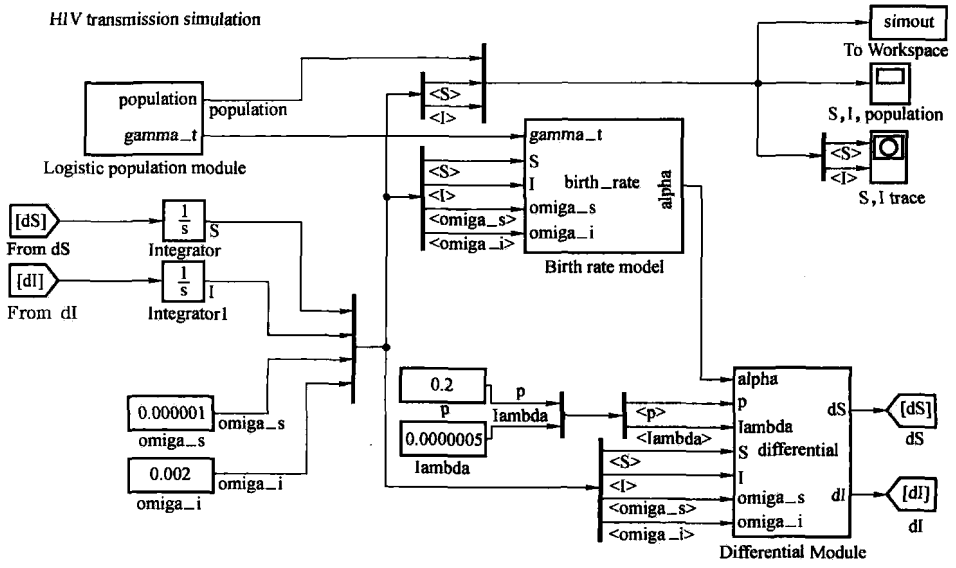


图 6-45　HIV 自由传播的 Simulink 仿真模型结构图

在仿真模型中，设置积分模块的初值，即可对不同时刻的 HIV 感染人群和易感人群数量仿真，结果如图 6-46 所示. 代入不同国家的实际参数，即可得到在不考虑 HIV 疫苗和 ARV 治疗的情况下，HIV 疫情的传播发展趋势.

从图 6-46 中可以看出，总人口数量呈现标准 Logistic 模型规律，最后人口总数趋于稳定. HIV 的感染人群逐步上升，易感人群逐步减少，最后稳定在一个固定的水平上.

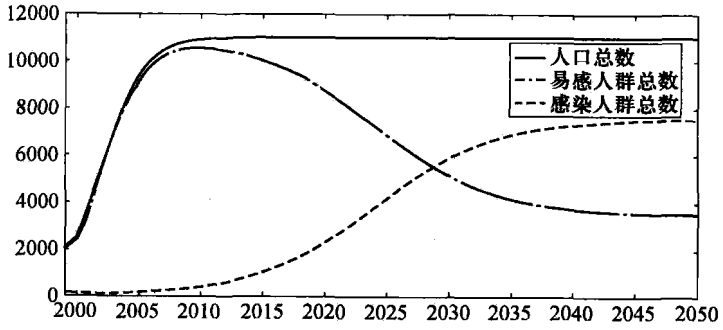


图 6-46　HIV 自由传播模型各类人群的变化规律

6.9.2　考虑 ARV 治疗传播的 Simulink 仿真模型

当实际中存在 ARV 治疗时，首先需要根据每年可投入的资金确定每年可开展的 ARV 治疗的病例数，然后将每年接受 ARV 治疗的病例数输入到仿真模型中. 修改相关仿真模块，即可实现考虑 ARV 治疗的 Simulink 仿真模型.

设政府每年可资助的 ARV 治疗的病例数为 $T(t)$，为了方便，这里设每年

接受治疗的人数占全部感染人数的 50%. 实际上，$T(t)$ 可以通过设置信号源的方式，根据政府可用防控资金总数来确定.

当考虑 ARV 治疗时，出生率模块和微分方程组模块中的内容需要按照模型 6.2 中的方法进行修改. 此外，在 6.9.1 节的仿真模型基础上，还需要增加常量输入：ARV 治疗人群的死亡率 omiga_t、母婴传播风险 M_t 和接受治疗后的传染率衰减系数 f. 这些数值都可以根据各个国家实际数据进行修改，以得到不同参数下的疫情发展的变化规律.

综上所述，可以得到在 ARV 治疗接受率为 50% 的情况下，疫情发展 Simulink 仿真模型结构如图 6-47 所示.

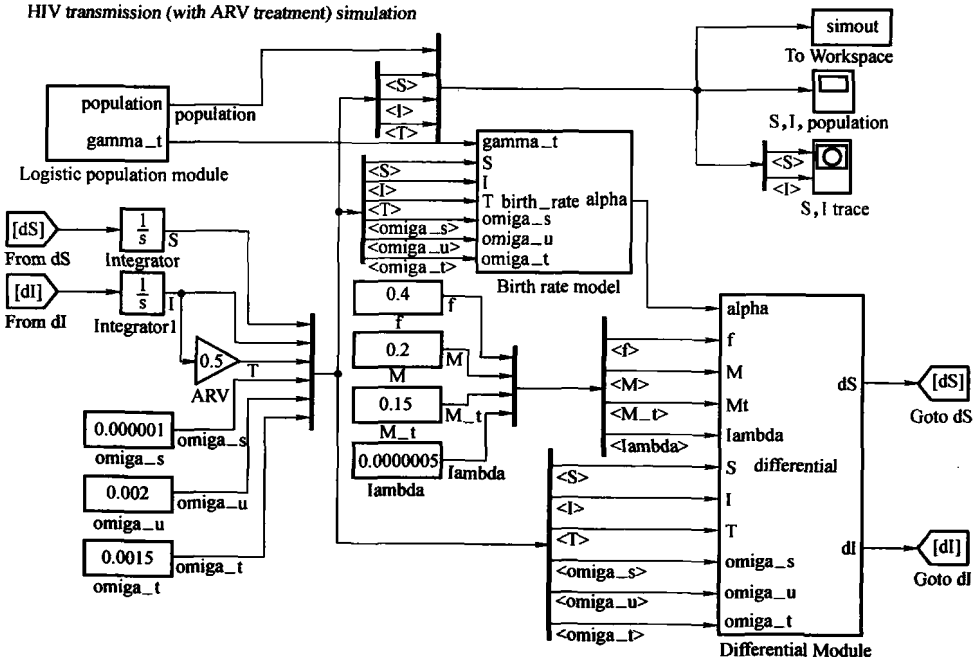


图 6-47 考虑 ARV 治疗的 Simulink 仿真模型结构图

对上述模型进行仿真计算得到易感人群 S，感染人群 I（包括未治疗人群 U 和治疗人群 T），各类人群的发展变化趋势如图 6-48 所示.

从图 6-48 可以看出，考虑 ARV 治疗后，各类人群在一定时间后开始稳定. 在这里来对比一下采用 ARV 治疗和不采用 ARV 治疗的两种情况，不妨取一组参数值，即 $f=0.4$，$M=0.2$，$M_T=0.15$，$parv=0.5$，$\Omega_T=0.15$，结果如图 6-49 所示.

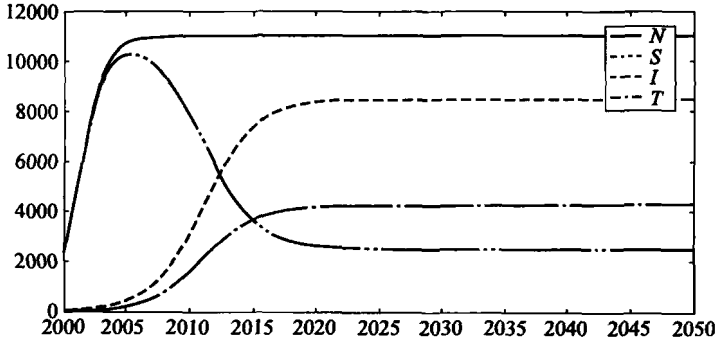


图 6-48 考虑 ARV 治疗时各类人群发展趋势

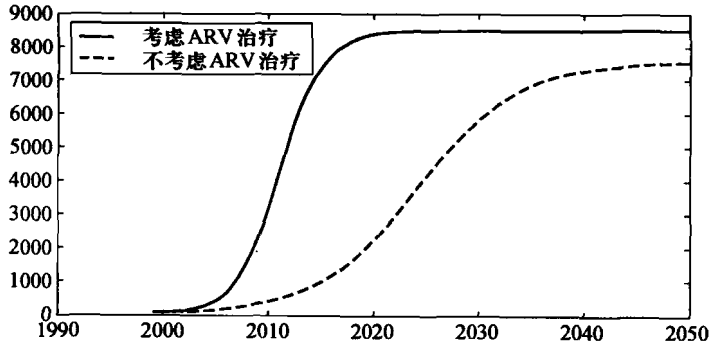


图 6-49 是否采用 ARV 治疗感染人群 $I(t)$ 的发展趋势

从图 6-49 中可以看出，开展 ARV 治疗并不能有效控制疫情，反而会导致疫情急剧上升，并且稳定在一个较高的水平上. 这是因为 ARV 治疗延长了 HIV 感染者的寿命，增加了感染易感者的可能性，抵消了 ARV 治疗而导致传染率的降低效果.

6.9.3 考虑 ARV 治疗以及 HIV 疫苗的 Simulink 仿真模型

在增加考虑 HIV 疫苗存在的情况下，需要在系统中增加一个状态变量 R，表示接种疫苗后的免疫人群数量. 同时加入控制输入参数 $p(t)$，表示 t 时刻新增接种疫苗的人数. 由于实际中所采取的控制策略不同，控制输入信号 $p(t)$ 也有所不同. 这里为了说明问题，不妨假设疫苗在 2015 年研制成功投入使用，并且每年新增接种数量为一常量.

将这一模块也封装到仿真系统中去，则得到同时考虑 ARV 治疗和 HIV 疫苗的 Simulink 模型结构图，如图 6-50 所示.

运行仿真模型，得到各类人群随时间变化的趋势如图 6-51 所示.

从图 6-51 中可以看出，接种疫苗后免疫人群 R 从 2015 年开始出现，并稳定增长. 随着免疫人群的增长，感染人口的比例开始出现增长变缓，直至开

始出现下降的趋势.

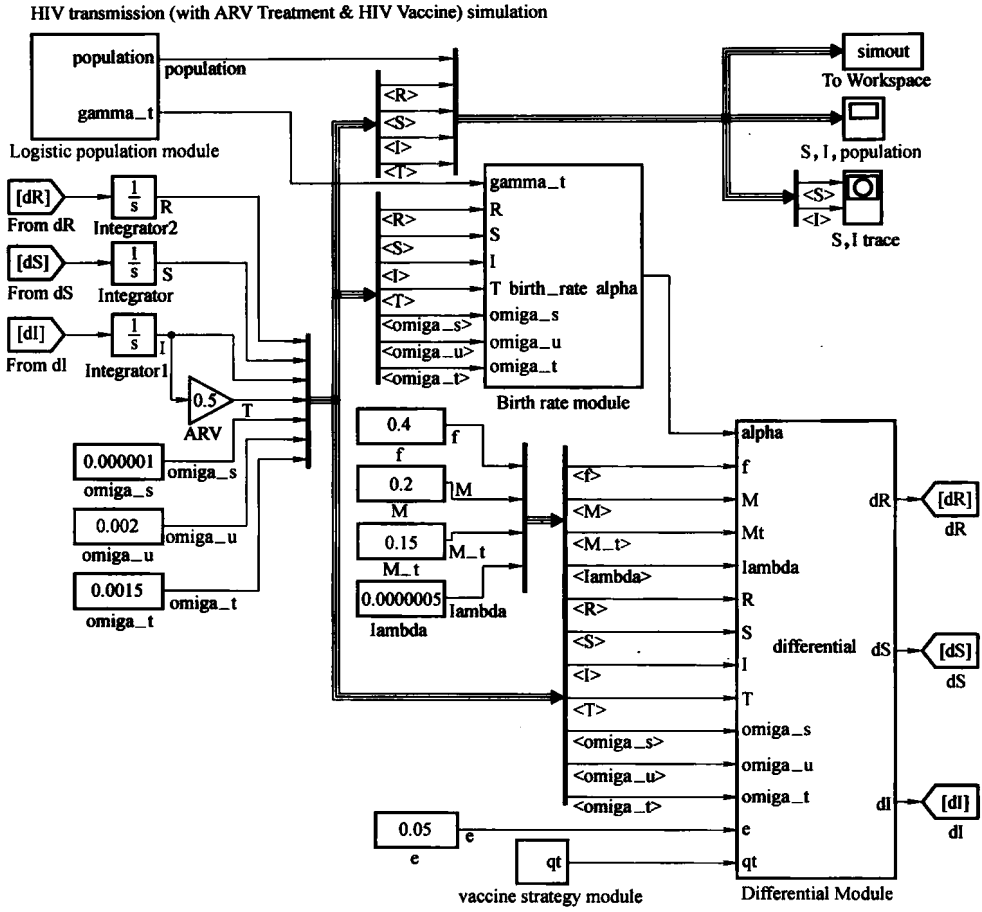


图 6-50 考虑 ARV 治疗和 HIV 疫苗的 Simulink 仿真结构图

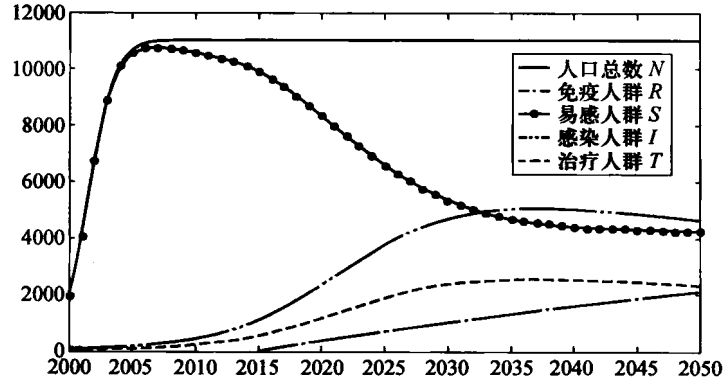


图 6-51 考虑 ARV 治疗和 HIV 疫苗各类人群的变化趋势

与图 6-48 相比，图 6-51 中感染人群 I 的数量比不考虑疫苗的情况要少，而且在后期出现了缓慢减少的趋势，这说明了疫苗在防控 HIV 中发挥的作用. 不同的疫苗推广时间及普及速度，都将会产生不同的控制效果.

6.9.4　考虑 ARV 抗药性与 HIV 疫苗的 Simulink 仿真模型

在 6.9.3 仿真模型的基础上，增加考虑 ARV 抗药性问题. 这时需要增加一类新的人群 W，表示抗药性人群的总数. 同时按照模型 6.6，更新出生率模块和微分方程模块的相关代码. 此外，为了表示在单位时间内产生抗药性的比例，增加一个输入参数 qw，表示单位时间内产生抗药性的病例占接受 ARV 治疗总人数的比例. 这样就可以得到如图 6-52 所示的仿真模型结构图.

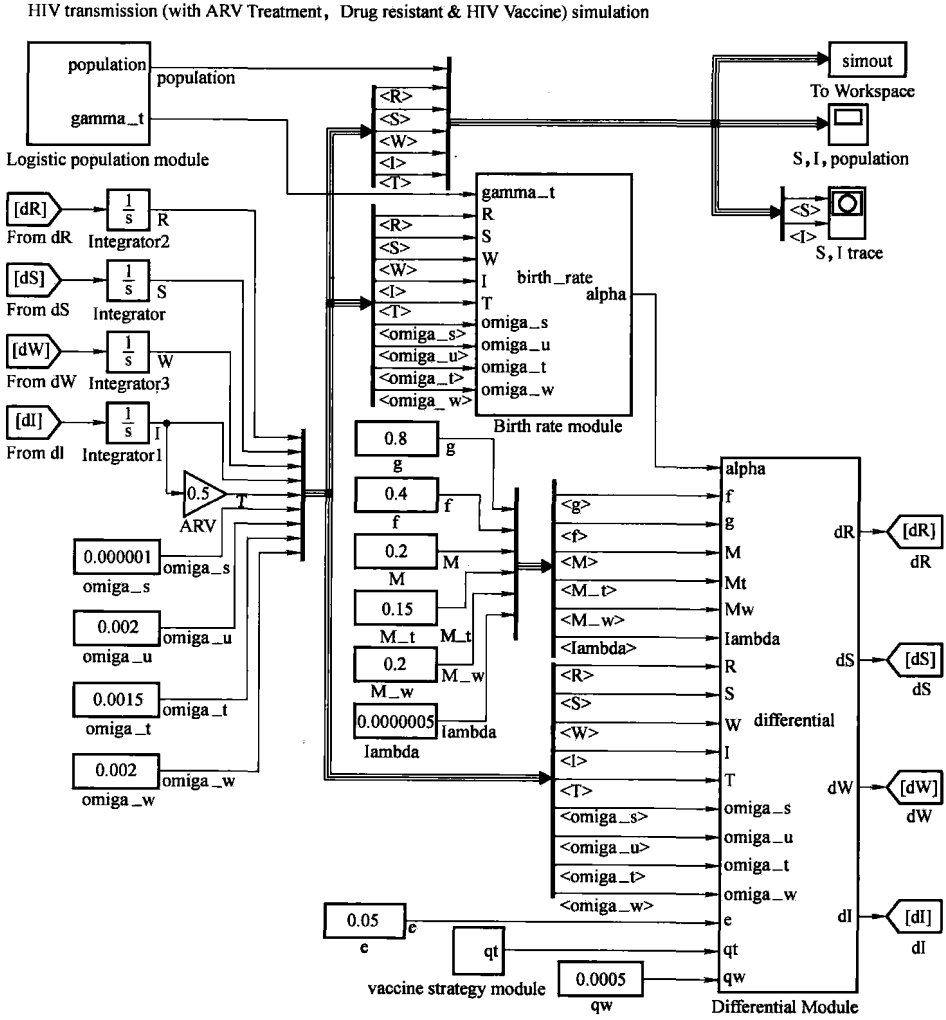


图 6-52　考虑 ARV 抗药性和 HIV 疫苗的 Simulink 仿真结构图

运行这个仿真模型，可以得到各类人群随时间变化趋势图，如图 6-53 所示.

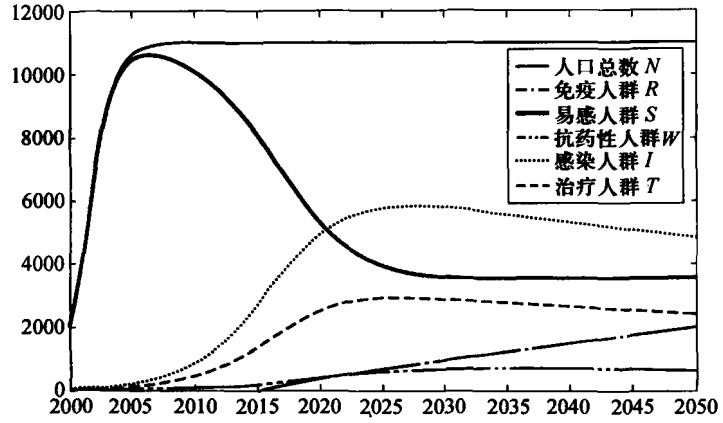


图 6-53 考虑 ARV 抗药性和 HIV 疫苗的各类人群随时间变化趋势图

从图 6-53 可以看出，抗药性人群一开始会有一个缓慢增长过程，之后开始逐渐下降. 这是因为随着时间的增长，HIV 感染人群的降低使得 ARV 治疗人群减少了，进一步地也减少了抗药性人群的数量. 同时，当新增抗药性的速度小于死亡的速度时，出现抗药性的人群数量也会逐步减少.

相比图 6-51 中不考虑抗药性的情况，图 6-53 中的感染人群 I 的人数要比图 6-51 中感染人群 I 的人数要多，这是因为考虑抗药性后，模型估计的疫情状况要比不考虑抗药性的更严重.

类似地，对其他各种具体模型都可以给出相应的 Simulink 仿真模型和结果分析.

6.10 HIV/AIDS 的预防控制策略

问题（4）要求给出 HIV/AIDS 防控资金的最优配置方案，并给联合国相关组织写一份建议书. 本节主要就防控资金用于开展 HIV 疫苗研发、推广使用计划以及对 AIDS 患者开展 ARV 治疗计划等方面进行讨论.

6.10.1 HIV/AIDS 防控资金的最优配置模型

HIV 疫苗研发推广计划与 ARV 治疗计划共同使用 HIV/AIDS 防控资金，该资金总量受该国可支配 AIDS 防控基金总量的限制. 若用于 HIV 疫苗基金的数量过少，则可能会使更多的人感染 HIV；若用于 HIV 疫苗基金的数量过多，就可能导致 ARV 治疗的基金数量不足，从而造成 HIV 感染者传染率的提高，最终导致 HIV 疫情恶化. 各国应该根据自己 HIV 疫情的特点以及可支配的

HIV 防控基金数量，合理配置 HIV 疫苗计划和 ARV 治疗计划.

通过前面的 Simulink 仿真模型，可以看出不同防控措施下的疫情发展趋势. 即通过传染病模型和仿真模型可以建立起 HIV 的防控策略与各类人群之间的函数关系：

$$[I,\ S,\ T,\ W,\ R]\ =F[p_{arv}(t),q(t)],$$

其中 I, S, T, W, R 表示近 50 年内各类人群随时间变化的数值，$p_{arv}(t)$ 表示每年开展 ARV 治疗的病例平均数，$q(t)$ 表示每年接种疫苗的数量，$F(\cdot)$ 为 6.9 节中的 Simulink 仿真模型所确定的函数关系.

防治策略指标 $p_{arv}(t)$ 和 $q(t)$ 都会受到一个国家可支配 HIV/AIDS 防控资金的限制，因此有

$$G[p_{arv}(t),\ q(t)]\leqslant g(t).$$

其中 $g(t)$ 是该国第 t 年的可支配 HIV/AIDS 防控资金总额，$G[p_{arv}(t),\ q(t)]$ 是防控策略 $p_{arv}(t)$ 和 $q(t)$ 所需要的资金额.

为了评价不同防控策略的优劣，需要建立疫情评价模型. 类似于问题(1)中选择危急国家所采用的疫情评价模型，可以针对各种疫情的危急程度进行评价，即取评价函数：

$$E=Q(I,\ S,\ T,\ W,\ R).$$

其中 $Q(I,\ S,\ T,\ W,\ R)$ 是根据各类人群总数而得到的疫情评价值的函数，函数值越大，疫情越严重.

综上所述，可以得到最佳防控资金配置的优化模型：

$$\min\ E=G(F[p_{arv}(t),\ q(t)])$$
$$\text{s. t. } G[p_{arv}(t),\ q(t)]\leqslant g(t).$$

求解上述模型，即可得到未来 50 年里的最佳资金配置策略.

由于一个国家每年的可支配金额是固定的，只要其中一个项目的资金确定后，剩余的资金就可以全部用于开展另一个项目. 实际中对于上面最优化模型的决策变量为 50 个，于是上面的最优化模型是包含 50 个变量的非线性规划问题.

为了考虑疫苗投入的时间对决策结果的影响，可以粗略估计疫苗研制经费与疫苗投入时间的关系模型. 为了方便，假设每年的疫苗研发费用是相同的，不妨设为 x，则疫苗投入使用的时间为 $T=V(x)$. 疫苗投入使用的策略要受到疫苗投入时间的限制，即

$$q(t)\ \begin{cases}=0, & t<T,\\ \geqslant 0, & \text{其他}.\end{cases}$$

如果疫苗研发投入的总经费为 xT，则在未来第 T（<50）年之前，用于 HIV/AIDS 的防控资金（包括 HIV 疫苗研发经费和 ARV 治疗项目经费）就有

$$G[p_{arv}(t), 0]+x \leqslant g(t), \ t \in [0, T).$$

而从第 T 年开始，HIV/AIDS 防控资金（包括 HIV 疫苗推广项目经费和 ARV 治疗项目经费）有

$$G[p_{arv}(t), q(t)] \leqslant g(t), \ t \in [T, 50].$$

于是得到考虑疫苗研发费用的最优资金配置模型：

$$\min E = Q(F[p_{arv}(t), q(t), x])$$

$$\text{s. t.} \begin{cases} G[p_{arv}(t), 0]+x \leqslant g(t), \ t \in [0, T), \\ G[p_{arv}(t), q(t)] \leqslant g(t), \ t \in [T, 50]. \end{cases}$$

求解该模型即可得到最优的资金配置策略.

6.10.2 给联合国相关部门的建议书

问题（4）要求给联合国相关部门写一份建议书，在撰写建议书时候应注意建议书的阅读对象为 UNAIDS，而不是数学专业的读者，因此建议书中不应出现各种数学符号和各种推导过程. 应有明确的结论以及各种可靠的数据和图表，并且这些结论、数据和图表都要基于前面建立的模型，而不是自己的理解和照搬现有文献的结论. 要针对不同国家的国情有针对性地给出建议，不能泛泛而谈，撰写时力求做到有理有据.

下面给出西点军校优秀论文的建议书的内容.

非洲撒哈拉沙漠以南地区是控制艾滋病传播的关键区域. 2006 年全世界有超过 70 亿美元的资金可用于防控艾滋病，我们建议将其中的 45 亿美元用于非洲撒哈拉沙漠以南地区开展 HIV 检测、艾滋病预防教育、分发预防器具、对感染者进行能够降低传染率的治疗，以及为感染 HIV 母亲的婴儿提供婴儿配方奶粉. 该计划将持续 3 年，每年投入的比例相同.

对非洲开展 3 年的特别投入项目后，需要对该项目的投入回报进行评价分析. 如果该项目不成功，就需要更改计划；如果项目成功，在保证非洲得到足够的资金用于 HIV 教育、预防、治疗和健康母乳的前提下，逐渐将防控重点转移到印度和东南亚. 印度的防控资金的分配与非洲的类似，主要用于检测、教育、预防、控制性治疗以及健康母乳. 该计划的实施时间长度要比在非洲的更长，将达到 5 年的时间.

东南亚和非洲撒哈拉沙漠以南地区的艾滋病问题是世界人口健康的最大威胁，当这些地区的感染者得到了治疗后，该项目将进入疫苗接种推广项目阶段. 该阶段应该将防控的精力集中在 HIV 检测、教育和预防上，并且谨慎地开展 ARV 治疗以避免患者产生抗药性以及产生病毒变异. 此时需要重新对计划的实施效果进行评价，并且需要

针对可能面临的新威胁及时调整计划.

6.11 写在后面：ICM 优秀论文评阅标准与启示

每年的 ICM 竞赛都吸引了大量的中国大学生参加，ICM 的问题和 MCM 的问题有着明显的不同，参赛论文的写作也不相同. 2006 年 ICM 的问题是非常有代表性的一个问题，之后命题人给出了一篇很具有启发性的评述文章，在这里推荐给读者，应该会对想参加竞赛的学生有一定的启示.

2006 年 ICM 的问题是研究如何在全世界范围内合理配置抗 HIV/AIDS 资源的策略问题，命题人在问题中提供了一些国家和地区在防控 HIV 方面的历史和预测数据. 这些数据包括 AIDS 的扩散范围、HIV/AIDS 的分布图、人口统计数据、生育率、年龄分布、出生率和平均寿命等. 要求学生利用这些数据，完成 4 项任务. 下面针对论文中的各个要点分别进行评述.

1. 关于摘要

所有的参赛队都知道一份内容正确、条理清晰的摘要的重要性，但是往往不知道如何才能写好摘要. 一种比较可取的做法是对各种结论的必要条件进行充分的叙述，采用诸如“通过我们的分析，若能够有效开展 ARV 治疗以及/或者预防性 HIV 疫苗的研发成功，将能够实现 HIV/AIDS 病例数量的减少.”而更好的做法是采用“高度概括”的写法，在摘要中将文章中最重要的结论表述出来，使用类似“我们发现 ARV 的治疗成本过于高昂，对于贫穷国家来说根本是无法承受的，而富裕一些的国家则可以通过他们自己国家的财政经费开展费用高昂的 ARV 治疗计划，因此联合国应该将精力集中在预防性疫苗的开发和推广上，这样就可以让每个国家都能够使用到价格低廉、有效的 HIV 疫苗了.”好的文章摘要需要花费大量的精力总结结论，而不仅仅是叙述问题.

2. 关于背景知识的表述

大多数论文都分章节地叙述了 HIV/AIDS 的传播机理，包括疾病的生物学特征、传播机制、流行病学特征、当前 ARV 治疗的作用以及 HIV 预防疫苗的潜在价值. 作为一种传染病，HIV/AIDS 的全球扩散性、高传染率、ARV 治疗后 HIV 的高突变率以及至今仍然没有有效的疫苗等特征，使得 HIV/AIDS 已经成为当前最值得关注的疾病之一.

虽然大多数的论文都给出了 HIV/AIDS 的传播机理，但是有些论文在叙述上更为准确和具体. 那些不成功的参赛论文只是罗列了大量网上的数据，甚至包括了大量对解决问题根本就没有用的数据，没有将精力集中在问题需要研究的几个重点区域和国家中，没有很好地组织这些信息. 这些论文在评审的第一个阶段就被淘汰了.

被评为中等的论文，主要问题是各个章节之间缺乏联系. 这些文章出现了这种现象，一个组员写背景知识，这一部分条理较为清晰；而后面的模型建立和求解的章节由其他组员撰写. 虽然每个章节单独看上去都很不错，但是这些章节之间却没有很好地联系起来，后面的模型建立过程中并没有很好地与前面的背景知识联系起来，各个章节内容相对孤立.

一篇优秀的论文应该包含丰富但不是全部的问题背景知识，并且能够和后面的各个模型对应起来. 例如，当一个组员在背景知识中写到 ARV 治疗能够延长 HIV 感染者的寿命而不能阻止病毒的传播时，基于这个背景知识在模型假设中假设 ARV 治疗无法降低 HIV/AIDS 的传染率后，就需要在模型建立过程中反复重申并利用这一结论. 一篇优秀的论文只给出对于建模必要的背景知识，而对于不必要的背景知识几乎不提.

优秀论文的模型假设和他们的模型一样条理清晰. 这些论文在不违背问题真实机理的前提下，做出适当的假设，简化了问题.

3. 关于模型的选择

从提交的论文上看，模型主要分为两大类：基于微分方程的模型（微分方程模型或者差分方程模型）和基于统计的模型（概率模型或转移概率模型）. 两类模型都可以，评阅人对两类模型没有任何的偏好，两类模型都有中等和优秀的论文，不会以此作为依据评阅论文.

对于这两类模型，被评为中等水平论文的特点有以下几点：

（1）模型的细节过多.

（2）问题的假设不够简练. 模型考虑的要素过多，而对于问题本质的要素没有把握好.

（3）陈列出很多的结果，没有突出其中重要的结果，而且缺乏对这些结果含义的解释.

（4）没有对模型结果的合理性进行足够的讨论和解释，也没有在问题分析中给出模型合理性的解释.

在一些论文中，模型选择上生搬硬套，缺乏充分的理由. 例如在某个章节套用微分方程模型时，给人的感觉是他们选择这个模型是因为他们在竞赛前准备过相关模型，或者是这是他们在查阅资料时第一个看到的模型，而没有思考这个模型是否是最适合的. 就像一些评阅人之比喻的：“一些论文看上去像是拿着锤子找钉子，而不是看到钉子去思考用什么工具去敲它.”

一个好的模型应该具备如下特征：

（1）针对每个假设给出合理的解释.

（2）为了使模型更易处理，给出一些合理的假设.

（3）只给出后面有具体分析的必要结果.

（4）结合一些具体的案例，对各个结果进行详细的解释.

(5) 对模型结果的合理性进行讨论，并且在进一步的分析中要用到这些结果.

在每次竞赛中，各个参赛队的模型往往会集中到几个大类中，但也偶尔有个别组提交了独具创新的模型. 在2006年的竞赛中就出现了一个参赛队用元胞自动机的方法研究HIV/AIDS的扩散规律. 这个队仿真了一个包括10 000个个体的人群的6个方面的特征（年龄、性别、社会联系活跃度、社会身份、健康状况和生活圈）. 仿真过程的每一步，每个元胞都根据它周围的8个元胞的状态，依据问题参数所建立起的变化规则，确定状态发生改变（从未感染到感染）或者没有改变（仍然没有感染或者仍为感染）. 经过多轮的仿真计算，统计问题所需要的数据，完成了相关任务. 一些评阅人认为这篇论文思路比较独特，虽然这个模型需要很多的参量设置，并且需要细致地调整各个参数才能使模型正常地运行，但是这个模型能够提供其他模型所不具备的制定策略的结果.

4. 关于结果分析

一篇优秀的论文会花很多时间研究和讨论他们的模型针对问题的需求，在什么条件下更为有说服力，什么条件下缺乏说服力. 他们同时花时间去验证模型的灵敏性和鲁棒性. 例如，同时改变一些参数的值，计算结果是如何改变的，或者通过一些简单的讨论（例如，让系统的线性常微分方程组各个参数进行10%的扰动，会导致结果也有相应的变化，这是因为方程组的各个特征值的差异很大）来分析灵敏性和鲁棒性.

这个问题的解决还需要更多的时间，甚至再给一个长周末时间也是不够的. 即使是Outstanding论文，我们也发现在资金方面的讨论稍显不足，这一点让评阅人感到比较失望，因为这是竞赛的主要问题之一.

5. 关于建议书

一篇中等水平的论文，往往看上去像是一个对于建模和分析不完全清楚的组员撰写的论文. 这些论文的建议书更多地依赖于自己的理解和问题的背景研究，而不是依据他们模型的结果来写. 通常这些论文的执行建议都没有区分不同的国家、区域和社会类型.

优秀论文的建议书是基于对问题的深入分析写成的. 他们提出的措施有效性可以得到自己建立模型的验证. 这些论文在他们陈述自己的建议时引用了很多他们分析的结论，使得他们的建议更有说服力. 这些论文能够根据各个国家的国力和人口特征给出相关的策略和资源配置方案. 并且在建议书中给出了建议的假设前提. 例如，若因为HIV的高发病率，目标是减少病人的痛苦，就应该优先考虑开展ARV治疗，如果目标是控制HIV在世界范围的传播，减少感染人数，就应该优先考虑开展疫苗项目的研究.

6. 关于论文表述

优秀的论文要确保各个章节之间有密切关联，不是仅仅通过几句话过渡，而是要让各个章节之间有逻辑联系. 尤其是在模型建立、问题分析和建议书部分. 合适的图表能够有效地引起注意，但是不必要的图表会分散评阅人的注意力.

此外，一篇简明扼要、论据丰富的论文是论文写作的首选. 一篇冗长、凌乱的论文会因为过多的细节和重复而导致难以阅读，最终被淘汰.

参考文献

[1] Wikipedia. AIDS [OL]. http://en.wikipedia.org/wiki/AIDS. 2011-09-03.

[2] Koopman J S, et al. The role of early HIV infection in the spread of HIV through populations [J]. J Acq Imm Def Syn Hum Retroviral, 1997, 14: 249-258.

[3] AVERT. org [OL]. http://www.avert.org.2006-02-06.

[4] Martin Gayle. A Comparative Analysis of the Financing of HIV/AIDS Pro-grammes [M]. Chicago, IL: Independent Publishing Co., 2003.

[5] Anderson Deborah J, Thomas R O'Brien, et al. Effects of disease stage and Zidovudine therapy on the detection of human immune-deficiency virus Type 1 in semen [J]. Journal of the American Medical Association, 1992, 267: 2769-2774.

[6] Krieger John N, Robert W Coombs, et al. Recovery of human immuno-deficiency virus Type I from semen: Minimal impact of stage of infection and current antiviral chemotherapy [J]. Journal of Infectious Diseases, 1991, 163: 386-388.

[7] Royce Rachel A, Arlene Seina, et al. Sexual Transmission of HIV [J]. New England Journal of Medicine, 1997, 336: 1072-1078.

[8] Hyman James M, Jia Li, and E. Ann Stanley. Modeling the impact of randomscreening and contact tracing in reducing the spread of HIV [J]. Mathe-matical Biosciences, 2003, 181: 17-54.

[9] Leynaert Benedicte, Angela M Downs, and Isabelle de Vincenzi, For the Euro-pean Study Group on Heterosexual Transmission of HIV. Heterosexual transmission of human immunodeficiency virus: Variability of infectivity throughout the course of infection [J]. American Journal of Epidemiology, 1998, 148 (1): 88-96.

[10] Orendi J M, K Boer K, et al. Vertical HIV-I-transmission. I. Risk and prevention in pregnancy [J]. Nederlands Tijdschrift voor Geneeskunde, 1998, 142: 2720-2724.

[11] UnitedNations PopulationDivision. AIDS/HIV effect on life expectancy. In World Population Prospects: The 2000 Revision Highlights [OL]. http://www.overpopulation.com/faq/health/infectious_diseases/AIDS/life_expectancy.html,2001-02-28.

[12] United Nations Office on Drugs and Crime. 2005 World Drug Report. Vol. 1: Analysis [OL]. http://www.unodc.org/pdf/WDR_2005/volume_1_web.pdf.2006-02-04.

[13] Francoeur Robert T, Raymond J Noonan, et al. (eds.). The Continuum Complete International Encyclopedia of Sexuality [M]. New York: Continuum, 2004.

[14] Mackay Judith. The Penguin Atlas of Human Sexual Behavior [M]. New York: Penguin Reference, 2000.

郑重声明

高等教育出版社依法对本书享有专有出版权。任何未经许可的复制、销售行为均违反《中华人民共和国著作权法》，其行为人将承担相应的民事责任和行政责任；构成犯罪的，将被依法追究刑事责任。为了维护市场秩序，保护读者的合法权益，避免读者误用盗版书造成不良后果，我社将配合行政执法部门和司法机关对违法犯罪的单位和个人进行严厉打击。社会各界人士如发现上述侵权行为，希望及时举报，本社将奖励举报有功人员。

反盗版举报电话　(010) 58581897　58582371　58581879

反盗版举报传真　(010) 82086060

反盗版举报邮箱　dd@ hep. com. cn

通信地址　北京市西城区德外大街 4 号　高等教育出版社法务部

邮政编码　100120