arXiv:2301.04141v1 [cs.LG] 11 Jan 2023

机器学习在天然气燃烧中的应用

by

Rong Lu

© 版权所有：Rong Lu，2020年保留所有权利

提交给科罗拉多矿业学院教员委员会和董事会的论文，旨在部分满足获得哲学博士学位（石油工程）的要求。

科罗拉多州戈尔登

Date _____

签名：~~Rong Lu~~

签名：~~Jennifer L. Miskimins~~ 博士
论文指导教师

科罗拉多州戈尔登

Date _____

签署人：~~Jennifer L.~~ Miskimins 博士，教授兼石油工程系主任

摘要

目前，在石油工业中，运营商常常将伴生气火炬燃烧，而不是将其商品化。有些州的燃烧放空量很大，由此造成能源浪费和 CO_2 排放等问题。在北达科他州，运营商必须估算并报告火炬燃烧量。问题是，这些报告的质量如何？我们能从中得出哪些见解?

除可从北达科他州工业委员会（NDIC）获取的公司报告统计数据外，还可以通过卫星遥感估算燃烧量，作为无偏基准。由于辉光造成的伪影会妨碍对陆地卫星8号影像的解读，因此本文采用基于可见光红外成像辐射计套件（VIIRS）的估算体积。通过逆地理编码，在县和油田等不同层级对比 NDIC 与 VIIRS 数据。

收集并预处理完所有数据后，采用马尔可夫链蒙特卡罗方法实现贝叶斯学习，以解决三个问题：县级模型开发、燃烧放空时间序列分析和分布估计。首先，不同县的NDIC与VIIRS体积之间的关联存在异质性。鉴于此，利用层级模型为每个县开发模型。其次，尽管燃烧放空时间序列含有噪声，但其中包含有关趋势和模式的信息，可为了解运营商方法提供一些洞见。研究发现，高斯过程在许多不同的模式识别场景中都很有效。第三，通过无监督学习获得分布洞见。研究发现，负二项分布和高斯混合模型分别能有效描述油田放空燃烧次数和燃烧气体体积的分布。最后，介绍一种基于最近邻的方法，用于运营商级监测与分析。

目录

摘要.iii 图目录.vii 表目录.xii 符号表.xiii 缩略语表.xvii 致谢.xix 献辞.xxi

第1章 引言.1

1.1 研究目标.3 1.2 学位论文目标.4 1.3 大纲与贡献.5

第2章 文献综述.7

2.1 卫星图像处理.9 2.2 贝叶斯推断.11 2.3 马尔可夫链蒙特卡罗.12 2.4 机器学习.14 2.5 分析工具集.15

第3章 数据预处理与探索性数据分析 . 17 3.1 数据收集.17 3.1.1 Landsat 8图像.17

3.1.2 VIIRS估算体积.18 3.1.3 NDIC月度生产报告.18 3.1.4 NDIC形状文件.18
3.2 卫星图像处理.20 3.3 逆地理编码.23 3.4 相关性分析.23 3.5 州级火炬燃烧模型.26

第4章 县级火炬燃烧模型 .32

4.1 学习异质性.32 4.2 层次模型.32 4.3 数据描述.35 4.4 模型设定.37 4.5 模型重新参数化.41 4.6 模型拟合.42 4.7 模型可扩展性.47

第5章 燃烧放空时间序列分析.52

5.1 学习燃烧放空模式与行为. 52

5.2 高斯 过程 53

5.2.1 Mean 函数 54

5.2.2 协方差函数 . 54

5.2.3 推断与模型重新参数化 . 56

5.3 用于模式识别的模型系列 . 57

5.3.1 燃烧天然气比例建模 . 57

5.3.2 建模井燃烧比例.63 5.3.3 建模火炬检测次数.66 5.3.4 建模燃烧原油比例.70 5.3.5 建模 VIIRS 与 NDIC 之间的比例因子.74 5.3.6 预测 NDIC 火炬燃烧量.80 5.3.7 回顾先验选择.81

第6章 多视角无监督学习 . 83

6.1 学习分布 . 83

6.2 概率模型估计 . 84

6.3 VIIRS检测次数建模 . 85

6.4 燃除量建模 . 91

6.4.1 模型设定.93 6.4.2 模型比较.96 6.4.3 聚类.99

第7章 讨论 . 102

7.1 运营商级监测与分析.102 7.2 不一致警告.106 7.3 石油数据分析中的注意事项.111

第8章 结论与建议.117

8.1 结论.117

8.2 未来工作.119

图表目录

图1.1 2018年按火炬燃烧气体体积排名的前30个国家.1 图1.2 时间序列显示了美国火炬燃烧量排名前两位的州的趋势（美国能源信息署 2019a）.2 图1.3 这张谷歌地球影像显示了北达科他州某油井位置的天然气燃烧情况（谷歌地球 2019）.3 图1.4 原始海报的一部分（佩恩研究所地球观测组 2019），使用一年累积的 VIIRS低光成像数据展示人类活动，例如天然气燃烧、捕鱼和城市灯光。 . . . . 4 图2.1 陆地卫星8号的空间分辨率（美国国家航空航天局 2020）.8 图 2.2 四条随机游走 Metropolis 马尔可夫链的演变（Carpenter 2020），每条链均从不同位置开始.14 图 2.3 改编自（Betancourt 2020）的流程图展示了严谨的贝叶斯工作流程.16 图3.1 2018年10月生产报告中前 ~ 50 行的截图。每行对应一口油井。此电子表格共包含17，135行，第一行为表头.19 图3.2 VIIRS（上）和Landsat 8（下）的夜间燃烧源探测限。对于温度通常高于1500开尔文的天然气燃烧，Landsat 8检测到的火炬源区面积（约 $10^{-2}\,\mathrm{m}^2$）比VIIRS（约1 m^2）小几个数量级.21 图3.3 簇大小分布的计数图：显然存在一定数量的大簇（如右侧长尾所示）。 . . . . 22 图3.4 一个由许多热像元（检测结果）组成的大型火炬，这是通过在L8 图像上运行Nightfire算法发现的。图中展示了波段6（灰度图像）和KMZ视图，由Christopher D. Elvidge提供（私人通信）. 22 图 3.5 热图显示了各项之间的成对斯皮尔曼相关性

g最终时间序列的月度 y 观测值.25

图 3.6 对时间序列进行一阶差分后，显示其两两斯皮尔曼相关性的热图.26 图3.7 NDIC 和 VIIRS 报告数据的可视化结果.27 图 3.8 州级火炬燃烧模型的后验分布（左列）和轨迹图（右列）.29 图 3.9 使用后验预测样本构建的区间.30 图 4.1 不同县的 NDIC 和 VIIRS 报告数据散点图. 36

图 4.2 不同县的 NDIC 和 VIIRS 报告数据散点图，所有子图均不共享 x 轴或 y 轴.38

图4.3 LKJcorr($\eta = \texttt{eta}$) 概率密度.40

图4.4 每个县的斜率后验分布和轨迹图。. . . . 43

图4.5 每个县的截距后验分布和轨迹图。. 44

图 4.6 显示每个县斜率估计不确定性的森林图.45

图4.7 显示每个县截距估计不确定性的森林图.45

图 4.8 截距与斜率之间的相关性.48 图 5.1 Blue Buttes油田天然气燃烧比例模型的后验分布和轨迹图.60 图 5.2 显示 Blue Buttes油田天然气燃烧比例变化的后验预测样本.60 图 5.3 运营商A天然气燃烧比例模型的后验分布和轨迹图.61 图5.4 显示运营商A天然气燃烧比例变化的后验预测样本.62 图 5.5 Blue Buttes油田油井燃烧比例模型的后验分布和轨迹图.64

图 5.6 显示 Blue Buttes油田井燃烧比例变化的后验预测样本.64 图5.7 运营商A油井燃烧比例模型的后验分布和轨迹图.65 图5.8 显示运营商A井燃烧比例变化的后验预测样本.66

图 5.9 Blue Buttes油田燃除计数模型的后验分布和轨迹图.68

图 5.10 展示 Blue Buttes油田放空燃烧次数变化的后验预测样本.68

图5.11 北达科他州放空燃烧次数模型的后验分布和轨迹图.69

图5.12 展示北达科他州放空燃烧次数变化的后验预测样本.70

图 5.13 Blue Buttes油田 BOE 放空比例模型的后验分布和轨迹图.72

图 5.14 显示 Blue Buttes油田 BOE 火炬燃烧比例变化的后验预测样本.72

图 5.15 运营商A BOE 放空比例模型的后验分布和轨迹图.73

图 5.16 显示运营商A BOE 火炬燃烧比例变化的后验预测样本.74

图5.17 北达科他州 VIIRS-NDIC 比例因子模型的后验分布和轨迹图.77

图 5.18 显示北达科他州比例因子变化的后验预测样本.78

图5.17 Blue Buttes油田 VIIRS-NDIC 比例因子模型的后验分布和轨迹图.79

图5.20 显示 Blue Buttes油田比例因子变化的后验预测样本.80

图 5.21 显示未来六个月比例因子预测的后验预测样本.81 图 6.1 直方图的有效使用可能出人意料地微妙.84 图 6.2 2018年10月油田检测次数分布的直方图.87 图 6.3 油田检测计数分布的后验分布和轨迹图，该分布使用2018年10月的数据进行拟合.88

图 6.4 2018年10月油田探测次数分布的直方图.89

图 6.5 2018年10月油田探测次数分布的直方图，纵轴经过截断，以便更清晰地呈现大于零的计数.90

图6.6 2018年第四季度油田燃除气体体积分布的直方图。 92

图6.7 2018年第四季度油田燃烧量大小的分布。 . . . . 93 图 6.8 从具有 $\boldsymbol{\alpha} = (6, 6, 6, 6, 6, 6, 6)$ 的狄利克雷先验中随机抽取的十个样本。 . . . . 95 图 6.9 不同 K 下的高斯混合模型推断结果.97

图 6.10 不同 K 的 WAIC值.98

图 6.11 2018年第四季度不同油田石油产量与燃除气体体积的散点图.100

图7.1 在运营商层面，NDIC 与 VIIRS 报告体积之间良好拟合的示例.105

图7.2 在运营商层面，NDIC 与 VIIRS 报告体积之间拟合较差的示例.106

图7.3 两个示例运营商的时间序列，其报告结果与 VIIRS 检测到的趋势/模式未能完全吻合。公司报告数据与卫星检测结果存在显著差异的时间点或周期已作注释.107 图7.4 运营商D的更完整时间序列图。. . . . . 109 图7.5 运营商E的更完整时间序列图。. 110

图 7.6 为假设的井性能分类问题设计的神经网络.115

表格目录

表 2.1 陆地卫星8号和 VIIRS 的分辨率.7 表 3.1 州级燃烧模型的参数估计值.28 表4.1 北达科他州县缩写.37 表 4.2 县级燃烧模型的参数估计值.49 表 6.1 油田检测次数分布的参数估计值.88 表 7.1 图 7.4 和图 7.5 中运营商时间序列的个体.111 表 7.2 OnePetro 上的出版物数量增长.112 表 8.1 本学位论文中建立的模型.119

符号表

向量与矩阵采用粗体表示。下标星号（例如 y_*）表示引用测试集数量或预测值。

一般符号说明

ℓ_2 范数. $\|\cdot\|$数据集： $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid i = 1, \dots, n\}$. $\mathcal{D}$含零自然数.$\mathbb{N}_0$斜体 Pi，表示变量.π 正体 Pi，表示超越常数（$3.14159\cdots$）.π对测试输入 $\mathbf{x}_*$ 的预测. y_*正比于；例如， $p(x \mid y) \propto p(x, y)$ 表示 $p(x \mid y)$ 等于 $p(x, y)$ 乘以一个与 x 无关的因子 .$\propto$ 实数 .$\mathbb{R}$ 全称量词：对所有 x . $\forall xa$ 定义为 b . $a \coloneqq bb$ 定义为 a . $a =: bn$ 维实数向量空间.$\mathbb{R}^n$

概率与统计

条件概率密度函数. $p(\cdot \mid \cdot)$ 期望；当 $x \sim p$ 时 $g(x)$ 的期望.$\mathbb{E}$ 或 $\mathbb{E}_p[g(x)]$概率密度函数.$p(\cdot)$

概率质量函数. $P(\cdot)$ 随机变量 X 服从 p . $X \sim p$方差; 当 $x \sim p$ 时 $g(x)$ 的方差.V 或 V $_p[g(x)]$

概率分布

参数为 n、 p 的二项分布.二项分布(n, p) 参数为 p 的分类分布.分类分布(p) 参数为 a、 b 的连续均匀分布 . 连续均匀分布 (a, b) 参数为 $\boldsymbol{\alpha}$ 的狄利克雷分布 . 狄利克雷分布 ($\boldsymbol{\alpha}$) 参数为 η、 $\boldsymbol{\sigma}$ 的经乔列斯基分解的协方差矩阵分布.LKJCholeskyCov(η, $\boldsymbol{\sigma}$) 参数为 λ 的指数分布.指数分布(λ) 参数为 α、 β 的伽马分布.伽马分布(α, β) 参数为 γ 的半柯西分布.半柯西(γ) 参数为 σ 的半正态分布.半正态(σ) 参数为 η 的 LKJ分布.LKJcorr(η) 参数为 $\boldsymbol{\mu}$、 $\boldsymbol{\Sigma}$ 的多元高斯分布.多元高斯分布($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) 参数为 μ、 ϕ 的负二项分布.负二项分布(μ, ϕ) 参数为 λ 的泊松分布.泊松(λ) 参数为 ν、 μ、 σ 的 Student- t 分布.Student-t(ν, μ, σ) 参数为 μ、 σ 的单变量高斯分布. $\mathcal{N}(\mu, \sigma)$

高斯过程

协方差函数 在$\mathbf{x}$ 和$\mathbf{x}'$ 处求值.k($\mathbf{x}$,$\mathbf{x}'$)

高斯过程： $f \sim \mathcal{GP}(m(\mathbf{x}),\ k(\mathbf{x},\mathbf{x}'))$，函数 f服从高斯过程.$\mathcal{GP}$在$\mathbf{x}$ 处求值的均值函数.$m(\mathbf{x})$ 潜在函数值向量，$\mathbf{f} = (f(\mathbf{x}_1)\ldots, f(\mathbf{x}_n))^\top$.$\mathbf{f}$

向量与矩阵

乔列斯基分解：$\mathbf{L}$ 是一个下三角矩阵，使得 $\mathbf{L} \cdot \mathbf{L}^\top = \mathbf{K}$.乔列斯基分解$(\mathbf{K})$ 大小为 $n \times n$ 的单位矩阵.$\mathbf{I}_n$矩阵 $\mathbf{L}$ 的转置.$\mathbf{L}^\top$长度为 n 的全零向量. $\mathbf{0}_n$长度为 n 的全一向量. $\mathbb{1}_n$参数向量.$\boldsymbol{\theta}$

标准函数

逆逻辑函数.$\text{logit}^{-1}(\cdot)$ 自然指数函数.$\exp(\cdot)$ 自然对数函数.$\log(\cdot)$

个体

桶油当量.BOE 桶/日.bbl/天 十亿立方米.bcm 天.d 美元/桶.$/bbl 美元/百万英热单位.$/MMBtu

开尔文.K 千米.km 兆瓦.MW 米.m 百万立方英尺.MMcf 千立方英尺.Mcf 每桶千立方英尺.Mcf/bbl

缩略语列表

自回归积分滑动平均.差分自回归移动平均模型 可信区间.可信区间 深度学习.深度学习 基于密度的含噪声应用空间聚类.DBSCAN 能源信息署.美国能源信息署 例如（拉丁语：例如）.例如 气油比.气油比

高斯混合模型.高斯混合模型

高斯过程.GP 哈密顿蒙特卡洛.HMC 层次化基于密度的含噪声应用空间聚类.HDBSCAN 最高密度区间.HDI 即（拉丁语：也就是说）.i.e. 独立同分布.i.i.d. 四分位距.IQR 核密度估计.KDE 陆地卫星8号.L8 长短期记忆.LSTM 马尔可夫链蒙特卡罗.MCMC

最大后验.MAP 最大似然估计.MLE 美国国家航空航天局.NASA 美国国家海洋和大气管理局.NOAA 北美基准.NAD 北达科他州工业委员会.NDIC 预测区间.PI 辐射热.RH 右侧.RHS 短波红外.SWIR VIIRS Nightfire.VNF 可见光红外成像辐射计套件.VIIRS 世界大地测量系统.WGS 零膨胀泊松分布.ZIP 零膨胀负二项分布.ZINB

致谢

首先，我谨向我的导师 Jennifer L. Miskimins 博士致以最诚挚的谢意。自2011年以来，Miskimins 博士一直担任我的硕士/博士导师和指导老师。2017年我重返 Mines 攻读博士学位以来，Miskimins 博士给予我的指导、支持和机会之多，远超我的想象。第一学期，我在 High Bay 担任实验室助理；之后的一个学期，我在她的油井增产课程中担任助教；自从我开始对机器学习产生兴趣以来，她为我创造了大量机会，让我与不同团队的人交流，一起集思广益，并探索自己的研究兴趣。在某种程度上，我觉得自己终于成为一名“合格”的 FAST 学生成员，这一切都要归功于这些宝贵的经历。若没有 Miskimins 博士的指导与支持，我不可能取得如今的成就，包括完成这篇学位论文。她世界一流的专业技术、对待工作与生活的态度，以及管理不同层级团队的艺术，都是我希望在职业生涯和个人生活中学习的榜样。

我由衷感谢学位论文委员会的各位成员：Soutir Bandyopadhyay 博士、Alfred W. Eustes III 博士、Yilin Fan 博士和 Jim Crompton 教授。通过他们富有成效的讨论和深刻的见解，我得以提升在研究领域的专业能力，也使这篇学位论文逐步成形。

我感激佩恩公共政策研究所的导师、同事和朋友们。尤其要感谢 Mikhail N. Zhizhin 博士、Christopher D. Elvidge 博士和 Morgan D. Bazilian 博士。能与这些遥感和卫星影像领域的世界级专家共事，对我而言是一次大开眼界的经历。我还特别想感谢 Zhizhin 博士给予我的帮助、分享的见解以及投入的时间。

我非常感谢 AMS 系的 Bandyopadhyay 博士和 Luis Tenorio 博士，他们教学出色，了解我本人，并激励我努力学习。回

顾我在机器学习方面的所学，正是这些知识让这篇学位论文得以完成；他们的课程无疑是对我最重要的帮助（请原谅我此刻还不是一名概率学家）。通过学习他们的统计方法课程，我开始真正领会是什么机器学习，并逐渐爱上数学，更具体地说，是概率论和统计建模。

在 Mines 的助教经历让我成为更优秀的博士生。我所学到的知识，无论是技术方面还是非技术方面，都融入了这篇学位论文。我要感谢 Crompton 教授、Eustes 博士、Mark G. Miller 博士、Linda A. Battalora 博士和 Miskimins 博士，为我提供了宝贵的助教机会。感谢所有学生给予我的支持和反馈。

我要感谢 Yu-Shu Wu 博士、Xiaolong Yin 博士和 Yilin Fan 博士，在我攻读博士学位期间给予我的关怀、支持和鼓励。还要感谢 Denise Winn-Bower、Rachel McDonald 和 JoeChen 对我的帮助。

我非常感谢 FAST 成员公司代表提供的反馈。许多讨论以及之后的思考都被纳入了这篇学位论文。特别要感谢 Ty Woodworth 在为我收集柱塞气举数据的过程中付出的时间和提供的帮助。每次我拜访他们位于科罗拉多州北部 Windsor 的办公室时，都受到了热情欢迎。Ty 友善地向我介绍了他所带领的团队，让我有机会向不同领域的多位现场专家提问和交流。这些讨论给了我极大的帮助。

特别感谢开源社区。在开展这项研究以及排版这篇学位论文的过程中，我从 Linux/GNU、TEX/LATEX 和 Python 的生态系统中受益匪浅。尤其要感谢 PyMC3 背后的开发者；这是一种概率编程语言，本学位论文在很大程度上依赖于它。

最后但同样重要的是，我要感谢我的家人和朋友。感谢我挚爱的妻子 Xiaodan，感谢她给予我的爱与支持，也感谢她烹制的美味佳肴。我还要感谢我的父母和岳父母，感谢他们的支持、鼓励与理解。

谨以此作品献给我的母亲倪灵英博士和我的父亲卢洪刚先生。

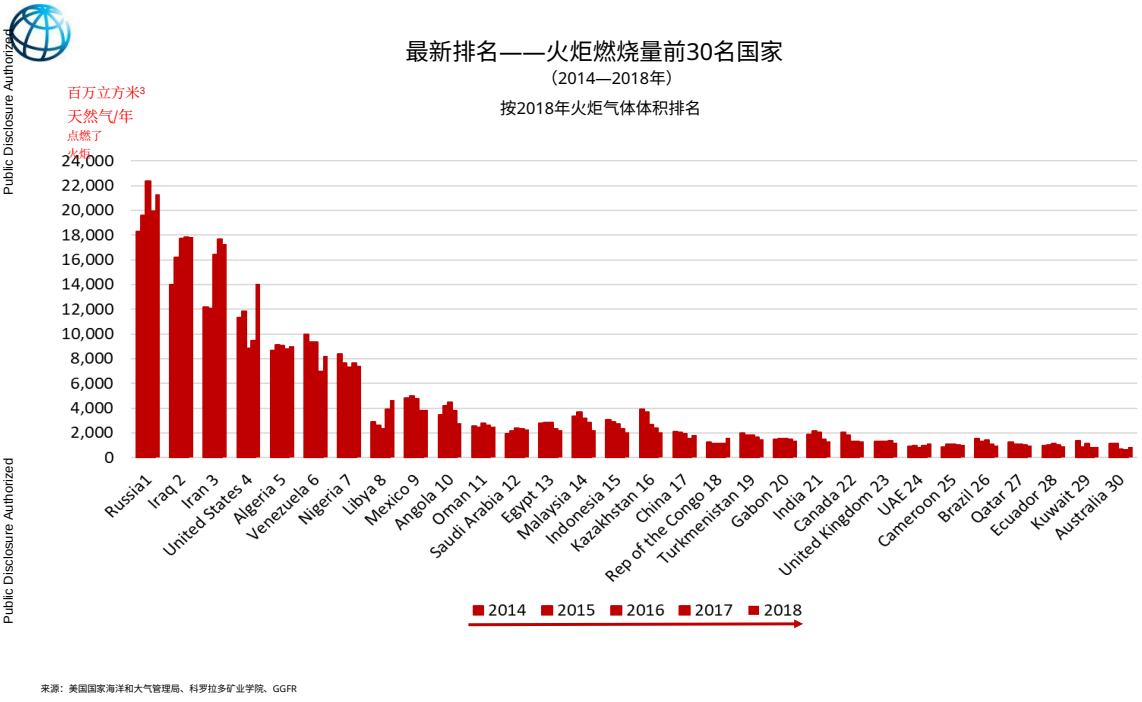
谁言寸草心，报得三春晖。

第1章

引言

目前，在石油工业中，对于同时生产原油和天然气的油井，运营商往往选择将伴生气进行火炬燃烧，而不是将其商品化。作出此类决定的原因多种多样。天然气价格的变化可能是一个重要因素，尤其是在加工和运输成本高于天然气价值时（Srivastava等，2019年）。在全国范围内，每年被火炬燃烧的气体数量巨大，而且火炬燃烧量排名靠前的国家呈现出不断增长的趋势（图1.1）。

Public Disclosure Authorized



Public Disclosure Authorized

来源：美国国家海洋和大气管理局、科罗拉多矿业学院、GGFR

图1.1：2018年火炬燃烧气体体积排名前30的国家。美国排名第4，且从2017年到2018年大幅增加（世界银行，2019年）。

由于近十年来非常规资源（例如页岩气藏）的开发热潮，美国一直是火炬燃烧总量最高的国家之一，

这一点得到美国能源信息署（EIA）2019年数据的佐证：自2013年以来，地下蕴藏巴肯组的北达科他州，以及拥有二叠纪盆地和鹰福特页岩的得克萨斯州，一直是火炬燃烧量最高的两个州。两州年度火炬燃烧体积时间序列如图1.2所示。从谷歌地球的影像中可以清晰识别出一些火炬燃烧点（图1.3）。

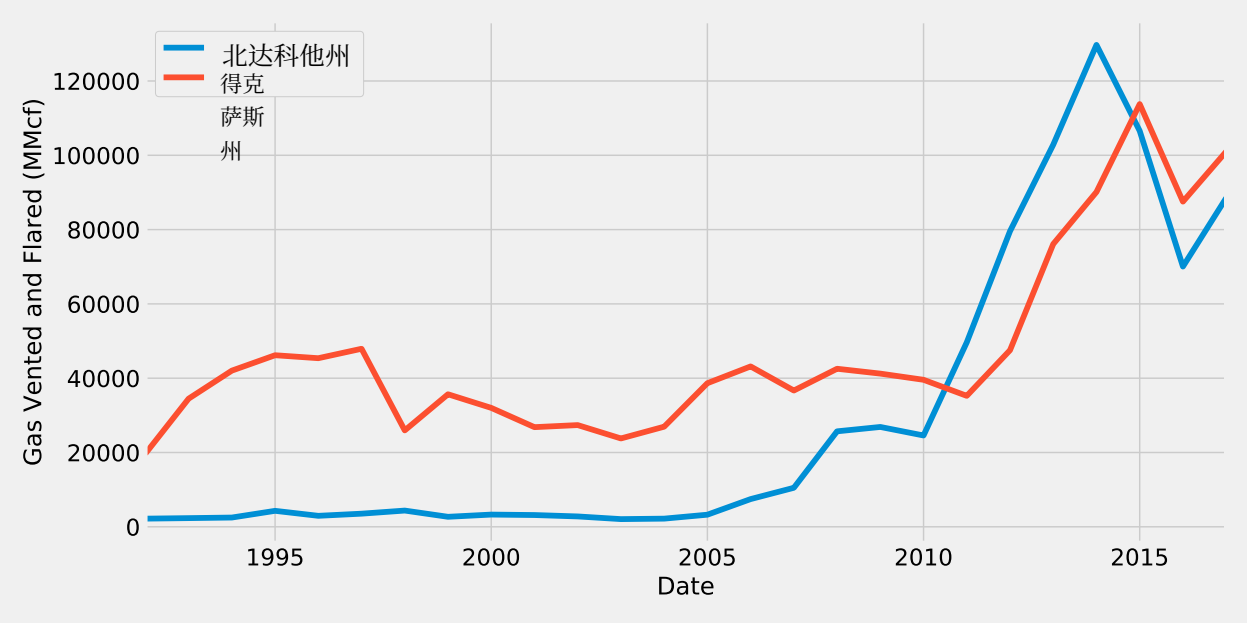


图 1.2：时间序列显示了美国天然气燃烧量排名前两位的州的趋势（美国能源信息署，2019年a）。得克萨斯州于2015年重新夺回领先地位。

天然气燃烧会造成能源浪费和 CO_2 排放。近年来，各类组织和政府机构一直倡导减少或消除常规天然气燃烧。例如，北达科他州工业委员会（NDIC）于2014年出台了天然气燃烧监管政策（第24665号命令），旨在从不同方面减少火炬燃烧（例如燃烧的天然气体积）。世界银行于2015年发起了“2030年前实现常规燃烧零排放”倡议。为了监测和评估火炬燃烧活动的规模，需要一种能够准确、精确地获取火炬燃烧定量信息的方法。然而，在某些情况下，这些信息只能通过自我报告机制获取。



图1.3：这幅谷歌地球影像展示了北达科他州某井位进行天然气燃烧的情况（谷歌地球，2019年）。

误差可能是有意或无意引入的 .

卫星遥感是解决这一问题的一种无偏方法。它有助于检测活跃火炬燃烧，尤其是在夜间，还可用于校准火炬燃烧气体体积的估计。本研究考虑了两种不同类型的传感器：陆地卫星8号（Landsat 8，L8）的陆地成像仪（OLI）和热红外传感器（TIRS），以及搭载于Suomi国家极轨伙伴卫星（NPP）和NOAA-20卫星上的可见光红外成像辐射计套件（VIIRS）。在本学位论文后文中，分别简称为L8和VIIRS。图1.4展示了利用VIIRS低光成像数据检测火炬燃烧的示例。

1.1 研究目标

本研究旨在实现以下目标：

- 评估利用卫星影像估算火炬燃烧气体体积的方法；并且，

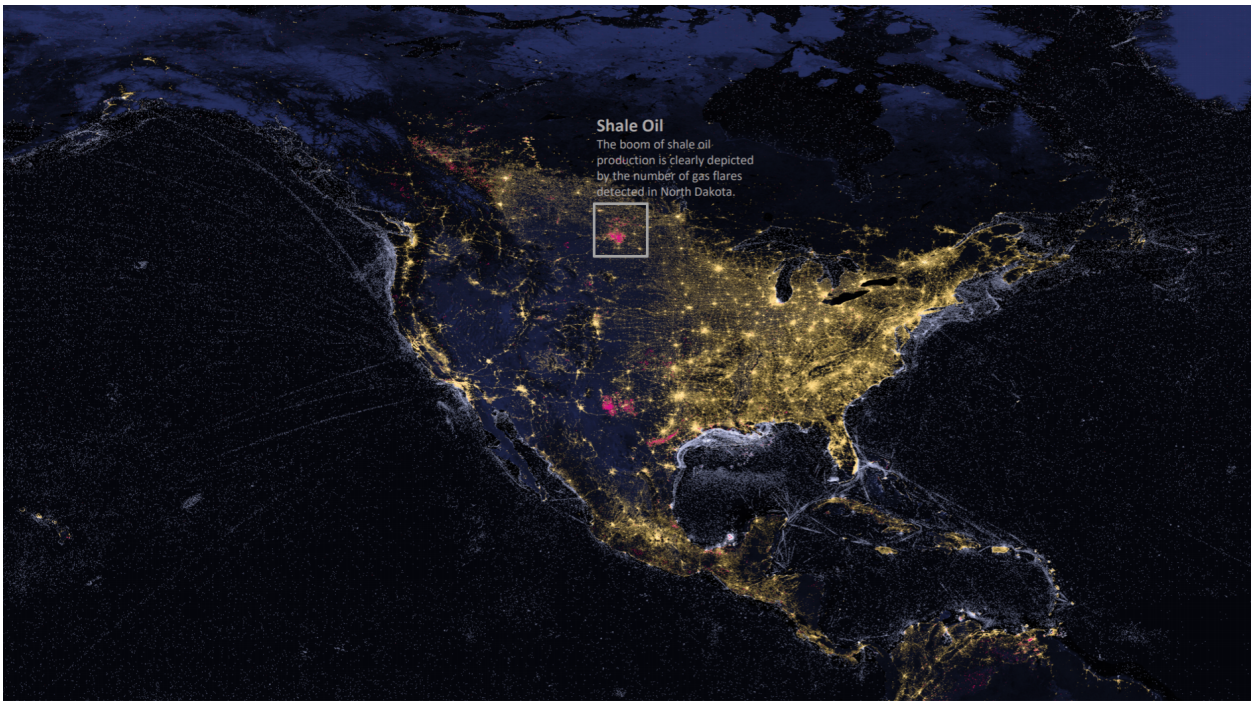


图1.4：原始海报的一部分（佩恩研究所地球观测组，2019年），该海报利用一年的VIIRS低光成像数据累积结果展示人类活动，例如天然气燃烧、捕鱼和城市灯光。如图中所注，北达科他州的燃烧活动在夜间从空间中清晰可见。

- 深入了解运营商的天然气燃烧行为。

1.2 学位论文目标

为实现第1.1节概述的目标，下面列出了更具体的目标：

1. 比较VIIRS和NDIC的燃烧数据，并分析其差异。

    - 比Compa 较VIIRS燃烧量与NDIC数据，并以NDIC作为基准 .

2. 评估使用陆地卫星8号夜间图像提升火炬检测和体积估算效果的有效性。

    - 确定陆地卫星8号的检测限，并将其与 VIIRS 的能力进行比较。

3. 调查运营商在天然气燃烧方面采用的方法。

- 确定天然气价格 / 油价 / 石油产量与火炬燃烧气体体积之间的相关性。

- 评估北达科他州的监管政策（第24665号命令）是否实现了其目标。
- 建立一个能够在州级预测火炬燃烧气体体积的模型。

4. 从所有生产实体中寻找任何隐藏的结构/簇。

1.3 概述与贡献

本学位论文的主要贡献在于证明，采用马尔可夫链蒙特卡罗方法实现的贝叶斯学习在燃烧数据分析中非常有效。针对不同的分析目标和数据粒度，本文开发了一系列参数化和非参数化机器学习模型，为今后的建模工作提供了直接指导。为证明这些模型的有效性和稳健性，本文均使用真实数据对其进行了测试。这种方法的优势在于，其推断阶段完全基于概率，因此能够恰当地刻画并量化可能模型带来的参数不确定性，以及含噪观测带来的随机不确定性。这使得提取出的洞见具有稳健性和可解释性，有助于决策和政策制定，例如州政府的决策与政策制定。

第二章回顾了卫星图像处理、贝叶斯推断、马尔可夫链蒙特卡罗方法和机器学习领域的最新进展。

第3章讨论了数据收集过程，并呈现了部分探索性数据分析结果。

第4章构建县级模型，以研究 VIIRS 与 NDIC 之间的相关性，并探究北达科他州各县之间的异质性。

第5章介绍天然气燃烧时间序列分析，旨在揭示不同层面的趋势和模式。

在第6章中，采用无监督学习对燃烧数据进行分析，以刻画潜在结构。

第7章介绍了一种运营商级监测与分析方法，并讨论了贝叶斯学习的应用。

第8章介绍了得出的主要结论，并提出了基于本研究的建议，同时概述了若干未来研究领域。

第2章

文献综述

20世纪90年代，世界银行开始收集夜间卫星图像，从中可以看到大城市和油田都很明亮，因此需要借助额外信息进行区分。2012年，随着VIIRS红外数据的问世，情况发生了变化（Rassenfoss 和 Zborowski，2018年）。其中一种数据产品VIIRS Nightfire（VNF）专门用于观测天然气燃烧，甚至能够区分生物质燃烧和天然气燃烧（Elvidge 等人，2017）。

VIIRS夜间火灾产品的开发基于VIIRS图像。为了提升火炬检测和气体体积估算的性能，可以利用其他信息来源，例如Landsat 8图像。表2.1比较了Landsat 8和VIIRS的空间分辨率与时间分辨率（美国国家航空航天局 2019；维基百科 2019）。图2.1展示了Landsat 8的空间分辨率。此外，Landsat 8会采集电磁波谱中11个不同光谱波段的数据，而VIIRS有22个波段。Landsat 8和VIIRS均运行在地球近极地轨道上，能够揭示地貌中的丰富特征。因此，与VIIRS的能力相比，Landsat 8应能识别更小的天然气火炬，尽管其较长的卫星重访时间给识别持续性较弱的火炬带来了挑战。第2.1节将进一步介绍VIIRS夜间火灾产品的处理步骤，其中的核心内容也将应用于Landsat 8。

表2.1：陆地卫星8号和VIIRS的分辨率

	分辨率类型	
	空间 [米]	时间 [天]
陆地卫星8号	15 至 100†	16.0‡
VIIRS	375 至 750†	0.5

†取决于电磁波谱的波段

‡ 适用于白天模式

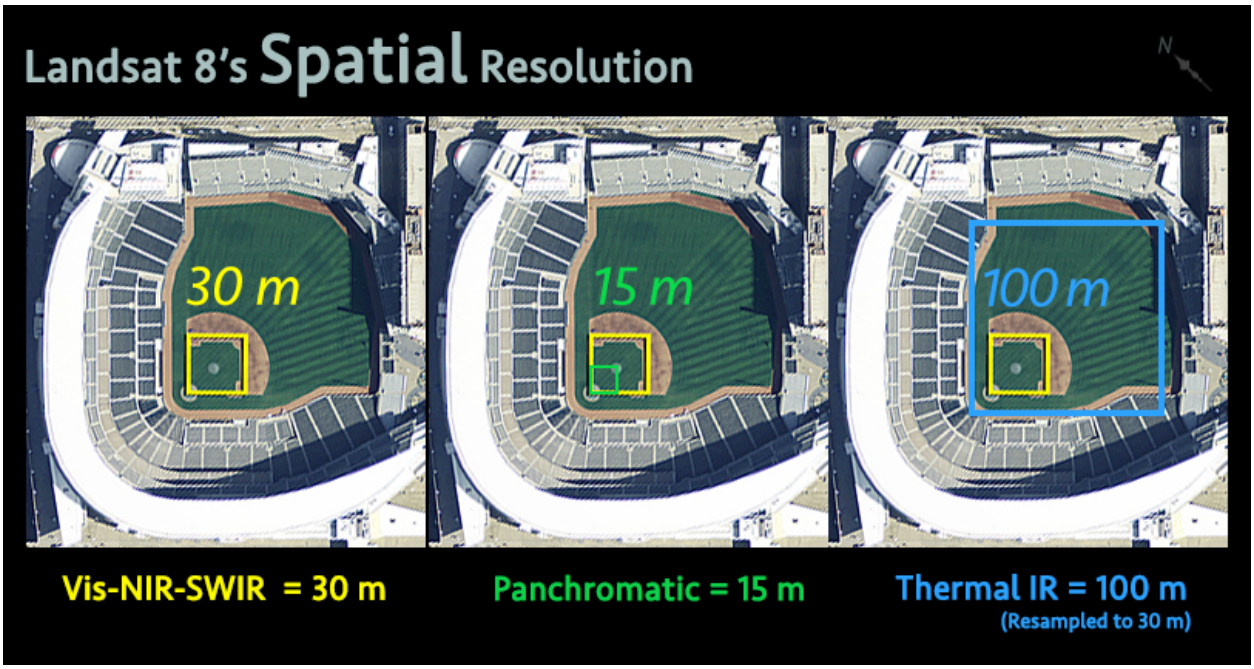


图2.1：陆地卫星8号的空间分辨率（美国国家航空航天局 2020年）。每个Landsat像元（30米×30米的区域）大致相当于一个棒球场的大小。

如今，一种极为丰富的资源就是数据。对于某个学科或研究领域而言，新的数据来源带来了新的信息维度，例如，卫星图像如今正用于天然气火炬燃烧分析。如何有效且智能地分析数据、从中获取洞见，是一个核心问题。以石油工程领域为例，已有数据驱动方法被提出，用于分析增产处理（Kazak- kov和Miskimins 2011）并预测砂堵（Yu等人2020）。机器学习是实现这一目标的有力工具。它是人工智能和数据科学的核心，处于统计学与计算机科学的交叉领域（Jordan和Mitchell 2015）。计算学习理论中的框架，例如Valiant（1984）提出的PAC学习，有助于为一些学习算法提供理论基础。

机器学习的一个子领域——深度学习（DL）于2006年首次亮相，当时 Hinton 和 Salakhutdinov 提出了深度信念网络（DBN）；但直到2012年，AlexNet 在 ImageNet竞赛中的分类准确率取得突破性表现后，深度学习才得到广泛认可（Krizhevsky等人，2012年）。AlexNet 是一种基于深度学习的模型（更

具体来说，它是一种卷积神经网络），错误率为15.3%，比亚军低10%以上。此后，深度学习一直主导着该竞赛；到2015年，基于深度学习的模型终于在分类数据集上超越了人类表现（He 等人，2015年）。

尽管基于神经网络的模型近年来取得了巨大成功，但需要注意的是，没有任何一种模型始终能够成为所有问题的最佳候选模型。Wolpert（1996年）对此进行了形式化证明，这通常被称为“无免费午餐”（NFL）定理。近期，Olson 等人（2017）对165个不同问题集上的13种分类算法进行了实证评估，结果与该定理一致：即使将表现最佳的五种算法объедин合起来，也无法支配所有问题集。

以下各节将对下述方面进行详细综述，这些方面构成了本研究的基础并为其提供了启发：

1. 卫星图像处理
2. 贝叶斯推断
3. 马尔可夫链蒙特卡罗
4. 机器学习
5. 分析工具集

2.1 卫星图像处理

利用卫星图像估算火炬燃烧气体体积。基于普朗克曲线拟合和物理定律的火灾检测算法称为 VIIRS Nightfire（VNF），由 Elvidge 等（2013）提出。本研究以此算法作为分析 L8 图像的起点。该方法包含几个主要步骤：

1. 热像元检测

   夜间，传感器主要记录仪器噪声，其大致符合高斯分布；少数包含红外线发射源的像元除外，

例如燃气火炬。因此，可以通过设定分布尾部的截断值来识别热像元，例如数字量化值高于均值加四个标准差的像元。

2. 噪声滤波

仅在一个光谱波段中检测到的热像元会被视为噪声并滤除。

3. 大气校正

散射和吸收效应导致的辐亮度损失可以进行校正。MOD- TRAN® 5 (Berk 等, 2006年)以大气水汽和温度廓线为参数，用于推导各光谱波段的校正系数。

4. 普朗克曲线拟合

针对燃气火炬建立普朗克曲线模型。由于它们是亚像元源，因此呈现为灰体。拟合的输出是对温度和发射缩放因子（普朗克函数中的发射率项）的估计。随后使用后者估算源面积。

5. 源面积的计算

源面积 S 的计算公式为

$$S = \varepsilon A\,, \tag{2.1}$$

其中， ε 为发射缩放因子， A 为像元覆盖范围的大小。

6. 辐射热的计算

辐射热根据斯特藩–玻尔兹曼定律计算：

$$\mathrm{RH} = \sigma T^4 S\,, \tag{2.2}$$

其中，RH 表示以 MW 为单位的辐射热， σ 表示斯特藩–玻尔兹曼常数， T 表示以 K 为单位的温度， S 表示以 m^2为单位的源面积。

获得辐射热（RH）后，Elvidge 等人（2015年）的先前研究建立了一种校准方法，用于估计火炬燃烧气体体积；该方法利用了 Cedigaz（2015年）提供的国家级火炬燃烧报告，以及得克萨斯州和北达科他州的州级报告。随后，可将该校准方法应用于全球各个火炬燃烧点，以估计火炬燃烧气体体积等。

2.2 贝叶斯推断

贝叶斯推断利用条件概率论，建立了一套从数据中学习的正式程序（Betancourt 2018年）。贝叶斯模型为可观测数据 $p(\mathcal{D},\ \boldsymbol{\theta})$ 以及不可观测的模型参数 $\mathcal{D}$ and $\boldsymbol{\theta}$ 提供完整的联合概率分布。贝叶斯分析的核心是获得后验分布 $p(\boldsymbol{\theta} \mid \mathcal{D})$，该分布刻画了给定某些数据 $\mathcal{D}$ 时参数 $\boldsymbol{\theta}$ 的条件概率。根据贝叶斯定理，可推导出：

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathcal{D})} \tag{2.3a}$$

$$= \frac{p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(\mathcal{D} \mid \boldsymbol{\theta}')\, p(\boldsymbol{\theta}')\, \mathrm{d}\boldsymbol{\theta}'} \tag{2.3b}$$

$$\propto p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\,, \tag{2.3c}$$

其中， $p(\mathcal{D} \mid \boldsymbol{\theta})$ 是似然（也称为观测模型），表示给定一组参数时数据出现的可能性； $p(\boldsymbol{\theta})$ 是先验，用于描述在观测到任何数据之前参数的概率。先验编码了领域专业知识。给定一些观测值后，先验会更新为后验分布，用以量化模型配置与领域知识和观测数据的一致程度（Betancourt 2018）。得到后验分布后，大多数乃至所有推断问题都可以通过某些函数的后验期望值来回答（Betancourt

2019年）：

$$\mathbb{E}_p[g(\boldsymbol{\theta})] = \int g(\boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\theta}\,, \tag{2.4}$$

其中，$g(\boldsymbol{\theta})$ 是对某个推断问题进行编码的函数（例如，后验分布在模型配置空间中的集中位置）。

预测可以用后验预测分布的形式给出：

$$p(y_* \mid \mathbf{x}_*, \mathcal{D}) = \int p(y_* \mid \boldsymbol{\theta}, \mathbf{x}_*)\, p(\boldsymbol{\theta} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\theta}\,, \tag{2.5}$$

其中，y_* 是基于训练集 $\mathcal{D}$ 对测试输入 $\mathbf{x}_*$ 所作的预测。本质上，这是对参数的后验分布进行积分，得到预测 $p(y_* \mid \boldsymbol{\theta}, \mathbf{x}_*)$（Rasmussen 和 Williams，2006年）。请注意，与仅通过最大似然估计或最大后验估计给出点估计相比（一些机器学习模型在频率学派框架下会如此处理），以概率分布的形式给出最终结果能够获取更丰富的信息，并得出更可靠的推断。这是通过将后验参数估计中的不确定性纳入推断过程来实现的。其他益处还包括后验预测检验，即检查生成数据（$\mathbf{y}_*$）与观测数据（$\mathbf{y}$）之间是否具有自洽性。

2.3 马尔可夫链蒙特卡罗

贝叶斯统计中的许多核心积分问题，包括公式 2.4 和 2.5 中的问题，在解析上都难以处理。一类称为马尔可夫链蒙特卡罗（MCMC）的采样算法可用于近似求解这些问题（Andrieu 等人，2003年）。假设对于某个感兴趣的函数 $f(x)$，目标是计算其相对于一个非常规目标分布 $p(x)$ 的积分，而无法直接从该分布中抽取样本：

$$I(f) = \int f(x)\, p(x)\, \mathrm{d}x\,. \tag{2.6}$$

通过构造以 $p(x)$ 为不变分布的马尔可夫链，MCMC 采样器在遍历样本空间 $\mathcal{X}$ 的同时，能够生成 $x^{(i)}$，其特征类似于样本

直接从目标分布 $p(x)$ 中抽取的样本。换言之，这种机制使得从 $p(x)$ 中抽取一组样本 $\{x^{(i)}\}_{i=1}^{N}$ 成为可能。

然后，根据蒙特卡罗原理，积分 $I(f)$ 可以用总和$I_N(f)$近似：

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}) \xrightarrow[N\to\infty]{\text{a.s.}} I(f) = \int f(x)\, p(x)\, \mathrm{d}x\,. \tag{2.7}$$

也就是说，估计值 $I_N(f)$ 是无偏的，并且根据强大数定律，它几乎必然（a.s.）收敛到 $I(f)$。因此，马尔可夫链蒙特卡罗是贝叶斯分析中的一种强大工具。在实践中，Metropolis-Hastings（MH）算法和吉布斯抽样一直是广受欢迎的马尔可夫链蒙特卡罗方法（Andrieu 等人 2003年），但前提是参数空间的维度不能太高（McElreath 2020年）。

由于计算资源有限，不可能无限期地运行马尔可夫链。换言之，推断必须基于有限次抽样得出。一种方法（本研究有效地采用了这种方法）是并行运行多条链，并监测各种统计量以诊断不收敛情况。除了马尔可夫链每次转移的有效样本量外，本学位论文还使用 Gelman-Rubin 统计量（Gelman 和 Rubin 1992年），记作 R。R 统计量用于衡量从参数空间中分散的点初始化的马尔可夫链集合最终是否收敛到相同的平衡态（Betancourt 2017）。当 R 足够接近 1（例如 $\hat{R} < 1.05$）时，即认为已实现收敛。例如，图 2.2 展示了四条链如何从不同的角落出发，并在经过一定次数的迭代后趋于平稳并实现收敛。

在许多实际问题中，包括本学位论文中的模型，参数空间都具有很高的维度，且包含高度弯曲的区域。在这种情况下，Metropolis-Hastings 算法和吉布斯抽样的效率远远不够。哈密顿蒙特卡洛（HMC）最初由 Duane 等人（1987）提出，在这类问题中表现尤为出色，是本学位论文采用的主要抽样策略。

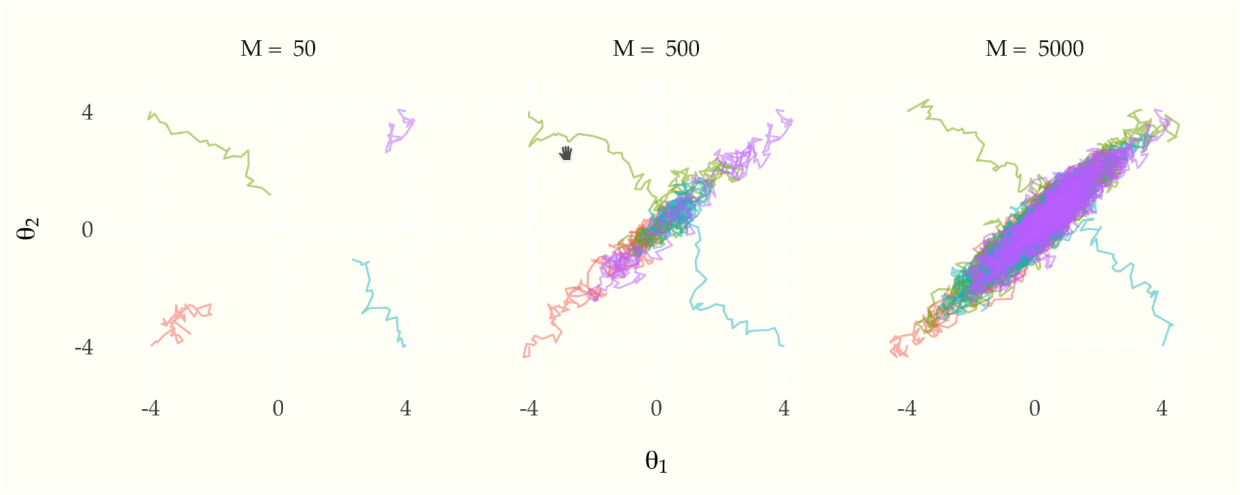


图 2.2：四条随机游走 Metropolis 马尔可夫链的演化过程（Carpenter 2020），每条链都从不同位置开始。目标密度是一个方差为 1、相关性为 0.9 的二元正态分布。经过 $M = 5000$ 次迭代后，四条链已充分混合，并探索了目标密度的大部分区域。

具体而言，本文采用 Hoffman 和 Gelman（2014）提出的无转向采样器（NUTS）从后验分布中进行抽样。NUTS 是 HMC 的一种扩展。

2.4 机器学习

Mitchell（1997）将机器学习定义为计算机通过经验自动改进的过程。它也可以被视为一个函数估计问题（Vapnik 2000），或从数据中提取重要模式和趋势的过程（Hastie 等人 2009）。

从任务类型来看，常见的学习类型包括监督式、无监督式、半监督式和强化学习（Burkov 2019）。令 $\mathbf{x}_i \in \mathcal{X} \subseteq \mathrm{R}^d$ 表示输入，$y_i \in \mathcal{Y}$表示目标，那么前两种类型的目标是：

- 监督学习旨在利用由 $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ 和 $\mathbf{y} = \{y_i\}_{i=1}^n$ 组成的数据集，构建一个模型，使其能够根据一些新的／未见过的输入（$\mathbf{x}j$）预测输出（y_j），即学习底层映射 f: $\mathcal{X} \to \mathcal{Y}$。

- 无监督学习用于在$\mathbf{X}$中寻找隐藏模式；在这种情况下，不存在任何标签（$\mathbf{y}$）或预先定义的目标。

另一种学习方式是在线学习，在这种方式中，训练数据会持续输入算法，或者一次输入一个样本（Abu-Mostafa 等，2012年）。换句话说，算法可以获取流数据，并需要随时对其进行处理。这与批量学习不同，批量学习会事先提供数据，并在学习过程中将其“冻结”。在线学习可应用于上文讨论的不同任务（监督式学习及其他任务）。

从模型特征来看，机器学习模型可分为参数模型和非参数模型。参数模型的特点是参数数量固定，而非参数模型则具有无限维参数空间。例如，在回归情境中，后者的参数空间可以是连续函数的集合（Orbanz and Teh 2010）。本学位论文运用监督学习和无监督学习，同时采用参数模型和非参数模型。

从贝叶斯的角度来看，机器学习本质上是在计算后验分布（de Freitas 2013），并以此进行推断和预测任务。这一过程通过方程 2.3a 实现。在实践中，基于贝叶斯框架开展机器学习遵循一套严谨的工作流程（图 2.3），本学位论文的建模工作也采用了这一流程。

2.5 分析工具集

在过去五到十年间，开源社区蓬勃发展，贡献不断涌现，取得了显著进步。围绕 Python、R 和 Julia 构建的生态系统已经过原型开发、测试，并在各行业的生产环境中部署。强大的概率编程语言（PPL），例如 Stan（Carpenter 等人 2017）和 PyMC3（Salvatier 等人 2016），已成为贝叶斯机器学习的主力工具。

本研究的大部分工作均使用 Python 实现。具体而言，贝叶斯学习借助 PyMC3 完成。一些分析性可视化图表使用 ArviZ（Kumar 等，2019）生成。地理空间运算借助 GeoPan-

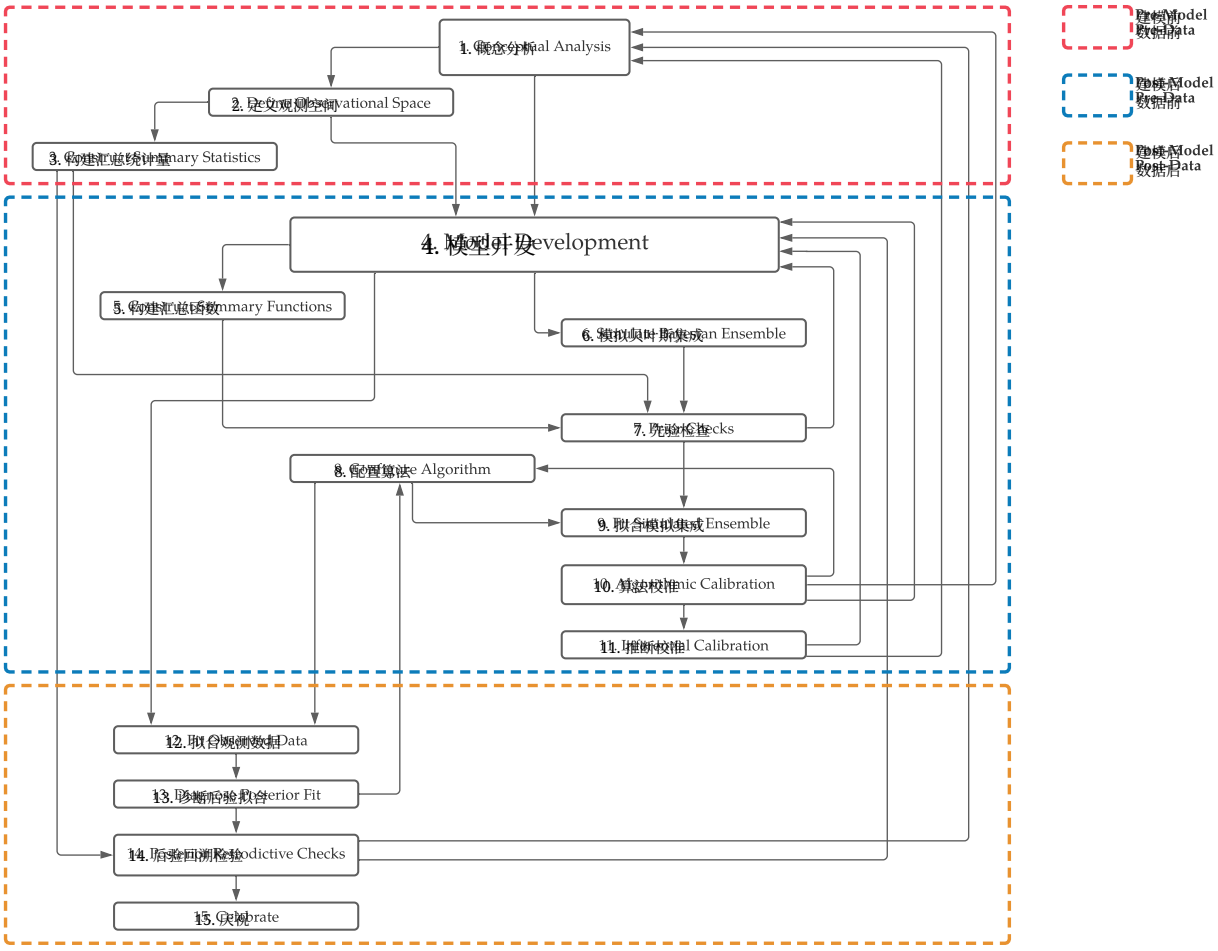


图 2.3：改编自（Betancourt 2020）的流程图展示了一套有原则的贝叶斯工作流程。

das（Jordahl 等，2020）完成。卫星影像在 MATLAB 中处理和分析，实现方法主要遵循 Elvidge 等（2013）的做法。

第3章

数据预处理与探索性数据分析

本章概述了燃烧数据，并考察了其他一些可能与燃烧统计数据相关的变量。通过探索性数据分析，选出作为本学位论文研究重点的变量子集。最后建立了一个州级模型，为接下来两章的研究工作提供依据。

3.1 数据收集

本研究分析所用的数据来自四个来源：Landsat 8卫星图像、VIIRS估算的燃烧量、NDIC月度生产报告，以及北达科他州的县/油田形状文件。

3.1.1 陆地卫星8号图像

总共从 Google Cloud 下载了 167 幅图像（自2013年起），所依据的标准如下：

- 来自五个路径/行号：126/216、126/217、126/218、127/216 和 127/21　7.

  根据全球参考系统（WRS），可以使用路径号和行号查询世界任何区域的卫星影像。这五个路径/行号覆盖了北达科他州大部分从事生产和火炬燃烧活动的区域。

- 夜间图像。

  火炬检测仅使用陆地卫星8号的夜间影像。

- 云覆盖率低于 10 %。

  低云量百分比较低的图像能更清晰地呈现地表特征，包括天然气火炬，因此非常适合用于验证所开发的方法。

- GeoTIFF数据产品。

  所有光谱波段的地理配准信息和原始图像均通过 GeoTIFF格式保存，这些信息对于分析至关重要。

3.1.2 VIIRS估算体积

本学位论文使用了 Mikhail N. Zhizhin（个人通信）提供的 VIIRS火炬燃烧清单和估算体积数据集。该数据集包含北美洲从2012年3月至2018年12月（含首尾月份）的月度火炬检测记录，以及相关的：

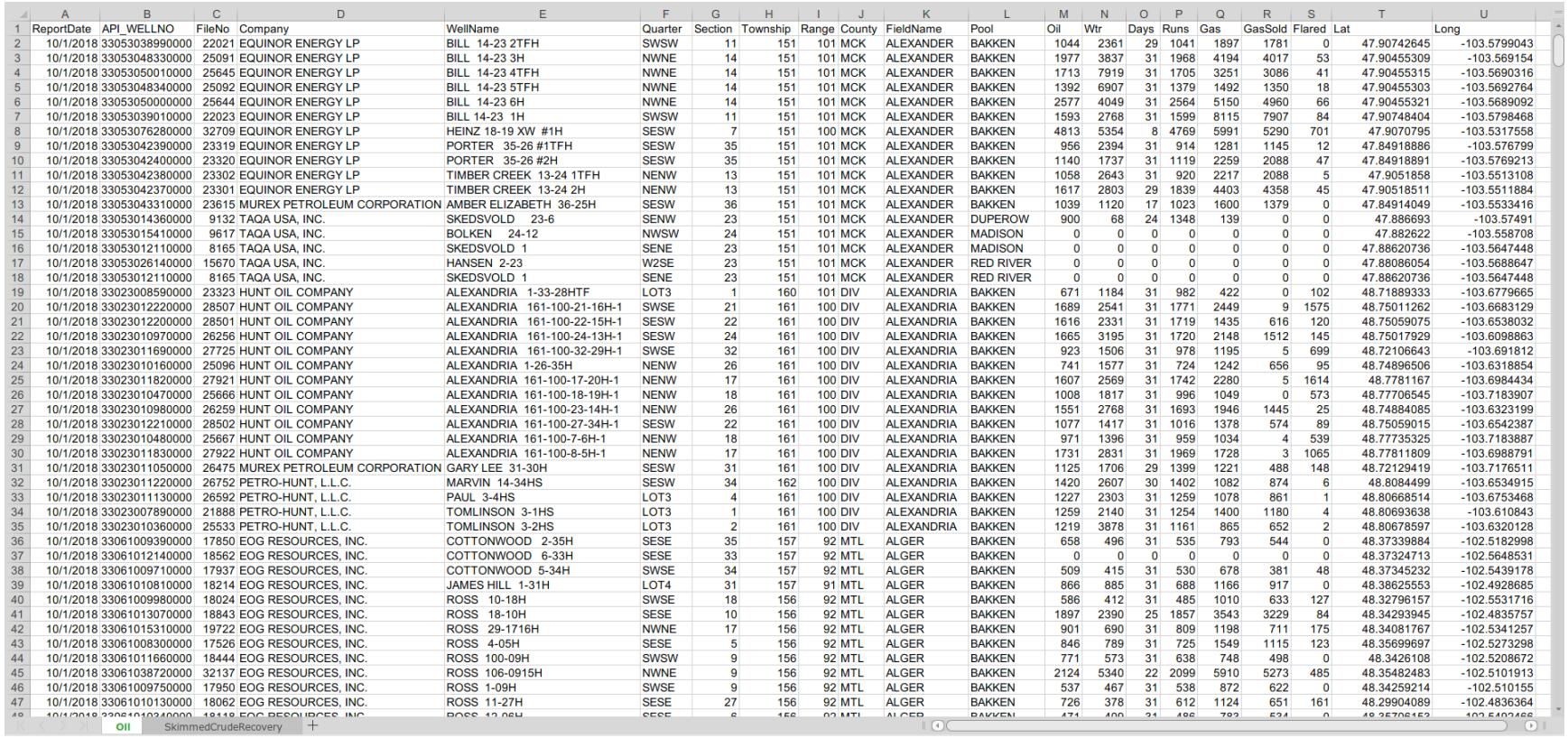
- 标明具体月份的时间戳
- WGS 84坐标中的纬度和经度
- 以十亿立方米为单位的火炬燃烧量估算值

3.1.3 NDIC月度生产报告

已从 NDIC 下载了 2015年5月至2020年4月（含首尾月份）所有包含火炬燃烧信息的月度生产报告。每个月份对应一个 Excel电子表格；每一行对应一口油井（该油井在该月份处于活跃状态），各列则包含不同类型的信息，包括火炬燃烧气体体积（由运营商估算并报告）、油田、石油产量等。图 3.1 展示了其中一个电子表格顶部 ~ 50 行的截图。

3.1.4 NDIC 形状文件

北达科他州各县和油田的形状文件从 NDIC GIS 地图服务器下载。所有多边形均以 NAD 27坐标表示。这些形状文件用于对卫星检测位置进行逆地理编码，以获取可读地址，具体而言，即确定火炬位于哪个县和油田。



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	ReportDate	API_WELLNO	FileNo	Company	WellName	Quarter	Section	Township	Range	County	FieldName	Pool	Oil	Wtr	Days	Runs	Gas	GasSold	Flared	Lat	Long
2	10/1/2018	33053038990000	22021	EQUINOR ENERGY LP	BILL 14-23 2TFH	SWSW	11	151	101	MCK	ALEXANDER	BAKKEN	1044	2361	29	1041	1897	1781	0	47.90742645	-103.5799043
3	10/1/2018	33053048330000	25091	EQUINOR ENERGY LP	BILL 14-23 3H	NWNE	14	151	101	MCK	ALEXANDER	BAKKEN	1977	3837	31	1968	4194	4017	53	47.90455309	-103.569154
4	10/1/2018	33053050010000	25645	EQUINOR ENERGY LP	BILL 14-23 4TFH	NWNE	14	151	101	MCK	ALEXANDER	BAKKEN	1713	7919	31	1705	3251	3086	41	47.90455315	-103.5690316
5	10/1/2018	33053048340000	25092	EQUINOR ENERGY LP	BILL 14-23 5TFH	NWNE	14	151	101	MCK	ALEXANDER	BAKKEN	1392	6907	31	1379	1492	1350	18	47.90455303	-103.5692764
6	10/1/2018	33053050000000	25644	EQUINOR ENERGY LP	BILL 14-23 6H	NWNE	14	151	101	MCK	ALEXANDER	BAKKEN	2577	4049	31	2564	5150	4960	66	47.90455321	-103.5689092
7	10/1/2018	33053039010000	22023	EQUINOR ENERGY LP	BILL 14-23 1H	SWSW	11	151	101	MCK	ALEXANDER	BAKKEN	1593	2768	31	1599	8115	7907	84	47.90748404	-103.5798468
8	10/1/2018	33053076280000	32709	EQUINOR ENERGY LP	HEINZ 18-19 XW #1H	SESW	7	151	100	MCK	ALEXANDER	BAKKEN	4813	5354	8	4769	5991	5290	701	47.9070795	-103.5317558
9	10/1/2018	33053042390000	23319	EQUINOR ENERGY LP	PORTER 35-26 #1TFH	SESW	35	151	101	MCK	ALEXANDER	BAKKEN	956	2394	31	914	1281	1145	12	47.84918886	-103.576799
10	10/1/2018	33053042400000	23320	EQUINOR ENERGY LP	PORTER 35-26 #2H	SESW	35	151	101	MCK	ALEXANDER	BAKKEN	1140	1737	31	1119	2259	2088	47	47.84918891	-103.5769213
11	10/1/2018	33053042380000	23302	EQUINOR ENERGY LP	TIMBER CREEK 13-24 1TFH	NENW	13	151	101	MCK	ALEXANDER	BAKKEN	1058	2643	31	920	2217	2088	5	47.9051858	-103.5513108
12	10/1/2018	33053042370000	23301	EQUINOR ENERGY LP	TIMBER CREEK 13-24 2H	NENW	13	151	101	MCK	ALEXANDER	BAKKEN	1617	2803	29	1839	4403	4358	45	47.90518511	-103.5511884
13	10/1/2018	33053043310000	23615	MUREX PETROLEUM CORPORATION	AMBER ELIZABETH 36-25H	SESW	36	151	101	MCK	ALEXANDER	BAKKEN	1039	1120	17	1023	1600	1379	0	47.84914049	-103.5533416
14	10/1/2018	33053014360000	9132	TAQA USA, INC.	SKEDSVOLD 23-6	SENW	23	151	101	MCK	ALEXANDER	DUPEROW	900	68	24	1348	139	0	0	47.886693	-103.57491
15	10/1/2018	33053015410000	9617	TAQA USA, INC.	BOLKEN 24-12	NWSW	24	151	101	MCK	ALEXANDER	MADISON	0	0	0	0	0	0	0	47.882622	-103.558708
16	10/1/2018	33053012110000	8165	TAQA USA, INC.	SKEDSVOLD 1	SENE	23	151	101	MCK	ALEXANDER	MADISON	0	0	0	0	0	0	0	47.88620736	-103.5647448
17	10/1/2018	33053026140000	15670	TAQA USA, INC.	HANSEN 2-23	W2SE	23	151	101	MCK	ALEXANDER	RED RIVER	0	0	0	0	0	0	0	47.88086054	-103.5688647
18	10/1/2018	33053012110000	8165	TAQA USA, INC.	SKEDSVOLD 1	SENE	23	151	101	MCK	ALEXANDER	RED RIVER	0	0	0	0	0	0	0	47.88620736	-103.5647448
19	10/1/2018	33023008590000	23323	HUNT OIL COMPANY	ALEXANDRIA 1-33-28HTF	LOT3	1	160	101	DIV	ALEXANDRIA	BAKKEN	671	1184	31	982	422	0	102	48.71889333	-103.6779665
20	10/1/2018	33023012220000	28507	HUNT OIL COMPANY	ALEXANDRIA 161-100-21-16H-1	SWSE	21	161	100	DIV	ALEXANDRIA	BAKKEN	1689	2541	31	1771	2449	9	1575	48.75011262	-103.6683129
21	10/1/2018	33023012200000	28501	HUNT OIL COMPANY	ALEXANDRIA 161-100-22-15H-1	SESW	22	161	100	DIV	ALEXANDRIA	BAKKEN	1616	2331	31	1719	1435	616	120	48.75059075	-103.6538032
22	10/1/2018	33023010970000	26256	HUNT OIL COMPANY	ALEXANDRIA 161-100-24-13H-1	SESW	24	161	100	DIV	ALEXANDRIA	BAKKEN	1665	3195	31	1720	2148	1512	145	48.75017929	-103.6098863
23	10/1/2018	33023011690000	27725	HUNT OIL COMPANY	ALEXANDRIA 161-100-32-29H-1	SWSE	32	161	100	DIV	ALEXANDRIA	BAKKEN	923	1506	31	978	1195	5	699	48.72106643	-103.691812
24	10/1/2018	33023010160000	25096	HUNT OIL COMPANY	ALEXANDRIA 1-26-35H	NENW	26	161	100	DIV	ALEXANDRIA	BAKKEN	741	1577	31	724	1242	656	95	48.74896506	-103.6318854
25	10/1/2018	33023011820000	27921	HUNT OIL COMPANY	ALEXANDRIA 161-100-17-20H-1	NENW	17	161	100	DIV	ALEXANDRIA	BAKKEN	1607	2569	31	1742	2280	5	1614	48.7781167	-103.6984434
26	10/1/2018	33023010470000	25666	HUNT OIL COMPANY	ALEXANDRIA 161-100-18-19H-1	NENW	18	161	100	DIV	ALEXANDRIA	BAKKEN	1008	1817	31	996	1049	0	573	48.77706545	-103.7183907
27	10/1/2018	33023010980000	26259	HUNT OIL COMPANY	ALEXANDRIA 161-100-23-14H-1	NENW	26	161	100	DIV	ALEXANDRIA	BAKKEN	1551	2768	31	1693	1946	1445	25	48.74884085	-103.6323199
28	10/1/2018	33023012210000	28502	HUNT OIL COMPANY	ALEXANDRIA 161-100-27-34H-1	SESW	22	161	100	DIV	ALEXANDRIA	BAKKEN	1077	1417	31	1016	1378	574	89	48.75059015	-103.6542387
29	10/1/2018	33023010480000	25667	HUNT OIL COMPANY	ALEXANDRIA 161-100-7-6H-1	NENW	18	161	100	DIV	ALEXANDRIA	BAKKEN	971	1396	31	959	1034	4	539	48.77735325	-103.7183887
30	10/1/2018	33023011830000	27922	HUNT OIL COMPANY	ALEXANDRIA 161-100-8-5H-1	NENW	17	161	100	DIV	ALEXANDRIA	BAKKEN	1731	2831	31	1969	1728	3	1065	48.77811809	-103.6988791
31	10/1/2018	33023011050000	26475	MUREX PETROLEUM CORPORATION	GARY LEE 31-30H	SESW	31	161	100	DIV	ALEXANDRIA	BAKKEN	1125	1706	29	1399	1221	488	148	48.72129419	-103.7176511
32	10/1/2018	33023011220000	26752	PETRO-HUNT, L.L.C.	MARVIN 14-34HS	SESW	34	162	100	DIV	ALEXANDRIA	BAKKEN	1420	2607	30	1402	1082	874	6	48.8084499	-103.6534915
33	10/1/2018	33023011130000	26592	PETRO-HUNT, L.L.C.	PAUL 3-4HS	LOT3	4	161	100	DIV	ALEXANDRIA	BAKKEN	1227	2303	31	1259	1078	861	1	48.80668514	-103.6753468
34	10/1/2018	33023007890000	21888	PETRO-HUNT, L.L.C.	TOMLINSON 3-1HS	LOT3	1	161	100	DIV	ALEXANDRIA	BAKKEN	1259	2140	31	1254	1400	1180	4	48.80693638	-103.610843
35	10/1/2018	33023010360000	25533	PETRO-HUNT, L.L.C.	TOMLINSON 3-2HS	LOT3	2	161	100	DIV	ALEXANDRIA	BAKKEN	1219	3878	31	1161	865	652	2	48.80678597	-103.6320128
36	10/1/2018	33061009390000	17850	EOG RESOURCES, INC.	COTTONWOOD 2-35H	SESE	35	157	92	MTL	ALGER	BAKKEN	658	496	31	535	793	544	0	48.37339884	-102.5182998
37	10/1/2018	33061012140000	18562	EOG RESOURCES, INC.	COTTONWOOD 6-33H	SESE	33	157	92	MTL	ALGER	BAKKEN	0	0	0	0	0	0	0	48.37324713	-102.5648531
38	10/1/2018	33061009710000	17937	EOG RESOURCES, INC.	COTTONWOOD 5-34H	SWSE	34	157	92	MTL	ALGER	BAKKEN	509	415	31	530	678	381	48	48.37345232	-102.5439178
39	10/1/2018	33061010810000	18214	EOG RESOURCES, INC.	JAMES HILL 1-31H	LOT4	31	157	91	MTL	ALGER	BAKKEN	866	885	31	688	1166	917	0	48.38625553	-102.4928685
40	10/1/2018	33061009980000	18024	EOG RESOURCES, INC.	ROSS 10-18H	SWSE	18	156	92	MTL	ALGER	BAKKEN	586	412	31	485	1010	633	127	48.32796157	-102.5531716
41	10/1/2018	33061013070000	18843	EOG RESOURCES, INC.	ROSS 18-10H	SESE	10	156	92	MTL	ALGER	BAKKEN	1897	2390	25	1857	3543	3229	84	48.34293945	-102.4835757
42	10/1/2018	33061015310000	19722	EOG RESOURCES, INC.	ROSS 29-1716H	NWNE	17	156	92	MTL	ALGER	BAKKEN	901	690	31	809	1198	711	175	48.34081767	-102.5341257
43	10/1/2018	33061008300000	17526	EOG RESOURCES, INC.	ROSS 4-05H	SESE	5	156	92	MTL	ALGER	BAKKEN	846	789	31	725	1549	1115	123	48.35699697	-102.5273298
44	10/1/2018	33061011660000	18444	EOG RESOURCES, INC.	ROSS 100-09H	SWSW	9	156	92	MTL	ALGER	BAKKEN	771	573	31	638	748	498	0	48.3426108	-102.5208672
45	10/1/2018	33061038720000	32137	EOG RESOURCES, INC.	ROSS 106-0915H	NWNE	9	156	92	MTL	ALGER	BAKKEN	2124	5340	22	2099	5910	5273	485	48.35482483	-102.5101913
46	10/1/2018	33061009750000	17950	EOG RESOURCES, INC.	ROSS 1-09H	SWSE	9	156	92	MTL	ALGER	BAKKEN	537	467	31	538	872	622	0	48.34259214	-102.510155
47	10/1/2018	33061010130000	18062	EOG RESOURCES, INC.	ROSS 11-27H	SESE	27	156	92	MTL	ALGER	BAKKEN	726	378	31	612	1124	651	161	48.29904089	-102.4836364
48	10/1/2018	33061010340000	18118	EOG RESOURCES, INC.	ROSS 12-06H	SESE	6	156	92	MTL	ALGER	BAKKEN	471	409	31	486	783	534	0	48.35706153	-102.5402466

Oil SkimmedCrudeRecovery +

Figure 3.1: A screenshot of the top ∼50 rows in the October 2018 production report. Each row corresponds to a well. There are in total 17,135 rows in this spreadsheet, with the first row being the header.

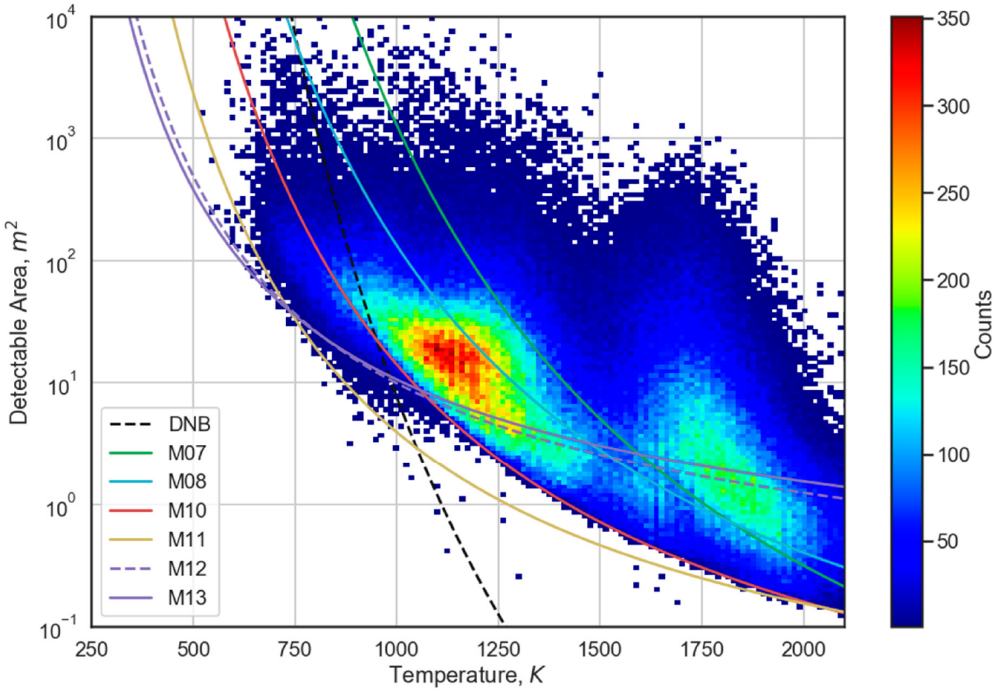
3.2 卫星图像处理

如第 3.1.1 节所述，所有可用的 L8 图像均已下载。按照第 2.1 节概述的工作流程，对这些图像进行批量处理。为与 VIIRS 的性能进行比较，特别是比较夜间燃烧源探测限，我们汇集了从所有 L8 图像中检测到的全部火炬，并据此生成图 3.2 所示的源面积与温度散点图。

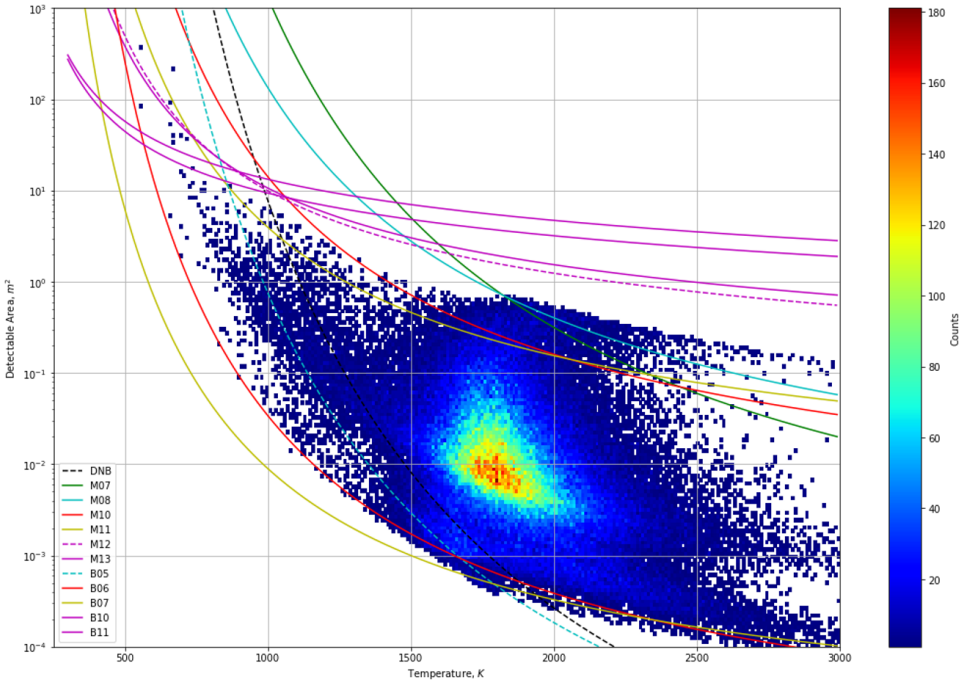
虽然预计 L8 能够检测到比 VIIRS 更小的火炬（VIIRS 可检测到面积大约相当于整个炉灶的火炬），但散点图显示，大多数检测结果对于天然气燃烧而言都太小。为确认某些热像元是否聚集在一起，实际上代表单个火炬或火炬燃烧点，我们使用 McInnes 等人（2017）提供的实现，在每幅 L8 检测图上运行 HDBSCAN（Campello 等人，2013），以查明是否存在由热像元构成的大型斑块。HDBSCAN 是一种基于密度的聚类算法，保留了原始 DBSCAN（Ester 等人，1996）的所有优点，例如能够发现任意形状的簇。它还优于 DBSCAN，因为它能够构建密度各异的簇（Burkov，2019）。此外，为了在此情况下获得尽可能准确的结果，我们选择哈弗辛度量来处理热像元之间的大圆距离；使用叶节点聚类而非默认的超额质量法，以生成更精细的簇。聚类结果如图 3.3 所示。

为验证这些簇究竟是单个火炬，还是实际上由大量邻井组成（这种情况下，每个热像元仍代表一个独立的火炬），我们进一步查看每张检测图（KMZ文件）进行追踪。结果发现，一些由热像元组成的大型斑块聚集在一起，确实代表单个（非常大的）火炬。图 3.4展示了其中一个例子。这给需要准确估算放空燃烧次数的情况带来了挑战。

造成这种处理伪影的原因是，对于大型火炬，其周围存在辉光，而这些辉光被当作许多独立的燃烧源处理。对此有一些可行的方法



(a) VIIRS 性能（Elvidge 等人，2019年）



(b) Landsat 8 性能；图由 Mikhail N. Zhizhin 提供（个人交流）

图3.2：VIIRS（上）和 Landsat 8（下）的夜间燃烧源探测限。对于温度通常高于 1500 开尔文的天然气燃烧，Landsat 8 检测到的火炬源面积（约为 10^{-2}m^2）比 VIIRS 检测到的（约为 1 m^2）小几个数量级。

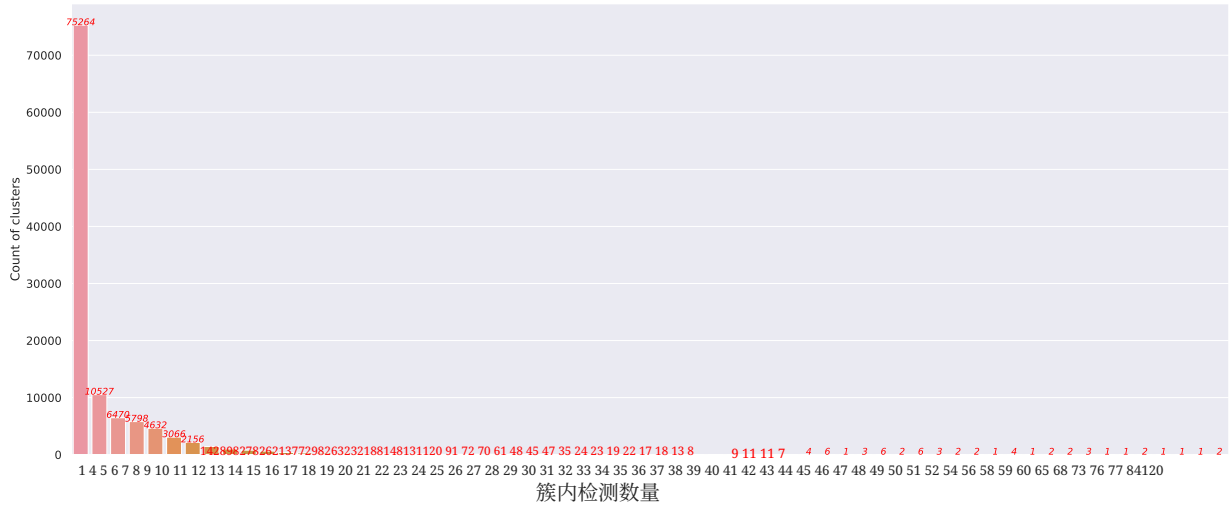
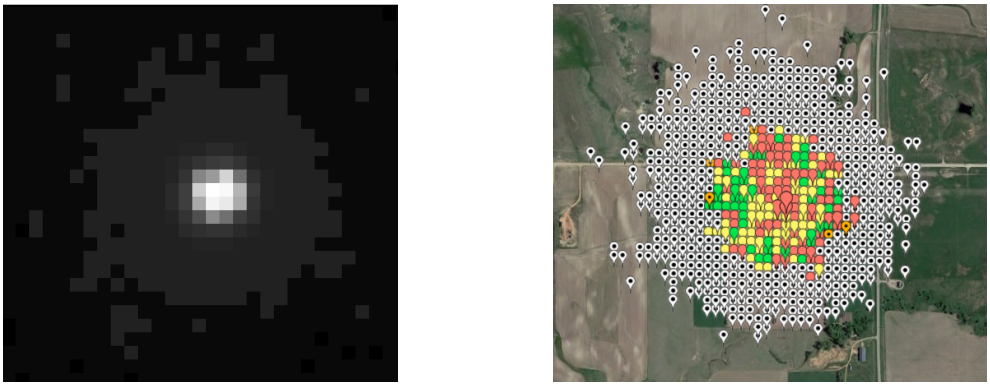


图3.3：展示簇大小分布的计数图：显然存在一定数量的大簇（如右侧尾部所示）。例如，有2个簇各包含120个热像元，另有一个簇包含84个热像元。



(a) 波段6（短波红外）　　(b) KMZ 视图

图3.4：一个由许多热像元（检测结果）组成的大型火炬，是通过对 L8 图像运行 Nightfire算法发现的。图中展示了波段6（灰度图像）和 KMZ 视图，由 Christopher D. Elvidge 提供（个人通信）。

为减轻这一问题，从而使对 Landsat 8 的解读和估计更加准确。在本研究中，通过 VIIRS 检测到的火炬以及根据这些火炬估计出的气体体积，是分析的重点。

3.3 逆地理编码

通过逆地理编码，可以获取北达科他州境内每个 VIIRS 火炬所在的县信息。对于大多数火炬，还可以获取其油田信息。此后，可以针对某个特定地点或时间周期，在不同层级上对 VIIRS 和 NDIC 的燃烧统计数据进行比较和对照。

使用第 3.1.4 节中讨论的形状文件。借助 GeoPandas，提取县和油田的步骤相同：

1. 将 VIIRS 记录读入地理空间数据对象，其原始坐标为 WGS 84。

2. 将形状文件读入地理空间数据对象，其原始坐标为 NAD 27。

3. 将形状文件中的所有几何对象转换为 WGS 84坐标。

4. 对这两个数据对象执行空间连接，以获取每个火炬对应的县或油田信息；如果某个特定县或油田的多边形与火炬相交（即二者具有任何共同的边界点或内部点），则视为相交。

3.4 相关性分析

为研究石油/天然气价格、燃烧统计数据与生产表现之间的相关性，提取了2015年5月至2018年12月（含首尾月份）的各类时间序列。以下列表列出了所用的所有变量及其对应标签：

VIIRS燃烧气体体积	VIIRS提供的月燃烧天然气体积
NDIC 火炬燃烧量	NDIC提供的月燃烧天然气体积
西德克萨斯中质原油价格	美国能源信息署提供的西德克萨斯中质原油价格（2020年b）
亨利枢纽天然气价格	美国能源信息署提供的亨利枢纽天然气价格（2020a）

NDIC石油产量：NDIC每月石油产量；NDIC天然气产量：NDIC每月天然气产量；VIIRS放空燃烧次数：VIIRS每月燃烧检测次数；NDIC燃烧气体井数：NDIC每月进行火炬燃烧的油井数量；NDIC气油比：NDIC天然气产量与NDIC石油产量之比

首先，从每个时间序列中提取月度观测值，并采用 Spearman 的 ρ来衡量变量之间的统计依赖关系。Spearman 的 ρ 是一种秩相关，用于量化两个变量排序之间的相关性。与 Pearson 的 r相比，它评估的是可能呈非线性的单调关系，并且对异常值更加稳健，因此本节采用这种方法。变量之间的两两相关性如图 3.5 所示。由于相关矩阵总是对称的，且对角线元素为 1，因此仅绘制不含对角线的下三角部分，以尽量减少信息冗余。

可以观察到，大多数变量对呈正相关。财务因素（即油气价格）并未出现在任何高度相关的变量对中（例如，相关性高于 0.80）。不过，结果表明 NDIC 与 VIIRS 报告数据呈正相关，石油产量也与火炬燃烧气体体积呈正相关。

在本分析中，由于该流程的性质（即提取月度数据，然后测量秩相关），时间尺度上的所有信息都被忽略了。为探究时间序列背景下的相关性，对每个变量进行一阶差分（即滞后一期差分）

$$y'_t = y_t - y_{t-1}, \tag{3.1}$$

然后评估两两之间的斯皮尔曼 ρ，并在图 3.6 中进行可视化。在本例中，高度相关的变量对并不多，只有石油和天然气产量在滞后一期差分上呈单调相关，这并不令人意外。在

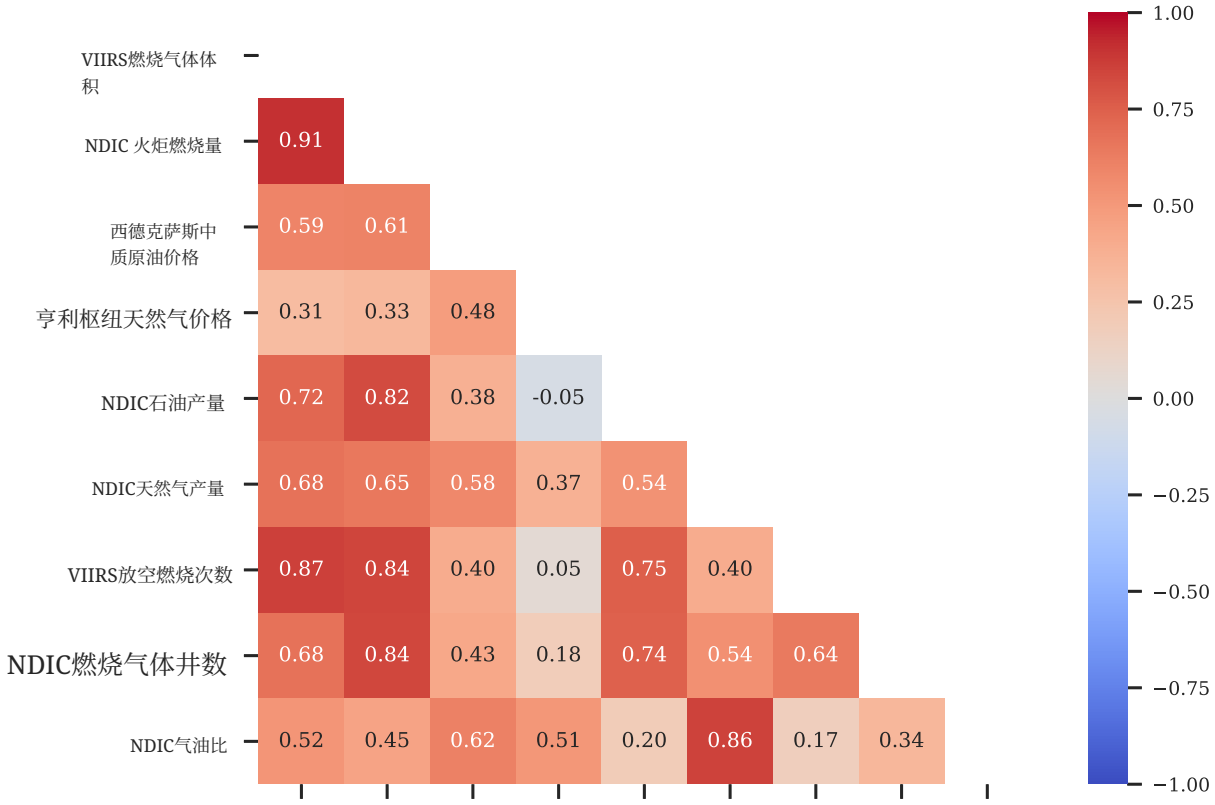


图 3.5：展示原始时间序列月度观测值两两斯皮尔曼相关性的热图。每个单元格中均标注了数值，对应的变量可通过读取纵轴和横轴上的刻度标签确定。

本学位论文的其余部分将重点关注火炬燃烧和生产相关统计数据，而非财务因素。

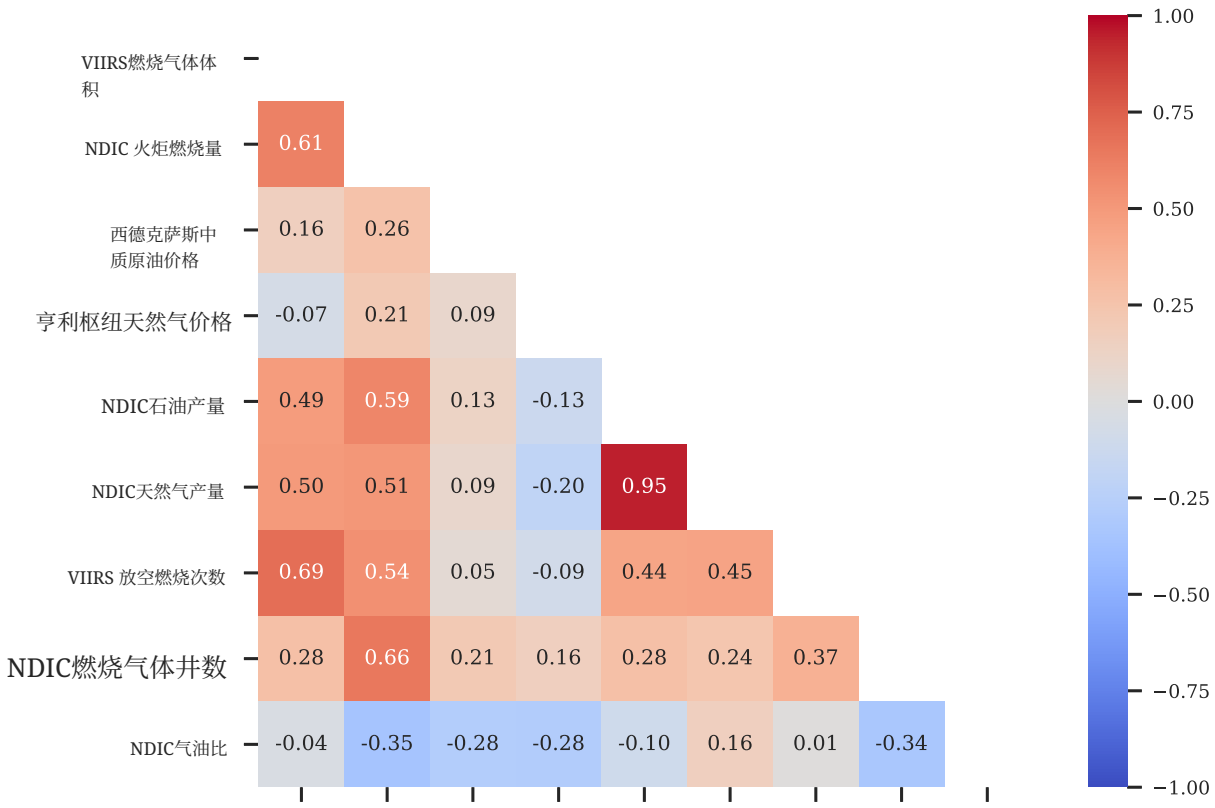


图 3.6：热图展示了对时间序列进行一阶差分后，各序列两两之间的斯皮尔曼相关性。每个单元格均标注了数值；对应变量可通过读取垂直和水平坐标轴上的刻度标签确定。

3.5 州级燃烧模型

本节构建了一个回归模型，用于研究 NDIC 与 VIIRS 报告数据之间的统计关系。图 3.7 展示了这两个来源的数据，表明二者呈正相关。

假设 NDIC报告采用高斯观测模型，并通过位置参数编码 VIIRS 的信息，该模型通过公式3.2a–



图3.7：NDIC 和 VIIRS 报告数据的可视化结果。左图显示时间序列，右图展示使用每个月份数据点绘制的散点图。

3.2e
进行
规定：

$$\alpha \sim \text{Half-Normal}(0.2) \tag{3.2a}$$

$$\beta \sim \text{Gamma}(2, 2) \tag{3.2b}$$

$$\sigma \sim \text{Half-Cauchy}(0.1) \tag{3.2c}$$

$$\mu_i = \alpha + \beta \times \text{VIIRS}_i \tag{3.2d}$$

$$\text{NDIC}_i \sim \mathcal{N}(\mu_i, \sigma) \tag{3.2e}$$

其中， α 是截距， β 是斜率；根据火炬燃烧体积的性质，二者都被约束为非负数； σ 是高斯似然函数中的标准差，也必须为非负数； μ_i 是第i月预期的NDIC报告，而NDIC_i 和 VIIRS_i 分别是NDIC和VIIRS在第 i月的观测数据（即报告体积）。用于定义该模型的符号清晰地表达了数据生成过程，并贯穿本学位论文始终。先验和超先验位于上方，观测模型位于下方。本文依据以下原则为该模型及本学位论文中的所有其他模型选择先验分布：

1. 优先采用弱信息先验，即在观测任何数据之前，根据现有领域专业知识选择先验。先验应足够强，能够体现领域专业知识；同时也应足够弱，以便“让数据说话”，也就是说，在数据量充足时由似然占主导。例如，在 β上设置一个均值为E$\beta = 2/2 = 1$ 的伽马分布先验，以反映这样一种假设：在观测任何数据之前，卫星解译流程得到的放空气量与NDIC报告相同。

2. 优先选择带有软约束而非硬约束的先验，即遵循克伦威尔定律。例如， α、 β 和 σ 的先验分布均以 $\mathrm{R}_{>0}$ 或 $\mathrm{R}_{\geq 0}$为支撑集。反例包括将三角分布或连续均匀分布用作此类量的先验，作者不建议这样做。

3. 优先选择最大熵分布，即根据现有的全部信息（遵守所有已知约束）做出最保守的假设。例如，高斯分布和二项分布都是最大熵分布，并在本学位论文中使用；利用库尔贝克–莱布勒（KL）散度的定义，可以对这一事实进行形式化证明。

确定先验分布和似然后，并行运行四条哈密顿蒙特卡洛马尔可夫链，从后验分布中进行采样。参数估计值列于表 3.1 中，后验分布和轨迹图见图 3.8。四条链分别以不同颜色绘制。轨迹图的 x 轴表示迭代次数。学位论文的其余部分将始终采用这种布局。

表 3.1：州级燃烧模型的参数估计值

参数	变量	点估计	90% 可信区间
α	截距	0.061	(0.044, 0.079)
β	斜率	0.535	(0.482, 0.590)
σ	报告变异性	0.030	(0.024, 0.035)

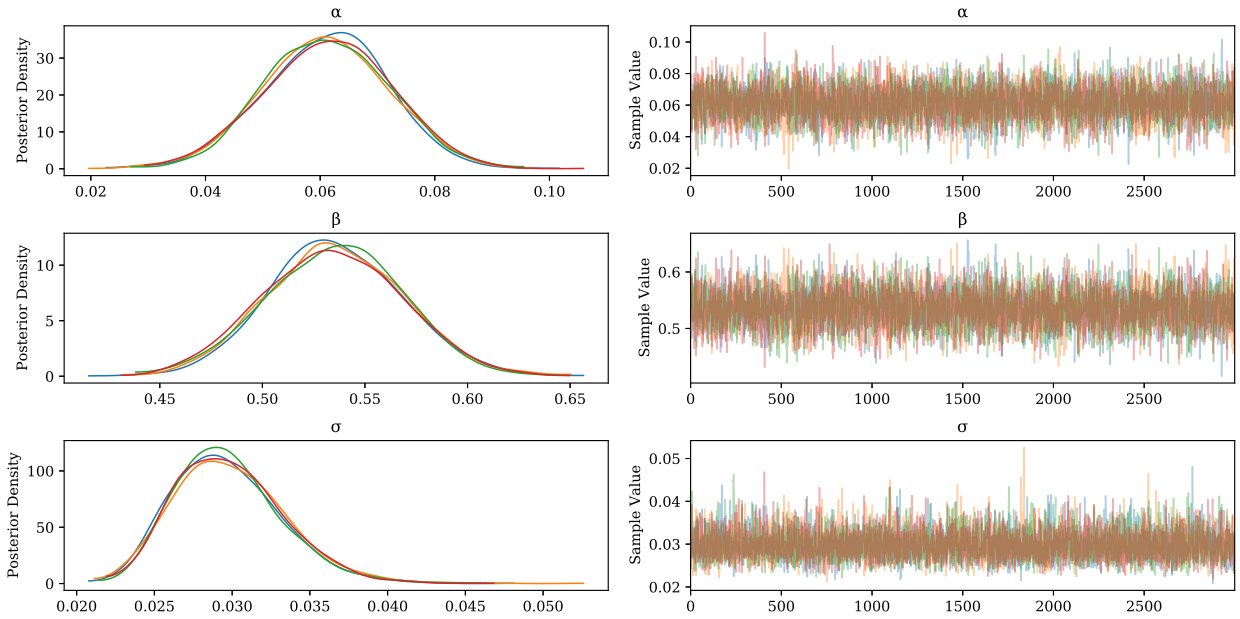


图 3.8：州级火炬燃烧模型的后验分布（左列）和轨迹图（右列）。如轨迹图所示，马尔可夫链已实现良好混合并收敛。

利用模型和轨迹图生成后验预测样本，以构建区间（图 3.9）。对于特定的机器学习模型而言，点估计和点预测很容易获得；然而，只有经过恰当构建的区间才能为决策提供有关不确定性的洞见。作者希望强调，在使用机器学习时量化不确定性的重要性，无论是用于推断、预测，还是构建可整合到基于物理的模型中的中间模型。在石油工程领域的一些应用和出版物中，这一点却不幸地被忽视或遗漏。恰当量化不确定性的重要性也将在接下来的章节中得到强调。

只要只需一个模型设定来进行点预测，就可以根据表 3.1 中的参数估计值还原该模型：

$$\mathrm{NDIC}_i = 0.061 + 0.535 \times \mathrm{VIIRS}_i \,, \tag{3.3}$$

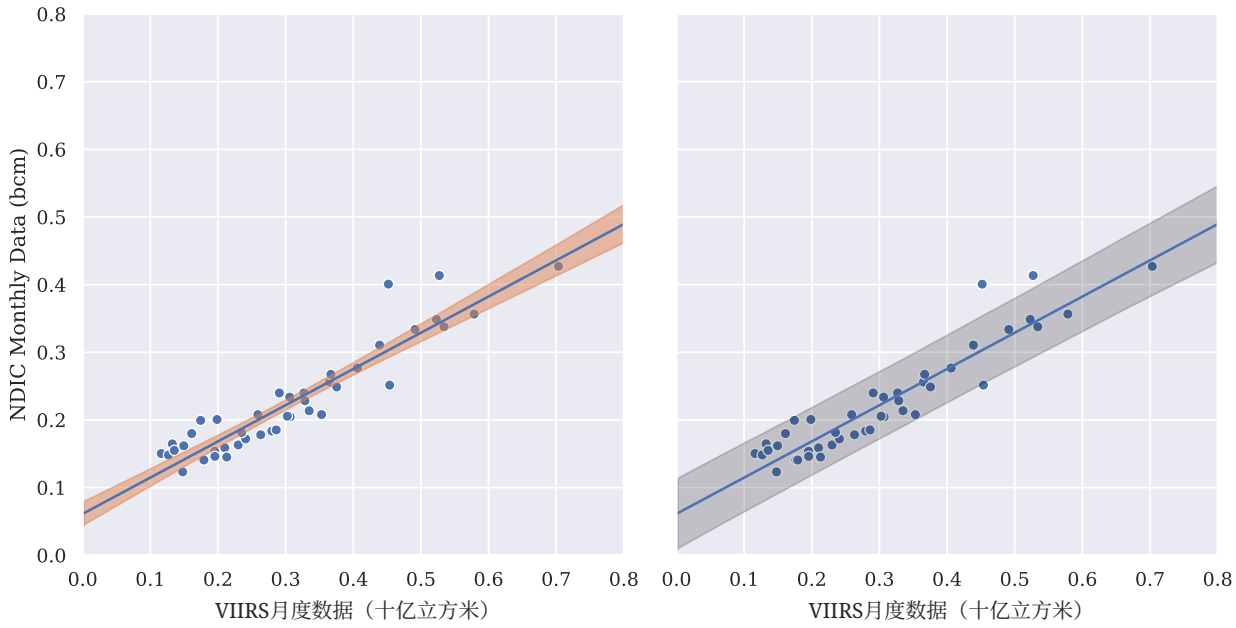


图 3.9：区间根据后验预测样本构建。在两幅图中，线条表示使用 α 和 β 的点估计（后验均值）得到的“最佳”拟合。左图中的阴影区域表示回归均值的 90 % 可信区间（CI）。右图中的阴影区域表示对未来 NDIC报告的 90 % 预测区间，大多数现有观测值都落在该区间内。

其中，NDIC_i 和 VIIRS_i 分别表示月份 i 的燃烧量，单位为十亿立方米。该模型还对全州范围内的 NDIC 报告回归均值提供了清晰的解释：

1. 截距表明，平均而言，每月有 90 % 的概率会有 0.04 至 0.08 十亿立方米的报告体积未被当前 VIIRS处理流程捕获。后验均值为 0.061 十亿立方米（≈ 2150 百万立方英尺）。

2. 斜率表明，平均而言，当卫星估算体积增加一个单位时，NDIC报告将以 90 % 的概率增加 0.48 至 0.59 个单位。后验均值为 0.535 个单位。

该模型虽然是用于州级 NDIC 报告的合适校准和估计工具，但假设州内的异质性（例如，不同县之间的差异）可以忽略不计，并且所有月度观测值都是条件独立的

且同分布（i.i.d.）。对于这些假设不成立的情形，可以构建其他类型的模型，相关讨论见第4章和第5章。

第4章

县级火炬燃烧模型

“多层回归应当成为回归的默认形式。”

——McElreath（2015年）

4.1 理解异质性

本章中，作者探讨了北达科他州不同县之间，州政府报告的火炬燃烧统计数据与卫星检测到的火炬燃烧统计数据之间相关性的异质性。其动机有三方面：

1. 提供比仅研究整个州的燃烧统计数据更细致的洞见。

2. 比较并对照不同县的报告与基准（即卫星检测结果）之间的一致性。

3. 为每个县建立专门的模型，用于校准和预测。

4.2 层次模型

《从数据中学习》中的一个常见问题，是对总体中的个体或单位进行建模。例如，为某个州的不同县，或某个油田中的不同井场建立模型。通常根据领域专业知识，人们预计这些单位会表现出一定差异；然而，它们不一定代表完全独立的数据生成过程。换言之，这些单位在某些方面有所不同，但在另一些方面又彼此相似。遗憾的是，以下两种常见的建模方法都较为极端，并不理想：

1. 完全汇聚

- 这种方法忽略了异质性，并假设所有个体的观测值均由完全相同的过程生成或描述。整个总体只学习一组参数。在这种情况下，方差可能较小，但偏差可能很大。

2. 不合并

- 这样一来，每个个体都可以利用自己的数据学习一组专属参数。这种做法假设，每个个体提供的信息与其他任何个体都无关。在这种情况下，偏差可能较小，但方差可能非常大。

实际上，由于总泛化误差较大，这两种方法都无法很好地用于洞见提取或预测任务。事实上，通过对所有个体构成的总体进行显式建模，可以折中这两种极端做法。也就是说，为了研究个体单位之间的相关性，需要为总体建立一个显式模型。在学习阶段，利用各个个体的后验分布来拟合某种总体分布，再将总体信息反馈给各个个体。在这种情况下，似然函数分布较宽的个体（例如数据较少的个体）会更多地向总体分布靠拢，而数据提供的信息较充分的个体，其后验分布则基本保持不变。通过这一过程，可以实现动态正则化：即通过部分合并数据，在偏差与方差之间取得平衡，从而大幅降低总泛化误差。

在县级模型开发的背景下，现在的问题是应如何对总体进行建模。为了说明选择某一特定类模型的依据，必须考察各县的一些特征。在本研究中，假定各县是可交换的，即联合概率 $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots\ldots, \boldsymbol{\theta}_n)$ 对指标的置换保持不变，其中$\boldsymbol{\theta}_i$、 $i = 1, 2, \ldots\ldots, n$ 是第 i 个县的参数。也就是说，对于任意置换 π，

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n) = p(\boldsymbol{\theta}_{\pi_1}, \boldsymbol{\theta}_{\pi_2}, \ldots, \boldsymbol{\theta}_{\pi_n}) \,. \tag{4.1}$$

此外，县的列表可以不断扩展，即尽管目前可能只考察少数几个县，但未来可能会纳入在火炬燃烧活动方面的新县。

如果被建模的总体具有可交换性，且总体可以任意增大，那么德·菲内蒂定理表明，唯一符合可交换性的分布是层级分布：

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n) = \int \left[\prod_{i=1}^{n} p(\boldsymbol{\theta}_i \mid \phi) \right] p(\phi) \, \mathrm{d}\phi \,, \tag{4.2}$$

其中，ϕ 是一个总体参数（可泛化为多个总体参数），而 $p(\phi)$是总体先验。该定理指出一个重要事实：如果分析中使用了可交换数据，那么必定存在一个总体模型（Jordan and Broderick 2010）。这为本章县级火炬燃烧模型的开发提供了指导。

等价地，可以联合拟合个体参数和总体参数，从而实现数据的动态汇聚：

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n, \phi) = \left[\prod_{i=1}^{n} p(\boldsymbol{\theta}_i \mid \phi) \right] p(\phi) \,, \tag{4.3}$$

在这一过程中，不仅学习 $\boldsymbol{\theta}$，也学习 ϕ 。将观测部分（$\mathcal{D} = \{(\mathbf{x}_j, \ y_j) \mid j = 1, \ldots\ldots, \ m\}$）加入后，联合模型变为：

$$p\Big(\big\{ y_j, \mathbf{x}_j, \boldsymbol{\theta}_{\text{COUNTY}[j]}, \psi_j \big\}_{j=1}^{m}, \phi \Big) = \left[\prod_{j=1}^{m} p(y_j \mid \mathbf{x}_j, \boldsymbol{\theta}_{\text{COUNTY}[j]}, \psi_j) \, p(\boldsymbol{\theta}_{\text{COUNTY}[j]} \mid \phi) \right] p(\phi) \,, \tag{4.4}$$

其中，$\boldsymbol{\theta}_{\text{COUNTY}[j]}$ 表示根据县级归属为第 j个观测设定的参数，而 ψ 是似然函数中的其他一些参数，它们不一定服从总体模型。公式 4.4 描述了一个层次模型，它与贝叶斯框架完美契合，并被用于构建本章中的模型。

作为建模异质性的一种基本方法，层级模型一直常规应用于包括生态学（Bolker 2008）、政治学（Gelman and Hill 2006）和生物学（McElreath 2015）在内的各个领域。作者认为

它们也应在石油工程领域得到广泛认可和应用，因为该领域的数据集通常以层级结构呈现。例如，某一给定盆地中的页岩气井由不同的油田服务公司完成。因此，可以在这些服务公司之间汇总信息。第7.3节将进一步讨论这一点。不过需要注意的是，德·菲内蒂定理基于这样一个假设：总体（由个体组成）是可交换的，并且可以任意增长。正如机器学习中的其他假设一样，这一假设不应被想当然地接受，也并非始终成立。在县级燃烧模型开发的背景下，有人可能会认为，北达科他州目前有53个县，并且在任何有限的时间内都不太可能出现许多新的县（作为行政区划）。在这一点上，作者同意 Box 等人（2009）的观点：既然假设“从来都不会完全正确”，那么应当寻求的是有用的 模型，而不是正确的 模型。这正是本章应用层级模型的目标。

值得注意的是，描述这类模型时所用的术语并不一致：有些人认为，层次模型和多层模型是同一种建模技术的不同名称（Bolker 2008；McElreath 2015），另一些人则试图对二者加以区分（Carpenter 2019）。在本学位论文中，我们通过数学结构而非术语来传达模型假设，并尽可能完整地写出模型定义。

4.3 数据描述

按照第3.3节所述进行逆地理编码后，发现有12个县同时在 VIIRS 和 NDIC 中报告了燃烧活动。对于每个县从2015年5月至2018年12月（含首尾月份）的历史数据，仅提取两个来源均有报告体积的月份。图 4.1展示了这12个县各自的散点图，其中县缩写遵循NDIC月度生产报告中的惯例。表4.1列出了每个缩写对应的完整县名。

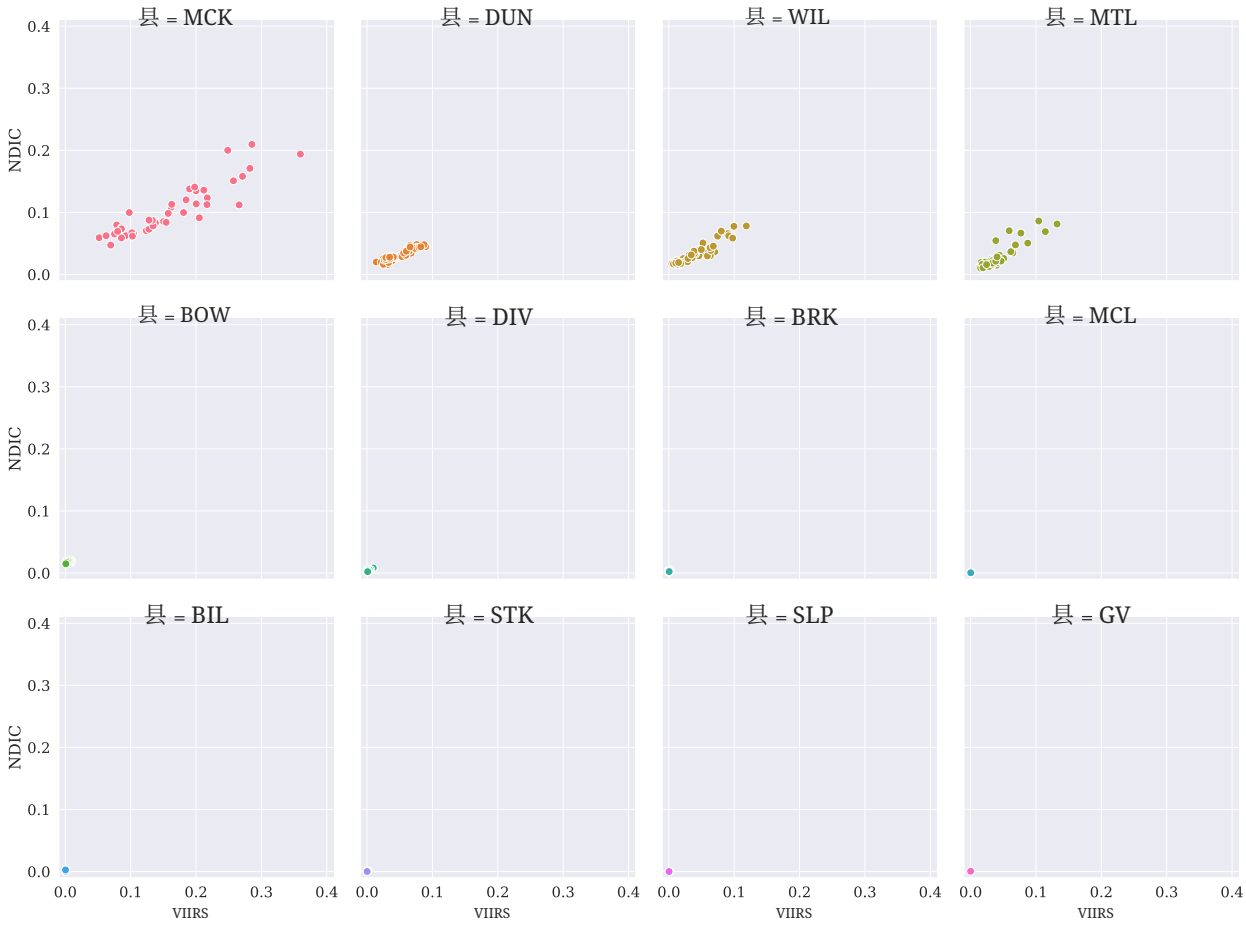


图 4.1：不同县的 NDIC 和 VIIRS 报告数据散点图。所有子图共用 x 轴和 y 轴。 x 轴表示以十亿立方米为单位的 VIIRS 月度燃烧气体体积报告数据， y 轴表示以相同单位计量的 NDIC 报告数据。

从燃烧气体体积来看，不同县的燃烧放空量差异很大。为更清晰地呈现这些数据，图 4.2 展示了各县的放大视图。可以看出，除 SLP 和 GV 外，大多数县的数据点都超过 ~ 12 ；不过，只有顶排的四个县（即MCK、DUN、WIL和MTL）拥有最多的数据，并显示出VIIRS与NDIC之间更强的正相关性。

为了构建县级模型并考察各县之间的异质性，上一节讨论的不合并选项并不适用。特别是对于 SLP 县（有 3 个观测值）和 GV 县（有 2 个观测值），如果拟合一个线性

表4.1：北达科他州县缩写

缩写	县
MCK	麦肯齐县
DUN	邓恩县
WIL	威廉斯县
MTL	芒特雷尔县
BOW	鲍曼县
DIV	迪瓦伊德县
BRK	伯克县
MCL	麦克莱恩县
BIL	比灵斯县
STK	斯塔克县
SLP	斯洛普县
GV	金谷县

模型（如公式 3.2d），学得的斜率参数 β_{COUNTY} 将得到点估计 $\beta_{\text{SLP}} \approx 0$ 和 $\beta_{\text{GV}} \gg 0$，并附有相应的样本。斜率参数的解释（在公式 3.3 之后讨论过）意味着这种推断是不可能的。其他一些县即使拥有更多数据点（例如 MCL），其观测值也会受到噪声水平的影响。仅使用各自的数据集会妨碍准确推断。因此，为了稳健地构建县级模型，我们采用上一节讨论的层次模型。

4.4 模型设定

受第 4.2 节讨论的启发，通过对整个县总体进行显式建模来实现部分汇聚。这样，诸如 MCL 这样的县便可以利用其他县的信息来学习自身的参数。具有“强数据”（即信息量非常丰富、足以使似然主导后验分布结构）的县，例如图 4.2 顶行中的县，表明 VIIRS 与 NDIC 之间存在正相关性。因此，对于各县采用与模型 3.2 类似的策略，即为每个县分别学习一组斜率和截距。

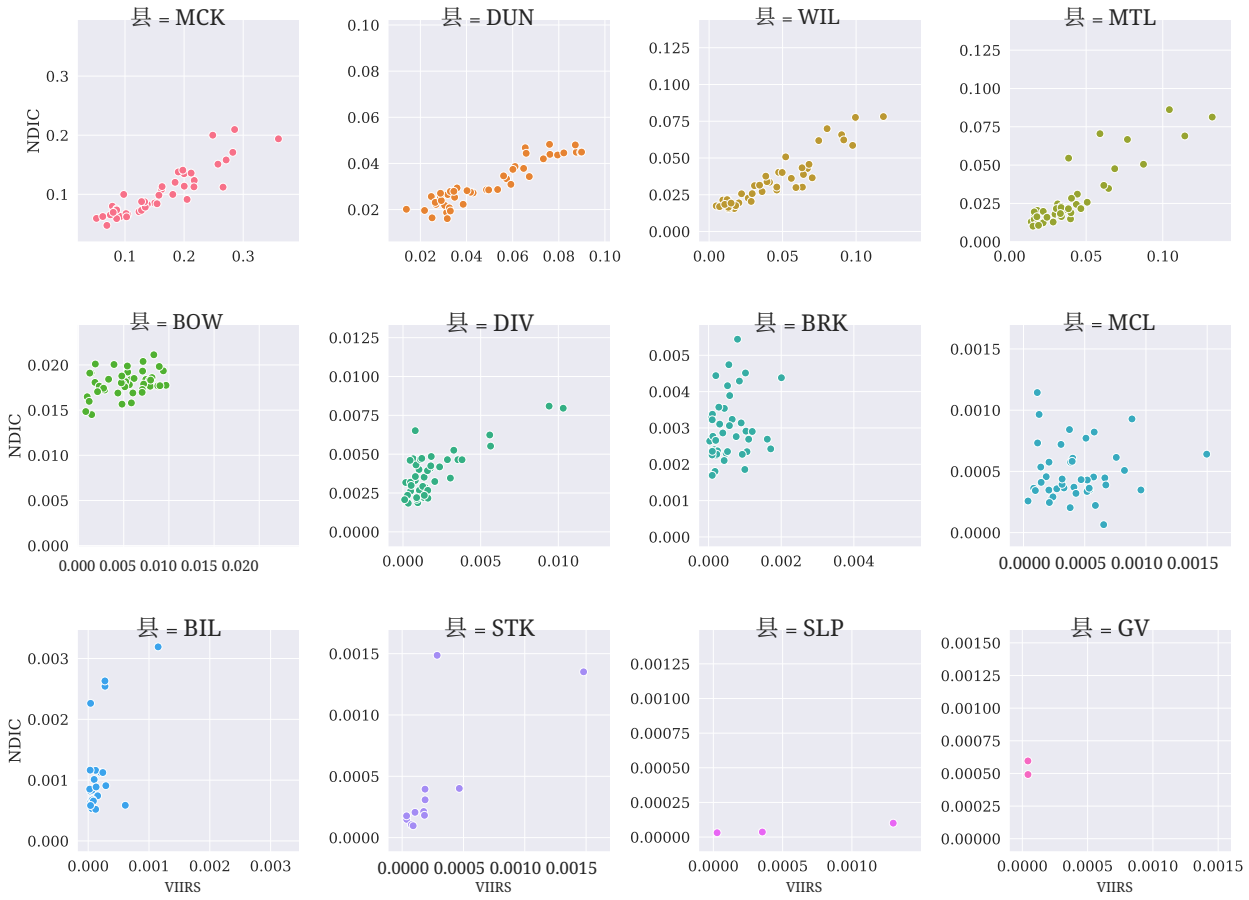


图 4.2：不同县的 NDIC 和 VIIRS 报告数据散点图，所有子图之间既不共享 x 轴，也不共享 y 轴。每个子图内，x 轴和 y 轴采用相同的比例和范围。坐标轴的含义与图 4.1 相同。

由于斜率和截距具有很强的可解释性，其含义已在公式 3.3 之后紧接着进行讨论，因此还可以跨参数类型（即截距和斜率）实现部分汇聚。换言之，了解 VIIRS 漏报了多少燃烧气体体积（即截距所携带的信息），可能有助于改进对 VIIRS 与 NDIC 将如何共同变化的学习（即斜率所携带的信息）。具体而言，对于不同县的参数，使用具有多元正态密度的总体模型。

层次模型通过表达式 4.5a–4.5j 指定：

$$\mu_\alpha \sim \text{Half-Normal}(0.1) \tag{4.5a}$$

$$\mu_\beta \sim \text{Gamma}(2, 2) \tag{4.5b}$$

$$\sigma_\alpha \sim \text{Half-Normal}(0.1) \tag{4.5c}$$

$$\sigma_\beta \sim \text{Half-Normal}(0.1) \tag{4.5d}$$

$$\sigma \sim \text{Half-Normal}(0.05) \tag{4.5e}$$

$$\mathbf{R} \sim \text{LKJcorr}(2) \tag{4.5f}$$

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_\alpha & 0 \\ 0 & \sigma_\beta \end{pmatrix} \cdot \mathbf{R} \cdot \begin{pmatrix} \sigma_\alpha & 0 \\ 0 & \sigma_\beta \end{pmatrix} \tag{4.5g}$$

$$\begin{bmatrix} \alpha_{\text{COUNTY}} \\ \beta_{\text{COUNTY}} \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} \mu_\alpha \\ \mu_\beta \end{bmatrix}, \mathbf{\Sigma} \right) \tag{4.5h}$$

$$\mu_j = \alpha_{\text{COUNTY}[j]} + \beta_{\text{COUNTY}[j]} \times \text{VIIRS}_j \tag{4.5i}$$

$$\text{NDIC}_j \sim \mathcal{N}(\mu_j, \sigma) \tag{4.5j}$$

其中：

μ_α 是所有县的平均截距；μ_β 是所有县的平均斜率；σ_α 是不同县截距之间的标准差；σ_β 是不同县斜率之间的标准差；σ 是县内 NDIC 报告的标准差；$\mathbf{R}$ 是服从 LKJ分布的相关矩阵。它是

2×2

维，并编码截距与斜率之间的相关性；$\mathbf{\Sigma}$ 是总体模型的协方差矩阵，由两侧分别乘以一个对角标准差矩阵构造而成——

将相关矩阵从两侧分别乘以一个由标准差构成的对角矩阵

偏差；

α_{COUNTY} 和 β_{COUNTY} 分别是每个县的截距和斜率，其先验分布

由二维高斯总体模型定义；

$\text{COUNTY}[j]$ （下标）表示县索引，即 $\text{COUNTY}[j] \in \{k \in \mathbf{N}_0 \mid k \leq 11\}$，其中 $\alpha_{\text{COUNTY}[j]}$ 和 $\beta_{\text{COUNTY}[j]}$ 分别是根据县归属确定的第 j 个观测的截距和斜率；

VIIRS_j 是第 j 次观测的 VIIRS 报告量；

μ_j 表示第 j 次观测的底层扩张体积；

NDIC$_j$ 是第 j个观测的NDIC报告体积。

Lewandowski、Kurowicka和Joe（2009年）提出的LKJ分布，是定义在对角线元素均为1的正定对称矩阵上的分布，即相关矩阵。在上述模型设定中，它直接影响协方差矩阵的先验。在该分布提出之前，且哈密顿蒙特卡洛尚未得到广泛应用时，由于Wishart分布和逆Wishart分布具有良好的共轭性质，它们通常用于对协方差矩阵建模。然而，LKJ分布更适用于现代贝叶斯计算环境（Betancourt，2015年；Lambert，2018年），因此本研究采用了该分布。

LKJ 只有一个参数 η，可以将其解释为对称贝塔分布的形状参数（Gelman 等，2013年）。随着 η 增大，该先验对矩阵中较大的相关性持更怀疑态度，即发挥正则化效应。图4.3展示了 LKJ 在若干个 η 取值下的概率密度。本文选择 LKJcorr($\eta = 2$) 来定义一个弱信息且具有正则化作用的先验。

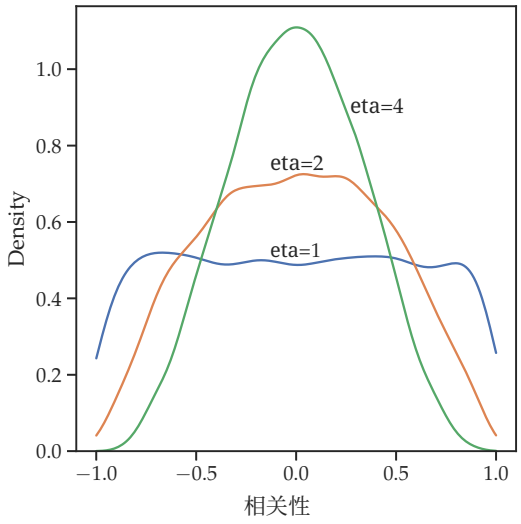


图4.3：LKJcorr($\eta =$ `eta`) 的概率密度。随着 η 增大，较大的相关性变得越来越不可信。

模型4.5虽然在数据生成过程方面具有较强的表达能力，但它是层次结构的一种中心化参数化（Papaspiliopoulos et al. 2007）。在这种参数化中，

层级参数（例如 β_{COUNTY}）与先验分布中的较低层级参数（例如，μ_β 和 σ_β）紧密耦合，并且它们在后验分布中高度相关。由于该模型的后验分布涉及复杂的几何对象和相互作用，因此采用哈密顿蒙特卡洛进行抽样。当数据量不大时（当前 NDIC 和 VIIRS 报告数据即属于这种情况），这种参数化会导致抽样效率极低且无法收敛（Stan 开发团队 2020）。在这些情况下，非中心化参数化更为适用，因此县级模型采用了这种参数化。

4.5 模型重新参数化

层次模型的重新参数化适用于位置-尺度族中的任意分布，而正态分布是其中的理想选择。对于多元正态先验分布的重新参数化，假设 $\boldsymbol{\theta}$ 的先验为均值向量为 $\boldsymbol{\mu}$ 、协方差矩阵为 $\boldsymbol{\Sigma}$ 的多元正态分布（例如表达式 4.5h），则非中心化参数化可表示为：

$$\widetilde{\boldsymbol{\theta}} \sim \text{MVNormal}(\mathbf{0}_n, \mathbf{I}_n) \tag{4.6a}$$

$$\boldsymbol{\varphi} = \boldsymbol{\mu} + \mathbf{L} \cdot \widetilde{\boldsymbol{\theta}} \tag{4.6b}$$

其中，$\widetilde{\boldsymbol{\theta}}$ 与 $\boldsymbol{\theta}$ 的维度相同，且其所有元素均独立同分布于 $\mathcal{N}(0,\ 1)$，$\mathbf{L}$ 满足 $\mathbf{L} \cdot \mathbf{L}^\top = \boldsymbol{\Sigma}$，而 $\boldsymbol{\varphi}$ 则恢复了 $\boldsymbol{\theta}$完全相同的先验分布。这种重新参数化通过降低$\widetilde{\boldsymbol{\mu}}$、$\mathbf{L}$和 $\boldsymbol{\theta}$之间的依赖性，使抽样更加高效。$\mathbf{L}$ 可以取 $\boldsymbol{\Sigma}$的 Cholesky因子，这为多元正态分布情形提供了便于实现的方案（Stan 开发团队 2020），本文采用了这一方案。

非中心化县级模型通过表达式 4.7a–4.7j 指定，其中以蓝色突出显示了重新参数化部分（对应模型4.5）：

$$\mu_\alpha \sim \text{Half-Normal}(0.1) \tag{4.7a}$$

$$\mu_\beta \sim \text{Gamma}(2, 2) \tag{4.7b}$$

$$\sigma_\alpha \sim \text{Half-Normal}(0.1) \tag{4.7c}$$

$$\sigma_\beta \sim \text{Half-Normal}(0.1) \tag{4.7d}$$

$$\sigma \sim \text{Half-Normal}(0.05) \tag{4.7e}$$

$$\mathbf{L} \sim \text{LKJCholeskyCov}\left(\eta = 2, \begin{bmatrix}\sigma_\alpha & \sigma_\beta\end{bmatrix}^\intercal\right) \tag{4.7f}$$

$$\begin{bmatrix} z_\alpha \\ z_\beta \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \tag{4.7g}$$

$$\begin{bmatrix} \alpha_{\text{COUNTY}} \\ \beta_{\text{COUNTY}} \end{bmatrix} = \begin{bmatrix} \mu_\alpha \\ \mu_\beta \end{bmatrix} + \mathbf{L} \cdot \begin{bmatrix} z_\alpha \\ z_\beta \end{bmatrix} \tag{4.7h}$$

$$\mu_j = \alpha_{\text{COUNTY}[j]} + \beta_{\text{COUNTY}[j]} \times \text{VIIRS}_j \tag{4.7i}$$

$$\text{NDIC}_j \sim \mathcal{N}(\mu_j, \sigma) \tag{4.7j}$$

其中：

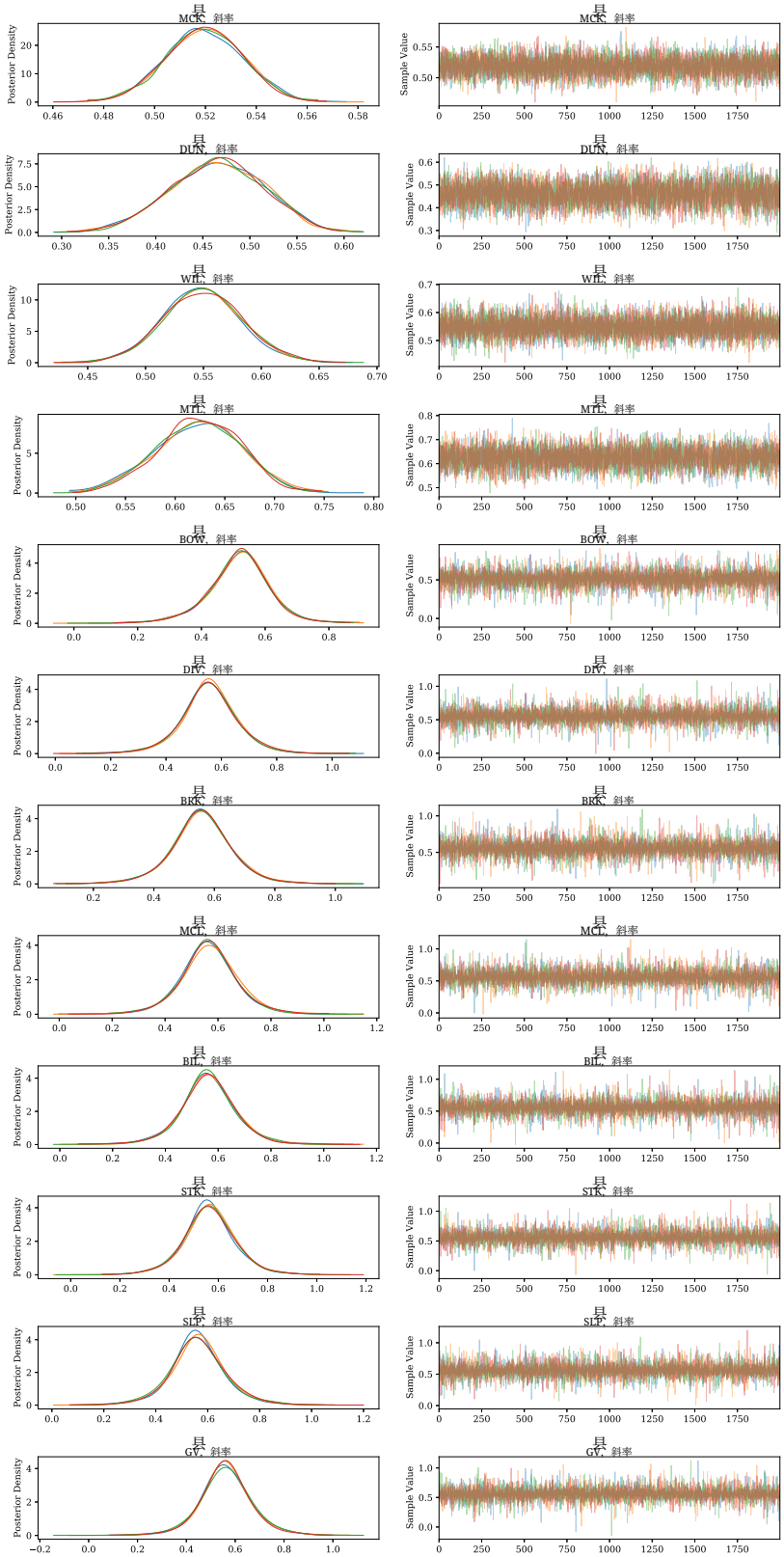
> **L** 是协方差矩阵的 Cholesky因子，该矩阵的相关系数服从 LKJ 分布；z_α 和 z_β 分别是每个县的标准化截距和斜率。

其余符号的含义与模型4.5中相同。非中心化模型采用与模型4.5完全相同的概率结构，用于推断各县的参数。

4.6 模型拟合

从后验分布中抽取了四条链。斜率和截距的后验分布及轨迹图分别见图4.4和图 4.5。从轨迹图可以看出，链已实现良好混合并收敛。

为了更好地比较各县的参数，图4.6和图4.7分别给出了斜率和截距的 90 % 最高密度区间（HDI）森林图。两幅图中，县按照 VIIRS 报告量排序，估计体积最少的县（如 SLP 和 GV）位于底部。细线表示 90 % 的最高密度区间，较粗的线段表示



Figu关于 **4.4**：各国斜率的后验分布和轨迹图 ty.

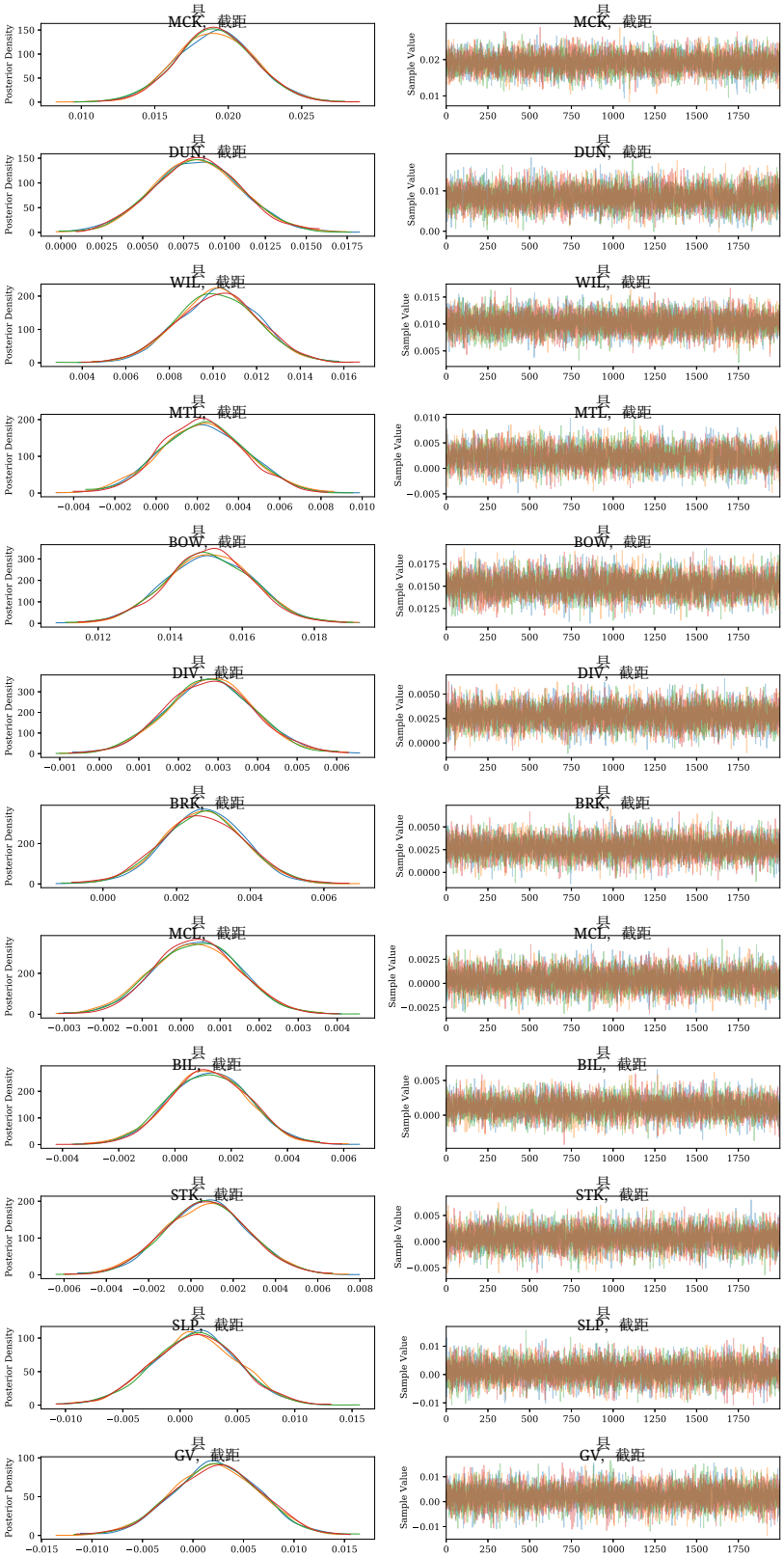


图 4.5：各县截距的后验分布和轨迹图。

四分位距（IQR）。这些点表示后验均值。

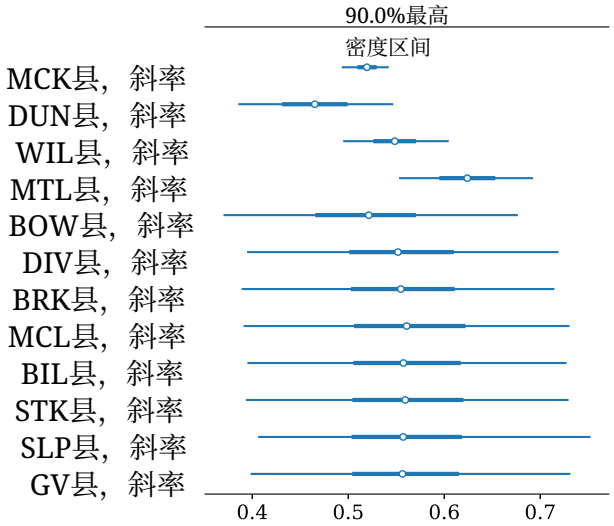


图4.6：森林图展示了各县斜率估计的不确定性。底部的县数据集不足或噪声较大，因此其估计值在很大程度上被拉向部分汇聚均值。

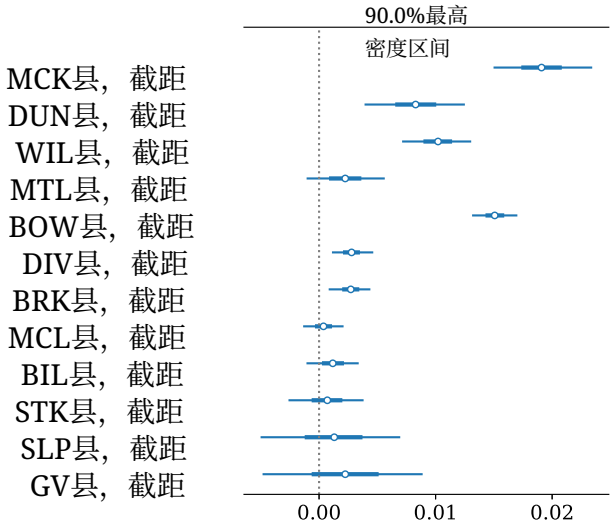


图4.7：森林图展示了各县截距估计的不确定性。虚线标示零截距，有些县的估计值与之没有显著差异。

就斜率而言（图 4.6），可以看出，前四个县之间差异颇大。MTL 在整个总体中的点估计最大（$\beta_{\text{MTL}} > 0.6$），而

DUN 的点估计最小（$\beta_{\mathrm{DUN}} < 0.5$）。此外，DUN 和 MTL 的最高密度区间很少重叠，这表明 MTL 的斜率几乎可以确定大于 DUN。观测值较少的县（其余八个县）的参数估计值不确定性更大，而它们的点估计值都被拉向介于 0.5 和 0.6 之间的部分汇聚均值。当某些县的数据不足时，层次模型会尽力加强不同县之间的信息共享，从而得出更合理的结果，同时量化这一过程中的不确定性。根据领域专业知识，与第 4.3 节讨论的非池化估计值（即 $\beta_{\mathrm{SLP}} \approx 0$ 和 $\beta_{\mathrm{GV}} \gg 0$）相比，这些结果在物理上更为合理。

对于截距（图4.7），各县之间也存在异质性。具体而言，通过绘制一条标示零截距的虚线，可以发现有些县的截距很可能为零（例如，零值落在四分位距或最高密度区间内），而另一些县的截距则与零有显著差异。对于数据较少的县（如 SLP 和 GV），得到接近零的截距以及较大的不确定性或许不足为奇；然而，MTL 的最高密度区间涵盖零值，这一点颇为有趣。请回想一下，截距参数可以解释为 NDIC 报告量中未被 VIIRS 捕获的部分。MTL 的这一发现，加上 MTL 的斜率点估计值最高（斜率越大，表示越接近卫星估算），令作者相信 MTL 过去曾有持续且较强的天然气火炬。它们通常使 VIIRS 能够捕捉到燃烧事件，因此 NDIC 与 VIIRS 的报告体积彼此更为接近。相反，DUN 较小的斜率和较大的截距表明，其火炬燃烧较为零星且强度较弱。值得一提的是，按照当前对截距的解释，负截距在物理意义上并不合理。尽管每个县的截距点估计值均为正，但有些县的最高密度区间涵盖了负值。这是选择二维高斯总体模型来描述截距和斜率所带来的局限。由于二维高斯分布在 R^2，上有定义，在某些县“数据较弱”的情况下，负值会

出现在它们的最高密度区间内。

上述讨论自然引出了一个问题：斜率与截距是否相关。事实证明，通过对不同参数类型进行部分合并，可以揭示斜率与截距之间可能存在负相关性（图 4.8）。这种相关性是从各县燃烧特征的异质性中学习得到的：

- 持续性燃烧产生较小的截距和较大的斜率。
- 零星燃烧产生较大的截距和较小的斜率。

换句话说，在整个县总体中，截距和斜率会共同变化。通过汇集不同参数类型之间的信息，模型从截距中学到的信息可以改善对斜率的学习，反之亦然。凭借这种“经验”或“知识”，层次模型即使在开始阶段只有少量观测值，也能够快速更新对任何新县参数的期望。需要注意的是，正相关值也具有一定的概率质量，即负相关并不是很强。这可能是因为目前一些县没有足够的数据。随着更多数据的纳入，后验分布将得到更新。

最后，参数估计值列于表 4.2，据此可以还原各县的参数模型，并将其用于校准和预测场景。

4.7 模型可扩展性

回顾前几节介绍的层次模型和重新参数化策略，有四种值得探讨的潜在部署场景。这些场景展示了所选方法在火炬燃烧数据分析方面的可扩展性和灵活性：

1. 根据 VIIRS 和 NDIC 报告的火炬燃烧统计数据，出现了新的县。

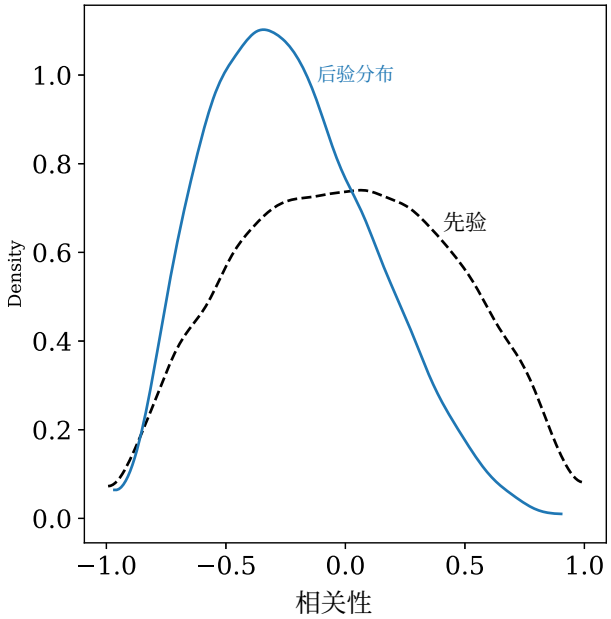


图 4.8：截距与斜率之间的相关性。蓝色：相关性的后验分布，其众数小于零。虚线：先验分布，即 LKJcorr(2) 密度。

目前，有 12 个县报告了来自 VIIRS 和 NDIC 的火炬燃烧统计数据。如果未来其他县也有火炬燃烧数据可用，层次模型便可立即扩展总体以纳入这些新县。这可以从方程 4.3 的条件结构中看出：通过为 $n+1$ 县建立一个模型

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}, \phi) = \left[\prod_{i=1}^{n+1} p(\boldsymbol{\theta}_i \mid \phi)\right] p(\phi)\,, \tag{4.8}$$

然后提取出右侧（RHS）中第 $(n+1)$ 个县对应的项)

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}, \phi) = p(\boldsymbol{\theta}_{n+1} \mid \phi) \left[\prod_{i=1}^{n} p(\boldsymbol{\theta}_i \mid \phi)\right] p(\phi)\,, \tag{4.9}$$

可以看出，右侧剩余部分是n 县的层次模型

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}, \phi) = p(\boldsymbol{\theta}_{n+1} \mid \phi)\, p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n, \phi)\,. \tag{4.10}$$

表 4.2：县级火炬燃烧模型参数估计值

参数	变量	县	点估计	90%置信区间
α_{COUNTY}	截距	MCK	0.019	(0.015，0.023)
		DUN	0.008	(0.004, 0.013)
		WIL	0.010	(0.007, 0.013)
		MTL	0.002	(−0.001，0.006)
		BOW	0.015	(0.013，0.017)
		DIV	0.003	(0.001, 0.005)
		BRK	0.003	(0.001, 0.004)
		MCL	0.000	(−0.001，0。002)
		BIL	0.001	(−0.001, 0.003)
		STK	0.001	(−0.003，0。004)
		SLP	0.001	(−0.005, 0.007)
		GV	0.002	(−0.005，0.009)
β_{COUNTY}	斜率	MCK	0.519	(0.493, 0.542)
		DUN	0.464	(0.385, 0.547)
		WIL	0.549	(0.495, 0.605)
		MTL	0.623	(0.553, 0.693)
		BOW	0.516	(0。370，0。677)
		DIV	0.554	(0.395, 0.719)
		BRK	0.556	(0.389, 0.715)
		MCL	0.563	(0.391, 0.730)
		BIL	0.560	(0.395, 0.727)
		STK	0.562	(0.393, 0.729)
		SLP	0.561	(0.406, 0.752)
		GV	0.560	(0.398, 0.731)

这表明，新引入的县仅取决于总体参数 ϕ，即新县与现有县（来自初始数据集）之间如何相互作用并未明确说明，而是通过 ϕ 进行中介。这种机制允许县的总体任意扩展。实际上，无需进行任何修改，即可使用新数据集整体重新拟合模型 4.7。

2. 过去观测值很少的那些县，现在有了更多数据。如果像 SLP 和 GV 这样最高密度区间较宽的县有了更多数据，后验分布将根据这些信息进行更新。它们的

最高密度区间会随着数据越来越多而逐渐变窄；由于层次模型会汇集各县的信息，这些县将有助于更新总体模型和其他县的参数。与上文的条目1类似，无需修改模型4.7，只需用新数据重新拟合即可。

3. 各县的样本量变得更加不均衡。

一般而言，当每个县都有大量数据时，中心化参数化（模型4.5）效率更高。当样本量不大时（目前 VIIRS 和 NDIC 报告数据的情况正是如此），非中心化参数化（模型 4.7）更合适。然而，层级模型的参数化并非单一策略。如果各县报告的火炬燃烧数据变得非常不均衡，例如，有些县的数据量极大，而另一些县的数据量极少，那么可以对各县采用不同的参数化。具体而言，

- 对于数据充分、似然函数占主导地位的县，可以通过表达式4.5f–4.5h采用中心化参数化。

- 对于数据不足、先验模型占主导地位的县，可以通过表达式4.7f–4.7h采用非中心化参数化。

总而言之，这仍然是一个层级模型，它定义的概率结构与模型4.5或模型 4.7完全相同，但避免了从后验分布中采样时的低效和不收敛问题。

4. 需要考察油田层级异质性。

在假设北达科他州的油田可交换且油田总体（其中的油田会进行火炬燃烧）可以增长的情况下，本章建立的层次模型可直接用于研究不同油田参数的异质性。按照第3.3节所述进行逆地理编码后，

共有258个油田在与本章相同的研究周期内同时具有 NDIC 和 VIIRS 报告数据。一些油田的观测值很少，可以通过在整个油田总体中汇集信息而受益于层次模型。此外，由于油田数量相对较多，总体模型可以更容易地学习（因为总体拥有更多信息）。对于本章建立的县级模型，由于总体中只有12个个体（县），关于总体的一些不确定性不可避免地存在，并通过后验分布体现出来。

本章构建的模型虽然捕捉到了北达科他州不同县之间的异质性，但依赖于这样一个假设：某个县内的所有月度观测均为条件独立同分布。对于必须考虑时间结构的情况，可以构建其他类型的模型，下一章将对此进行讨论。

第5章

燃烧放空时间序列分析

> “神经网络是否被过度炒作了，或者我们是否低估了平滑方法的能力？我认为这两种说法都成立。”
>
> — MacKay（2003年）

5.1 学习耀斑模式与行为

本章中，作者提出了一个用于揭示燃烧放空模式及其行为的通用框架。主要挑战有四个:

1. 观测数据存在噪声。

   - 公司估算火炬燃烧量并自行报告。卫星可能会漏掉一些事件。然而，在许多情况下，了解潜在过程至关重要，包括州和地方政府需要依据数据作出关键决策时。同时，了解潜在过程有助于通过区分报告中的真实异常与普通噪声或随机性来进行异常检测。

2. 宜采用概率方法。

   - 相比单一的最佳拟合函数，一组最可能的函数（用于刻画潜在过程）更为可取。

3. 某个实体的观测值构成时间序列。

   - 时间结构是数据集的内在属性，因此必须加以利用。

4. T该框架应足够通用，以支持自动化洞察提取 .

- 北达科他州有超过 200 家运营商和 500 个油田在运营。为每个实体选择特定的模型参数形式（例如差分自回归移动平均模型或长短期记忆），然后将模型拟合到数据，不仅耗时，而且会阻碍其轻松集成到自动化流程中（例如用于提取洞察）。

值得注意的是，高斯过程的优雅特性使其成为应对所有这些挑战的自然选择，因此本章采用了高斯过程。

5.2 高斯过程

高斯过程（GP）可视为函数的无限维希尔伯特空间上的一种分布。其正式定义为“由一组随机变量构成，其中任意有限个变量都服从联合高斯分布”（Rasmussen 和 Williams，2006年）。高斯过程是非常强大的非参数学习技术，兼具灵活性与可解释性。对于需要以严谨方式处理不确定性并进行解释、且数据集含噪并随时间动态变化的问题，这类技术尤为适用。此类情形包括平滑（Deisenroth 等，2012年）和时间序列建模（Roberts 等，2013年）。高斯过程在不同领域中也已得到广泛应用，并有各种不同名称，例如地统计学中的克里金法和卡尔曼滤波器，二者都对应于高斯过程（MacKay，1998年）。

本研究旨在开发一种通用框架，用于识别反映火炬燃烧策略与行为的底层未知过程 $f(\mathbf{x})$。因此，我们直接在函数空间中进行推断，并将高斯过程作为先验。高斯过程由其均值函数 $m(\mathbf{x})$ 和协方差函数 $k(\mathbf{x}, \mathbf{x}')$ 完全确定（Bandyopadhyay，2018年），定义如下：

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]\,, \tag{5.1}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]\,, \tag{5.2}$$

而服从高斯过程分布的函数记为

$$f(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}),\, k(\mathbf{x}, \mathbf{x}')\big)\,. \tag{5.3}$$

5.2.1 均值函数

在本研究中，均值函数始终设为零，因为对于潜在过程的均值没有先验知识。同时，对于均值函数为零的高斯过程，后验过程的均值并不局限于零（Rasmussen 和 Williams，2006年）。本学位论文中使用高斯过程先验建模的所有潜在函数均遵循

$$f(\mathbf{x}) \sim \mathcal{GP}\big(0,\, k(\mathbf{x}, \mathbf{x}')\big)\,, \tag{5.4}$$

其中， k 是某个协方差函数。

5.2.2 协方差函数

协方差函数也称为核函数，是高斯过程中的关键组成部分，因为它通过定义相似性来编码人们对函数应如何变化的假设。其基本假设是，输入$\mathbf{x}$ 彼此接近的数据点，其目标值也会相似 y。这一假设在包括时间序列建模在内的诸多领域通常都很合理，并且在理论上有托布勒地理学第一定律作为支撑。本学位论文中使用的协方差函数包括：

1. Mat´ern 类协方差函数，其形式如下：

$$k_\nu(r) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{r}{\ell}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\frac{r}{\ell}\right), \tag{5.5}$$

其中， $\Gamma(\cdot)$ 是伽马函数， K_ν 是阶数为 ν 的第二类修正贝塞尔函数， $r = \|\mathbf{x} - \mathbf{x}'\|$，而从某种角度来看， ℓ 是控制平滑性的长度尺度：较大的 ℓ 表示函数变化缓慢，因此可以更可靠地外推到更远的位置。

当 ν 为半整数时，Mat´ern 协方差函数可以写成指数函数与 p 阶多项式的乘积：$\nu = p + 1/2,\quad p \in \mathbb{N}_0$。超参数

ν 从另一个角度控制平滑性：当 $\nu = 1/2$ 时，Mat´ern 核函数变为指数核（连续但不可微）；当 $\nu \to \infty$ 时，它变为指数二次核（无限可微）。Rasmussen 和 Williams（2006年）认为，对于机器学习而言，最有趣的情形应当是$\nu = 3/2$ 和 $\nu = 5/2$。

对于天然气燃除时间序列，运营商可能会因政策变化、天然气处理设施部署、天然气价格波动等因素，在任何时候改变燃除策略，因此潜在过程可能并非光滑到无限可微。相反，我们采用 Mat´ern 核函数，它能够生成非光滑的函数实现，从而处理这些不连续性。具体而言，本学位论文选择了参数为 $\nu = 5/2$ 的 Mat´ern 核函数，输入空间为 $\mathcal{X} \subseteq \mathbb{R}^1$:

$$k_{\text{MATÉRN52}}(x, x'; \ell) \coloneqq \left(1 + \frac{\sqrt{5(x - x')^2}}{\ell} + \frac{5(x - x')^2}{3\ell^2}\right) \exp\left[-\frac{\sqrt{5(x - x')^2}}{\ell}\right], \quad (5.6)$$

其中 x 在时间域内变化。

2. MacKay（1998年）提出的标准周期核:

$$k_{\text{PERIODIC}}(x, x'; T, \ell) \coloneqq \exp\left(-\frac{\sin^2(\pi|x - x'|\frac{1}{T})}{2\ell^2}\right), \quad (5.7)$$

其中 T 表示周期。该核函数用于建模季节性行为 .

3. 白噪声核，定义如下:

$$k_{\text{WHITENOISE}}(x, x'; \delta) \coloneqq \delta^2 \mathbf{I}_n, \quad (5.8)$$

其中 δ^2 是噪声的方差。在本学位论文中，使用白噪声核是为了稳定协方差矩阵的计算。添加一个较小的对角线偏移值，旨在尽可能保证所得协方差矩阵始终为半正定。

一个良好的性质是，已构建核函数的和与乘积仍然是有效的核函数。本研究的模型构建过程也利用了这一点。

5.2.3 推断与模型重新参数化

实际应用中，我们处理的始终是有限大小的数据集。在这种情况下，会对函数值向量**f**赋予多元正态先验分布，

$$\mathbf{f} \sim \mathrm{MVNormal}(\mathbf{m}_x,\, \mathbf{K}_{xx})\,, \tag{5.9}$$

其中，向量$\mathbf{m}_x$ 和矩阵$\mathbf{K}_{xx}$ 分别是均值函数和协方差函数在输入$\mathbf{x}$上的取值。

一个对推断有重大影响的关键问题是如何从数据中学习超参数。一种自然且广受欢迎的方法是进行最大似然估计，即利用数据生成点估计。然而，正如 Betancourt（2017a）通过实验结果所示，无论是正则化还是未正则化的最大边际似然，在稳健拟合和恢复真实数据生成过程方面的表现都有限。从技术上讲，给定具有特定超参数的特定核函数，高斯过程并不涵盖整个希尔伯特空间，而只涵盖该空间中的一个切片；超参数即使只发生无穷小的变化，也会产生一个与原切片毫无重叠的不同切片。因此，本学位论文采用完整的贝叶斯方法进行高斯过程推断，即针对特定核函数考虑所有可能的超参数，从而纳入整个函数希尔伯特空间。

对于采用高斯观测模型的问题，GP 有优雅的闭式后验结果。然而，对于不采用高斯观测模型的情况，例如本学位论文中使用 Student-t 或泊松似然的情况，则不存在解析解。第 2.3 节所述的哈密顿蒙特卡洛用于从后验分布中采样。

具体而言，这里利用了潜在多元高斯分布的非中心化参数化。重参数化模型为

$$\widetilde{\mathbf{f}} \sim \text{MVNormal}(\mathbf{0}_n,\, \mathbf{I}_n) \tag{5.10a}$$

$$\mathbf{L} = \text{Cholesky}(\mathbf{K}_{xx}) \tag{5.10b}$$

$$\mathbf{f} = \mathbf{m}_x + \mathbf{L} \cdot \widetilde{\mathbf{f}} \tag{5.10c}$$

它定义的分布与表达式 5.9 相同，但能形成更适合哈密顿蒙特卡洛探索和抽样的后验几何（Betancourt 2017a）。

完成超参数学习后，可得到不属于原始数据集的潜在函数值的后验预测分布，具体如下：

$$\mathbf{f}_* \mid \mathbf{f} \sim \text{MVNormal}\Big(\mathbf{m}_* + \mathbf{K}_{x*}^\top \mathbf{K}_{xx}^{-1}(\mathbf{f} - \mathbf{m}_x),\, \mathbf{K}_{**} - \mathbf{K}_{x*}^\top \mathbf{K}_{xx}^{-1} \mathbf{K}_{x*}\Big)\,, \tag{5.11}$$

其中，$\mathbf{m}_*$ 是针对新输入计算的均值函数，$\mathbf{K}_{**}$ 是新输入之间的协方差，$\mathbf{K}_{x*}$ 是原始输入与新输入之间的协方差。

5.3 用于模式识别的模型套件

本节从多个角度构建模型，旨在以系统、合理的方式建立一个连贯的框架，用于学习燃烧放空模式与行为。每个模型均使用来自北达科他州的真实燃烧数据进行测试。在通过油田层面或运营商层面的研究展示更精细的分析能力时，所用数据来自一个主要产油油田 Blue Buttes 油田（ Alexeyev 等，2017 ）以及一家以 “ 运营商 A” 指代的运营商。

5.3.1 燃烧的天然气比例建模

天然气产量中被火炬燃烧的比例是燃烧强度和能源效率的指标。研究特定运营商和油田的这一比例是否随时间发生变化，很有意义。该模型由表达式 5.12a–5.12i 给出：

$$\ell \sim \text{Gamma}(2, 1) \tag{5.12a}$$

$$\eta \sim \text{Half-Cauchy}(5) \tag{5.12b}$$

$$\nu \sim \text{Gamma}(2, 0.1) \tag{5.12c}$$
$$\hat{\sigma}^2 \sim \text{Half-Cauchy}(5) \tag{5.12d}$$
$$k = \eta^2 \times k_{\text{MATÉRN52}}(x, x'; \ell) \tag{5.12e}$$
$$f \sim \mathcal{GP}(0, k) \tag{5.12f}$$
$$\pi_i = \text{logit}^{-1}\big(f(x_i)\big) \tag{5.12g}$$
$$\mu_i = \pi_i \times G_i \tag{5.12h}$$
$$F_i \sim \text{Student-}t(\nu, \mu_i, 1/\hat{\sigma}^2) \tag{5.12i}$$

其中:

ℓ 是 Mat´ern 核函数的长度尺度;

η 是边际偏差参数，用于控制潜在函数

在输出空间中的变化程度;

ν 是 Student-t 似然的自由度;

ˆσ^2 控制 Student-t 似然的逆尺度参数（类似于高斯分布的精度）;

k 是 GP 的协方差函数;

f 表示潜在过程，其服从高斯过程（GP）;

π_i 表示第 i个月的底层放空天然气比例。由于比例介于

0 和 1 之间，因此对潜在过程应用逆逻辑函数；G_i 表示第 i个月的天然气总产量;

μ_i 表示第 i个月的底层燃烧气体体积;

F_i 表示报告的燃烧气体体积，该数值采用 Student-t 观测模型进行建模。

选择 Student-t 观测模型的原因，是使模型设定能够泛化到尽可能多的实体，并能稳健地应对（可能大量存在的）异常值和含噪数据点。这是因为目前运营商必须按照各自的流程估算放空气量并进行报告，而在这一过程中会无意或有意地引入误差。Student 分布更厚的尾部

t-分布是应对这些现象的一种自然建模选择。这种思路，即设计通用且稳健的模型，确实体现在半柯西先验分布（其尾部较厚且信息性很弱）的选择，以及将 GP 用作非参数回归技术上。

为展示该模型处理真实数据的能力，我们对 Blue Buttes油田和运营商A进行了测试。使用的数据是 NDIC 提供的生产数据和燃烧量。对于该油田，图 5.1 展示了超参数的后验分布和轨迹图。图 5.2 展示了天然气燃烧比例（π_i）潜在过程的后验预测样本，清晰地描绘了其趋势。各色带对后验样本的覆盖范围如下：

- 颜色最深的色带（在某个 x 位置处位于中心）表示第 49 百分位数至第 51 百分位数；
- 颜色最浅的色带（在某个 x 位置处区间最宽）表示第 1 百分位数至第 99 百分位数。

此外，从高斯过程后验中抽取了 30 个随机样本，并以细线形式绘制在同一幅图中。潜在函数并未穿过所有观测数据点，否则模型就会过拟合；相反，它们呈现的是与数据以及模型内在假设最为相容的可能函数。一方面，我们已经获得了相关洞见，即推断出了潜在过程。另一方面，这也可作为异常检测工具。例如，州政府可能会想要仔细查看 2019 年下半年那些明显偏离“真实”过程的观测数据，例如审查该月份的报告，或调查导致火炬燃烧量在短短一个月内骤降的原因。

采用完全相同的模型设定，模型也使用该运营商的数据运行。超参数的后验分布和轨迹图见图5.3。

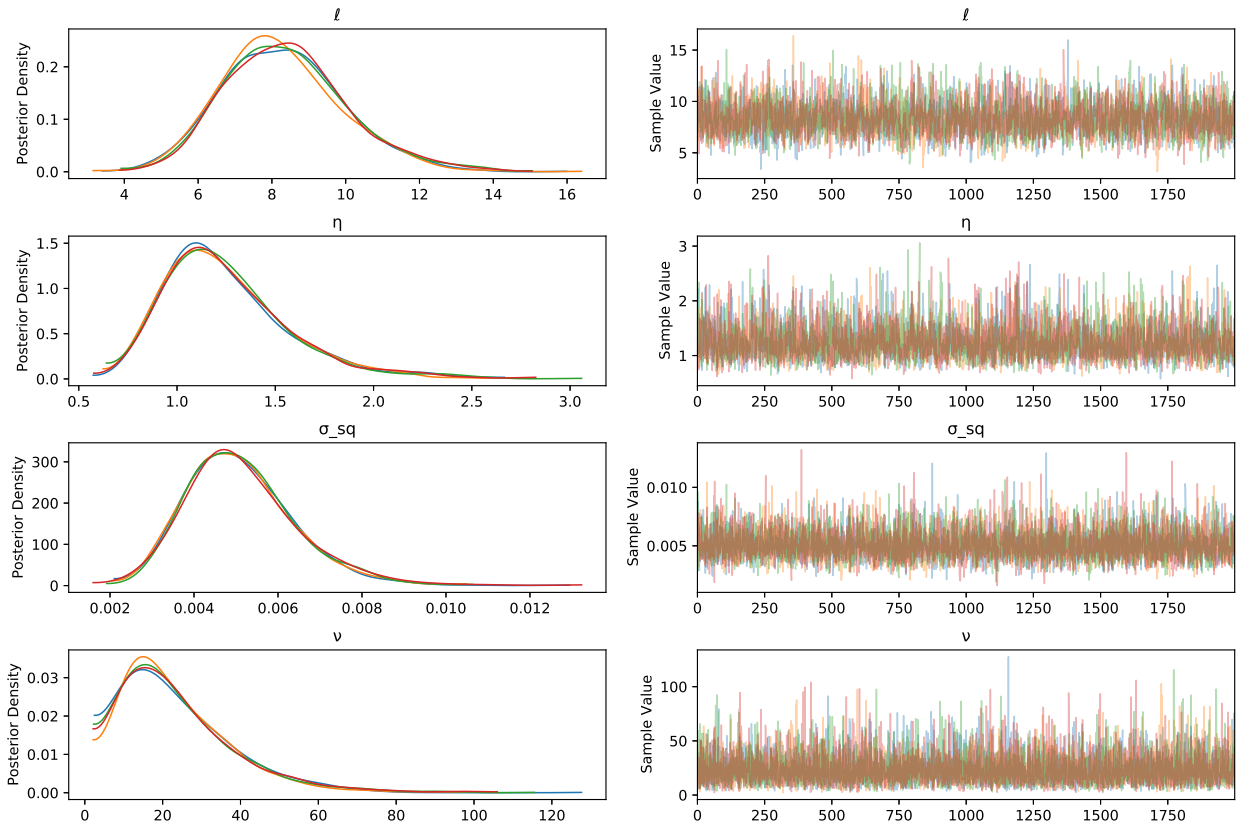


图 5.1：Blue Buttes油田天然气燃烧比例模型的后验分布和轨迹图。已实现良好混合并达到收敛。

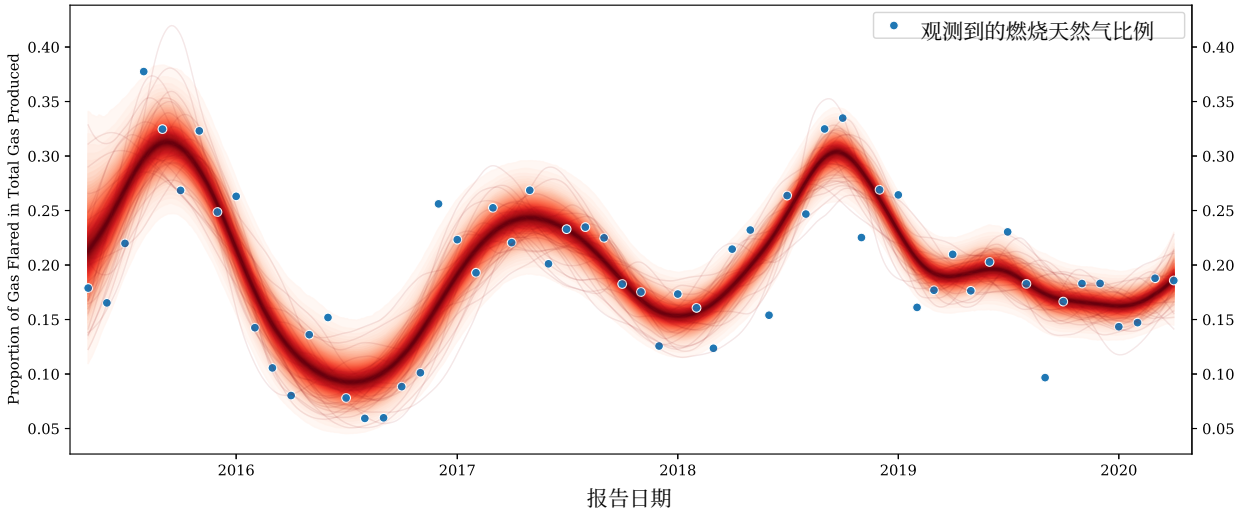


图 5.2：后验预测样本显示了 Blue Buttes油田天然气燃烧比例的变化。蓝色点表示观测数据，红色线条表示后验样本。

天然气燃烧比例潜在过程（π_i）的后验预测样本如图5.4所示。可以看出，这家运营商的燃除比例时间序列比 Blue Buttes油田（由五家以上公司运营）更加锯齿状。运营商可以更迅速地调整燃烧策略，这一点也能被捕捉到。尽管如此，长期趋势仍然清晰可见。比较图 5.1 和图 5.3，可以看出后验分布截然不同。然而，它们的先验分布完全采用相同的方式设定。这体现了贝叶斯方法的强大之处。以 ℓ 为例，我们为其设置了 Gamma(2，1)先验。然而，在根据数据进行条件化之后，运营商模型报告的长度尺度值平均更小（表明过程更锯齿状），而油田模型报告的长度尺度值则更大（表明过程更平滑）。

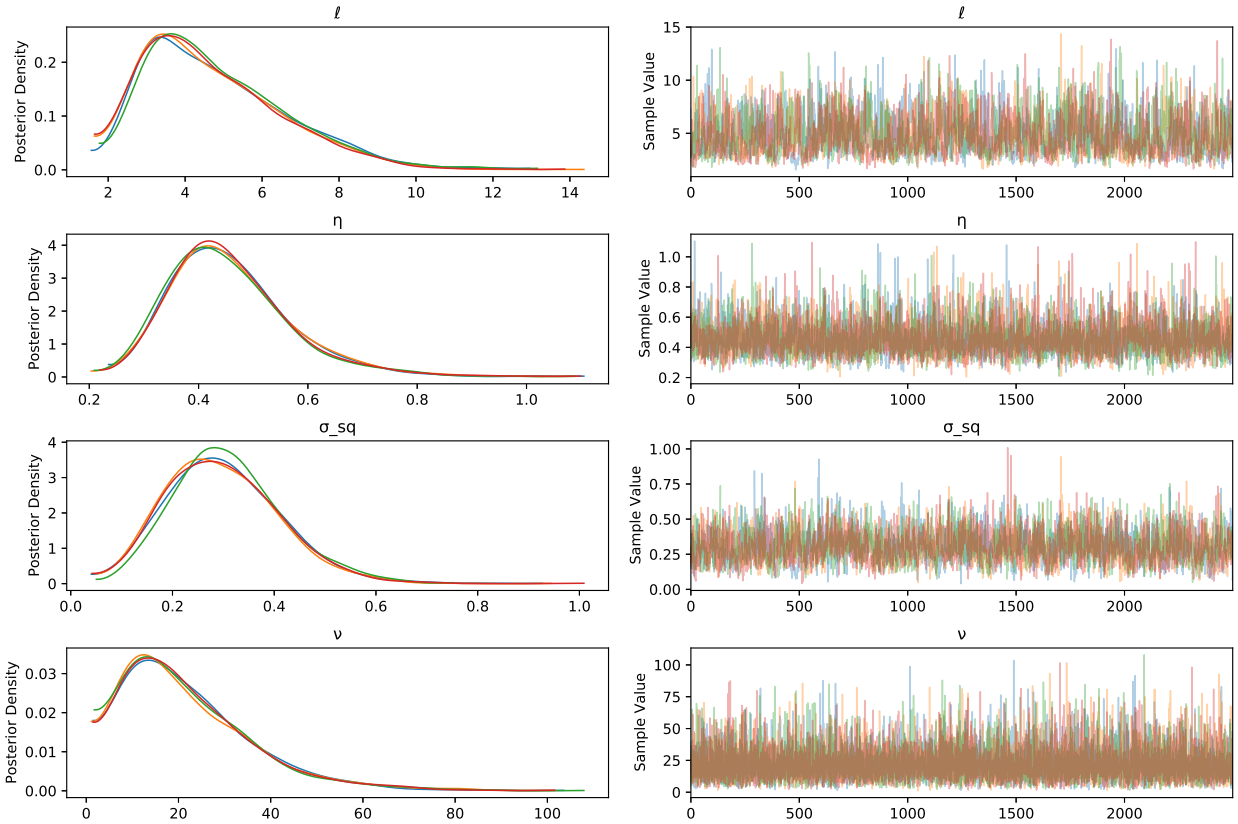


图5.3：运营商A天然气燃烧比例模型的后验分布和轨迹图。已实现良好混合并达到收敛。请注意，这些推断结果与图 5.1 中的结果有所不同，但两者完全基于相同的先验分布和似然，展现了该模型设定的广泛适用性。

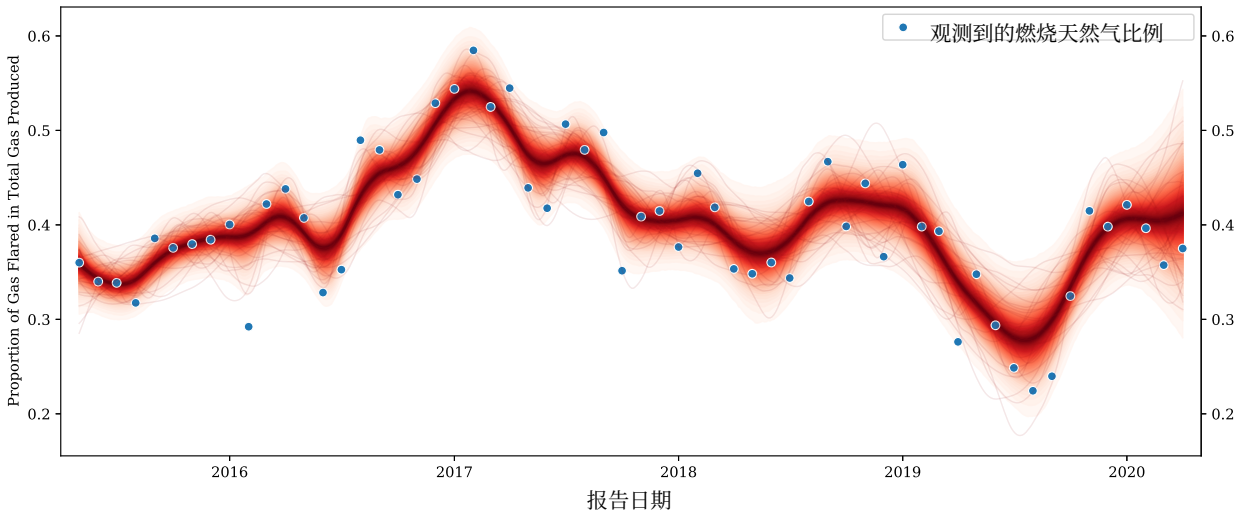


图5.4：显示运营商A天然气燃烧比例变化的后验预测样本。蓝色点表示观测数据，红色线条表示后验样本。

由北达科他州工业委员会制定的第24665号命令，将天然气捕集率 p_{CAP} 定义为

$$p_{\mathrm{CAP}} = \frac{G_{\mathrm{SOLD}} + G_{\mathrm{USED}} + G_{\mathrm{PROC}}}{G_{\mathrm{PROD}}}, \tag{5.13}$$

其中：

G_{SOLD} 表示月售天然气量；

G_{USED} 表示租区月用气量；

G_{PROC} 为月处理天然气量；

G_{PROD} 为月产天然气量。

由于北达科他州禁止天然气放空（美国能源部 2019年b），显然，本节开发的模型为 NDIC 提供了一个强大的工具，可用于评估天然气捕集目标的达成情况：在给定月份 i，$p_{\mathrm{CAP}} = 1 - \pi_i$。此外，在查看模型设定时，并没有任何特别之处对数据来源和位置信息进行编码。该模型的使用者可以自由地将卫星估算作为观测数据，或将模型应用于二叠纪盆地，并对燃烧比例进行推断。与黑箱模型相比，使用非参数且可解释的模型具有这一优势

（例如神经网络，其网络内部学习到的权重和偏差几乎无法提供领域洞见）。作者希望本节结合真实数据，全面阐述如何以及为何使用高斯过程。后续章节构建和介绍的模型也遵循类似的流程。

5.3.2 油井放空燃烧比例建模

一个月内进行火炬燃烧的油井比例可以反映公司的燃除策略，也是燃烧放空量的一项指标。研究某个实体在特定时间段内这一指标的变化情况颇有意义。模型由表达式 5.14a–5.14f 给出：

$$\ell \sim \mathrm{Gamma}(2, 1) \tag{5.14a}$$
$$\eta \sim \text{Half-Cauchy}(5) \tag{5.14b}$$
$$k = \eta^2 \times k_{\text{MATÉRN52}}(x, x'; \ell) \tag{5.14c}$$
$$f \sim \mathcal{GP}(0, k) \tag{5.14d}$$
$$p_i = \mathrm{logit}^{-1}\big(f(x_i)\big) \tag{5.14e}$$
$$W_i \sim \mathrm{Binomial}(N_i, p_i) \tag{5.14f}$$

其中， p_i 是月份 i内进行火炬燃烧的油井“真实”比例（不可观测）， N_i是月份 i内活跃井的总数， W_i 是月份 i内进行火炬燃烧的油井观测数量（即由公司估算并报告的数量）。其余符号的含义与模型 5.12 中相同。

为展示该模型在实际数据上的能力，我们对 Blue Buttes油田和运营商A进行了测试。对于该油田，超参数的后验分布和轨迹图见图 5.5。井燃烧比例 (p_i) 潜在过程的后验预测样本见图 5.6。可视化策略（不同颜色代表不同百分位数等）与第 5.3.1 节相同。

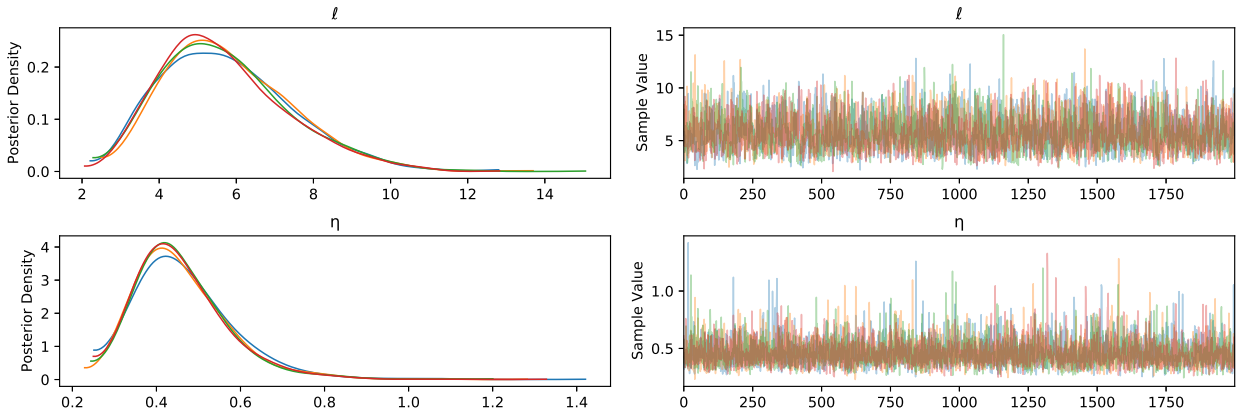


图 5.5：Blue Buttes油田油井燃烧比例模型的后验分布和轨迹图。模型已实现良好混合并收敛。

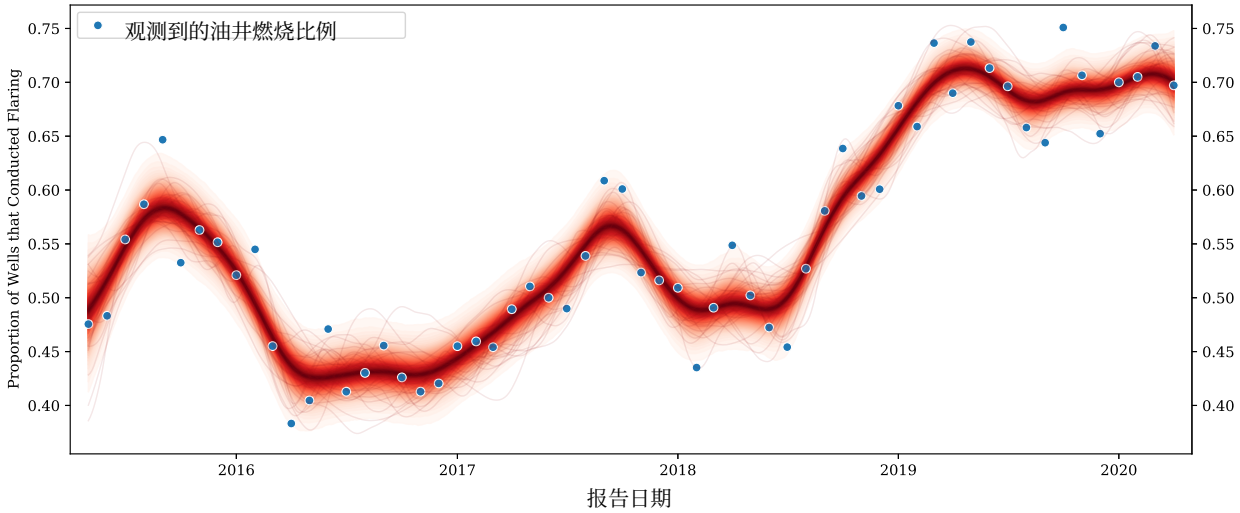


图 5.6：后验预测样本显示了 Blue Buttes油田油井燃烧比例的变化。蓝色点表示观测数据，红色线条表示后验样本。

采用完全相同的模型设定，我们也使用运营商的数据对该模型进行了测试。超参数的后验分布和轨迹图见图 5.7。井燃烧比例 (p_i) 潜在过程的后验预测样本见图5.8。对比油田和运营商的两组图，可以发现：

1. 在长度尺度 ℓ上设置相同的先验后，油田模型从数据中学习，得到的后验众数约为 5.5；而运营商模型得到的后验众数约为 10.0。这一点也反映在后验样本时间序列图中：油田的井燃烧比例在相对较短的时间段内发生了一些变化，而运营商则经历了跨度更长的变化。

2. 油田后验样本时间序列图中的百分位带较窄，而运营商图中的百分位带较宽。这是因为此处选取的运营商拥有的油井数量少于油田。由于每个月的火炬燃烧油井数量均采用二项观测模型，这自然地体现并量化了不确定性（即，二元数据所包含的信息较少，尤其是在样本量较小时），也符合这样的预期：数据越多，不确定性越小；数据越少，不确定性越大。

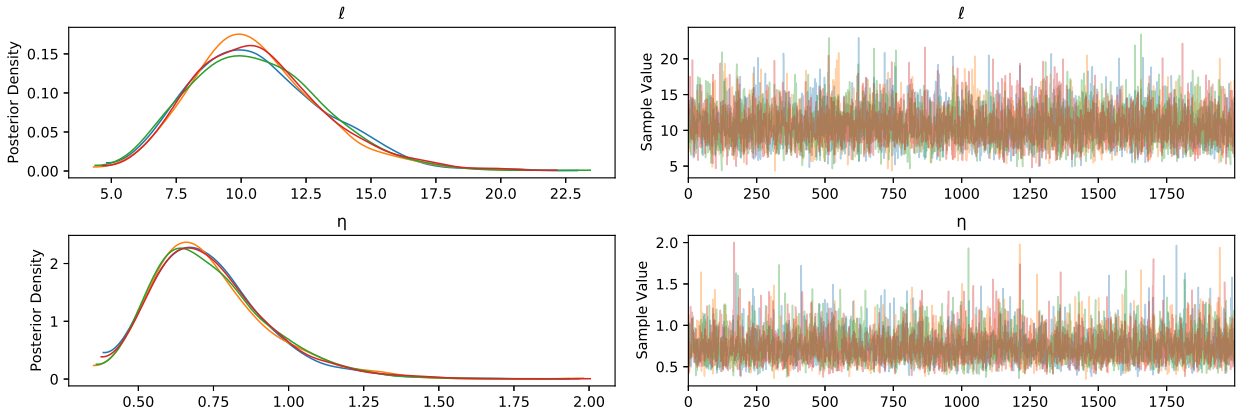


图 5.7：运营商A油井燃烧比例模型的后验分布和轨迹图。油井混合良好，且已达到收敛。请注意，这些推断结果与图 5.5 中的结果有所不同，而两者所依据的先验分布和似然完全相同，这表明该模型设定具有广泛的适用性。

这充分展示了如何以及为何在利用机器学习模型的同时，将领域专业知识融入燃烧数据分析，这也是选择贝叶斯方法的原因。可以使用目标值 $W_i \in \mathbb{R}$ 拟合一个黑箱模型，也可以完全不采用概率视角（例如，在假设空间中优化出最佳的确定性函数映射

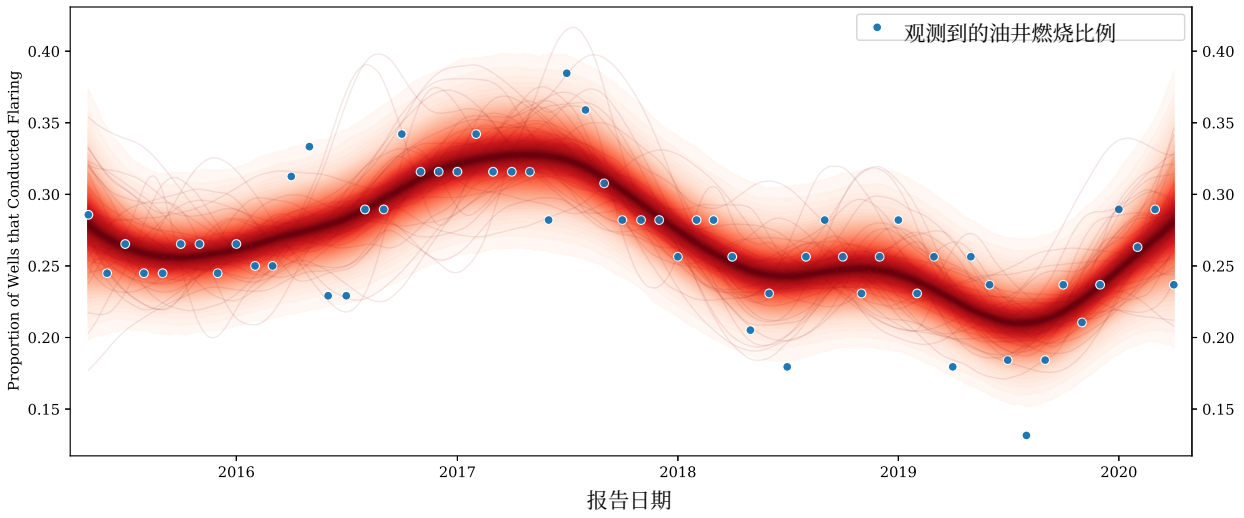


图5.8：展示运营商A油井燃烧比例变化的后验预测样本。蓝色点表示观测数据，红色线条表示后验样本。

）。但这两种做法从根本上来说都有缺陷。领域专业知识表明，油井数量必须是非正整数，即 $W_i \in \mathbb{N}_0$。此外，NDIC报告和卫星估算都不可能在无噪声环境中生成，因此必须进行概率建模。与频率学派机器学习相比，贝叶斯学习完全是概率性的，并赋予人们融入自身领域专业知识的能力与自由。

5.3.3 火炬检测次数建模

卫星检测到的燃烧次数是燃烧强度的无偏指标。了解这一指标在某个时间段内如何随特定实体而变化，具有重要价值。该模型通过表达式 5.15a–5.15f 进行定义。本质上，潜在过程被建模为高斯 Cox 过程（Adams 等人，2009年），其中泊松过程的强度随时间域变化，并且该强度服从高斯过程先验。

$$\ell \sim \text{Gamma}(2, 1) \tag{5.15a}$$

$$\eta \sim \text{Half-Cauchy}(5) \tag{5.15b}$$

$$k = \eta^2 \times k_{\text{MATÉRN52}}(x, x'; \ell) \tag{5.15c}$$

$$f \sim \mathcal{GP}(0, k) \tag{5.15d}$$
$$\lambda_i = \exp\big(f(x_i)\big) \tag{5.15e}$$
$$C_i \sim \text{Poisson}(\lambda_i) \tag{5.15f}$$

其中， λ_i 表示月份 i 中未观测到的燃烧强度（“真实”计数）， C_i 表示月份 i中报告的 VIIRS 检测次数。由于 λ_i 被限定为正值，因此对潜在过程应用自然指数函数。其余符号的含义与模型 5.12 相同。

对于耀斑模式识别任务，作者认为，这种方法（利用高斯 Cox 过程）比（ Davidson-Pilon 2015年；Salvatier 等人 2016；Stan 开发团队 2020年）中提出的一种流行的变点模型更适合作为替代方案，该模型定义如下：

$$e \sim \text{Exponential}(r_e) \tag{5.16a}$$
$$l \sim \text{Exponential}(r_l) \tag{5.16b}$$
$$s \sim \text{Uniform}(1, T) \tag{5.16c}$$
$$C_i \sim \text{Poisson}(i < s \; ? \; e \; : \; l) \tag{5.16d}$$

其中， e 和 l 分别是早期和后期的速率， r_e 和 r_l 控制早期和后期速率的先验分布， s 是变点， T 是总时间段，而泊松似然中的速率则通过三元条件运算符（?:）确定。原因在于，尽管该模型可以推广到多个变点，但其应用受到这样一个假设的限制：任意两个相邻变点之间的时间段内，速率都是恒定的。在下文讨论实际燃烧数据时，这一局限性显而易见，也是变点模型的一大缺点。

采用 Blue Buttes油田的数据对高斯 Cox 过程模型进行检验。由于仅使用 VIIRS 数据，因此从 2012年 开始分析整个时间序列。超参数的后验分布和轨迹图见图 5.9。潜在过程的放空燃烧次数 (C_i) 的后验预测样本展示于

图 5.10 中。可视化策略（不同颜色代表不同百分位数等）与第 5.3.1 节相同。从时间序列图可以看出，2014年到 2017 的观测值可能可以用变点模型来描述（2015年末可能是一个变点），但这一时间段前后持续增长的趋势会使得使用此类模型进行准确推断变得困难。

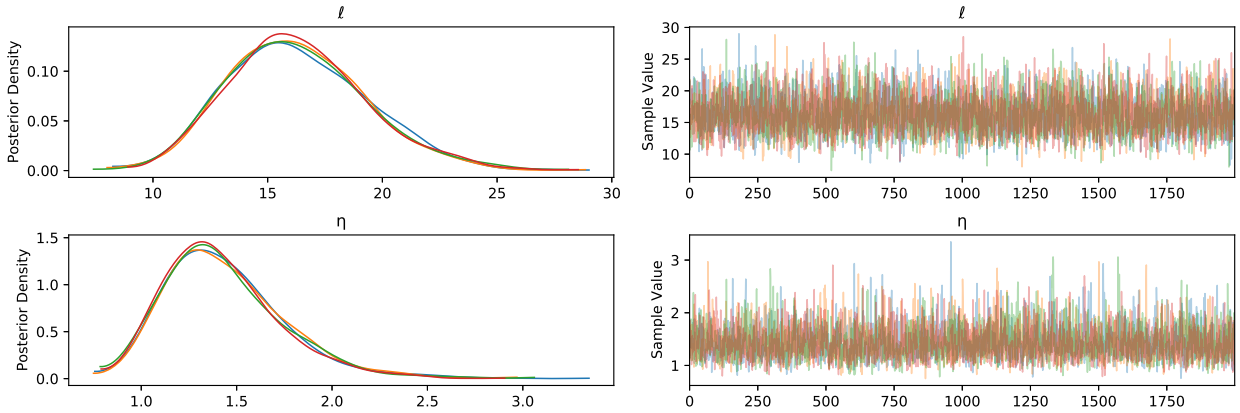


图 5.9：Blue Buttes油田燃除计数模型的后验分布和轨迹图。已实现良好混合并达到收敛。

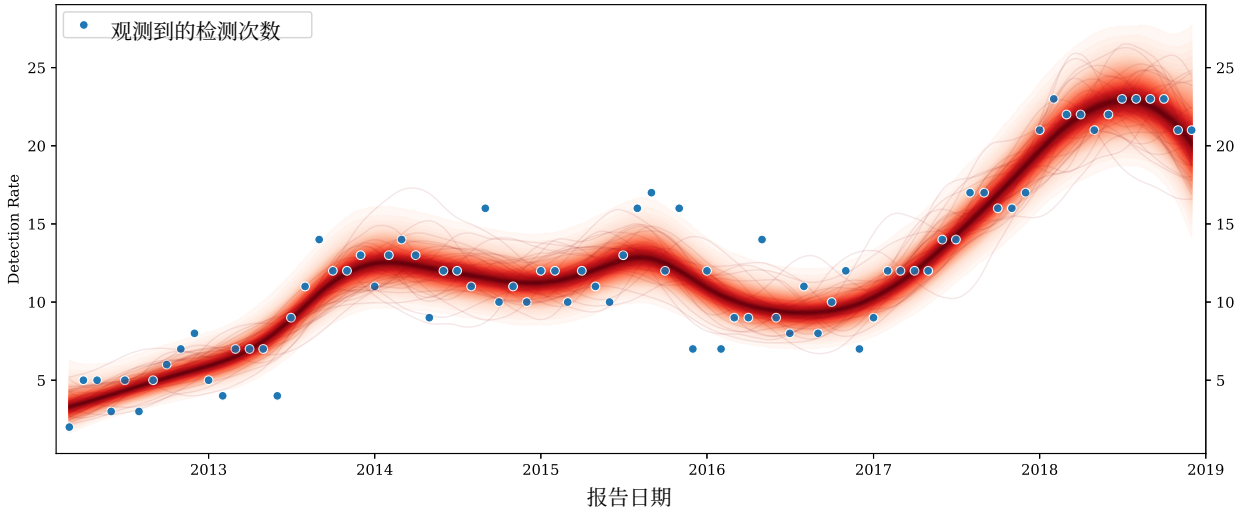


图 5.10：后验预测样本展示了 Blue Buttes油田的燃烧计数变化。蓝色点表示观测数据，红色线条表示后验样本。

该模型的推断结果可作为一种验证（即使不能算作证据），用于判断某个实体是否实现了减少放空燃烧油井数量这一目标；在此处，检测计数被用作放空燃烧油井数量的替代指标。在实践中，减少放空燃烧油井的数量正是北达科他州工业委员会于2014年提出的监管政策的第二项目标。如果州政府希望从宏观角度了解这项命令的有效性，也可以使用该模型对州级数据进行推断。在这种情况下，超参数的后验分布和轨迹图见图5.11。潜在过程中放空燃烧次数（C_i）的后验预测样本见图5.12。

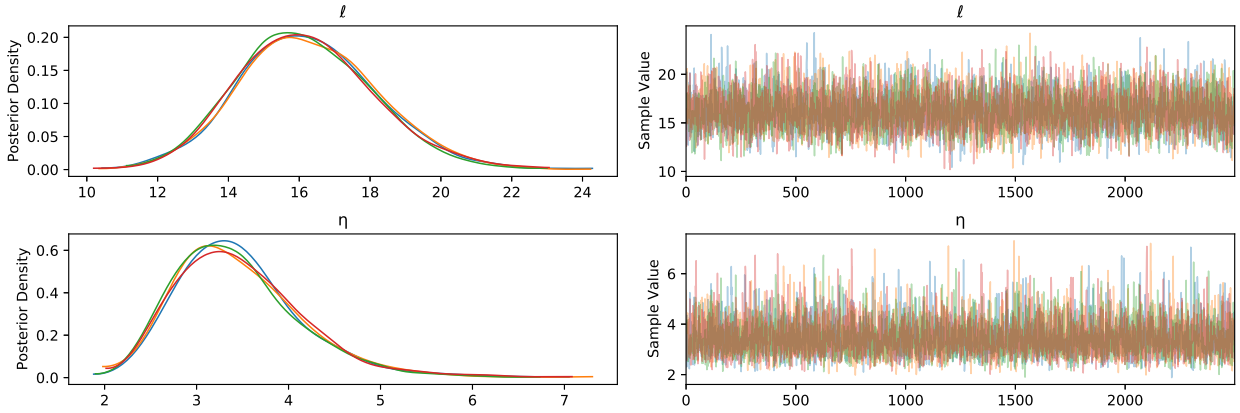


图 5.11：北达科他州放空燃烧次数模型的后验分布和轨迹图。模型已实现良好混合与收敛。请注意，这些推断结果与图 5.9 中的结果有所不同，尽管二者完全基于相同的先验分布和似然，这体现了该模型设定的广泛适用性。

本例中的百分位带相当狭窄，这表明在给定模型假设的情况下，我们对数据生成过程的推断更有信心。通过不对每一个观测都进行（过度）拟合，我们发现了有趣的模式，例如，每年都恰好只有一个峰值，出现在六月左右。值得指出的是，没有任何模型能告诉建模者其假设是否合理，只有领域专业知识或许能做到这一点。由于泊松似然只有一个参数 λ （用于

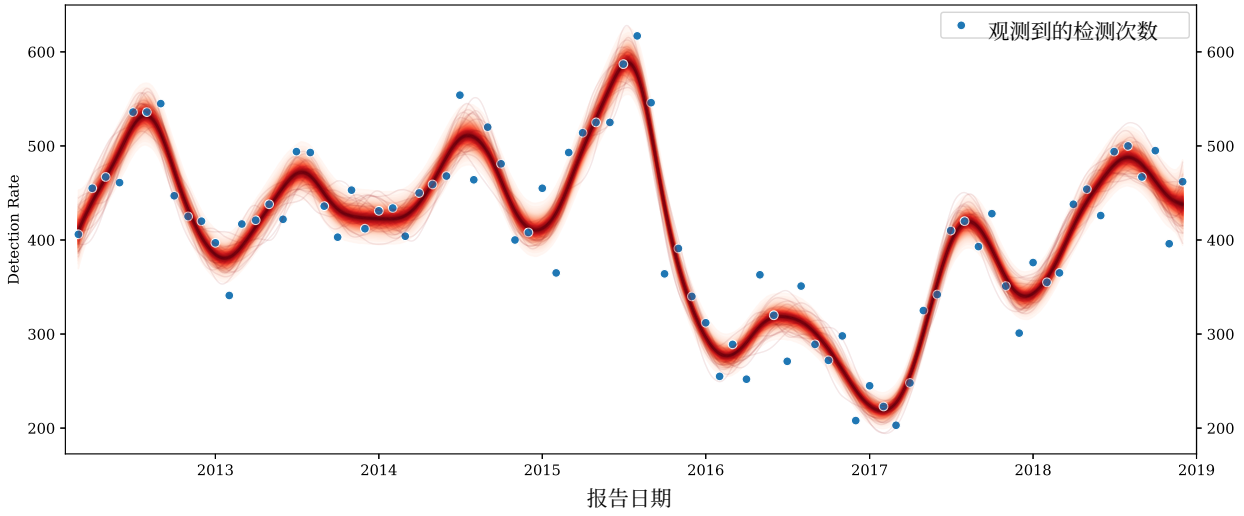


图5.12：展示北达科他州放空燃烧次数变化的后验预测样本。蓝色点为观测数据，红色线条表示后验样本。

同时控制均值和方差），而且当 λ 像本例中这样较大时，泊松分布可由正态分布很好地近似，因此采用泊松观测模型的模型可被视为“僵化”。如果州政府认为可能存在过度离散，则可以考虑其他观测模型，例如负二项分布。在这种情况下，只需将表达式5.15f改为负二项似然，并添加过度离散参数的先验。具体参数化方式见第6.3节中的公式6.4。这充分展现了在包括火炬燃烧数据分析在内的高风险决策领域采用贝叶斯方法所具有的灵活性和可解释性。

5.3.4 火炬燃烧比例建模

由于此时的主要商品是原油（而非天然气），因此，火炬燃烧的每桶油当量（BOE）所含天然气量可作为衡量火炬燃烧导致的生产效率的指标。在本研究中，采用归一化量——火炬燃烧的石油产量占比，使模型设定适用于大型和小型实体。模型通过表达式5.17a–5.17j进行设定：

$$\ell \sim \text{Gamma}(2, 1) \tag{5.17a}$$

$$\eta \sim \text{Half-Cauchy}(5) \tag{5.17b}$$
$$\nu \sim \text{Gamma}(2, 0.1) \tag{5.17c}$$
$$\hat{\sigma}^2 \sim \text{Half-Cauchy}(5) \tag{5.17d}$$
$$k = \eta^2 \times k_{\text{MATÉRN52}}(x, x'; \ell) \tag{5.17e}$$
$$f \sim \mathcal{GP}(0, k) \tag{5.17f}$$
$$\pi_i = \text{logit}^{-1}\big(f(x_i)\big) \tag{5.17g}$$
$$\mu_i = \pi_i \times O_i \tag{5.17h}$$
$$c = \frac{6\,\text{Mcf}}{1\,\text{BOE}} \tag{5.17i}$$
$$F_i/c =: E_i \sim \text{Student-}t(\nu, \mu_i, 1/\hat{\sigma}^2) \tag{5.17j}$$

其中：

> π_i 表示 i月潜在的桶油当量燃烧比例；O_i 表示 i月的石油总产量；μ_i 表示 i月“真实”的燃烧桶油当量；c 表示换算系数，即 6 千立方英尺等于 1 桶油当量，该系数由美国
>
> 地质调查局（2000）给出；

E_i 表示报告的燃烧桶油当量，并使用 Student-t 观测模型进行建模。其余符号的含义与模型 5.12 相同。为检验该模型在真实数据上的表现，采用了 Blue Buttes油田和运营商A的数据。对于该油田，超参数的后验分布和轨迹图见图 5.13。潜在过程的桶油当量燃烧比例（π_i）的后验预测样本如图 5.14 所示。可视化策略（不同颜色代表不同百分位数等）与第 5.3.1 节相同。

使用完全相同的模型设定，也用该运营商的数据对这个模型进行了检验。超参数的后验分布和轨迹图见图 5.15。BOE燃烧比例（π_i）潜在过程的后验预测样本见图 5.16。

比　　较油田和运营商的两组图，可以观察到　　：

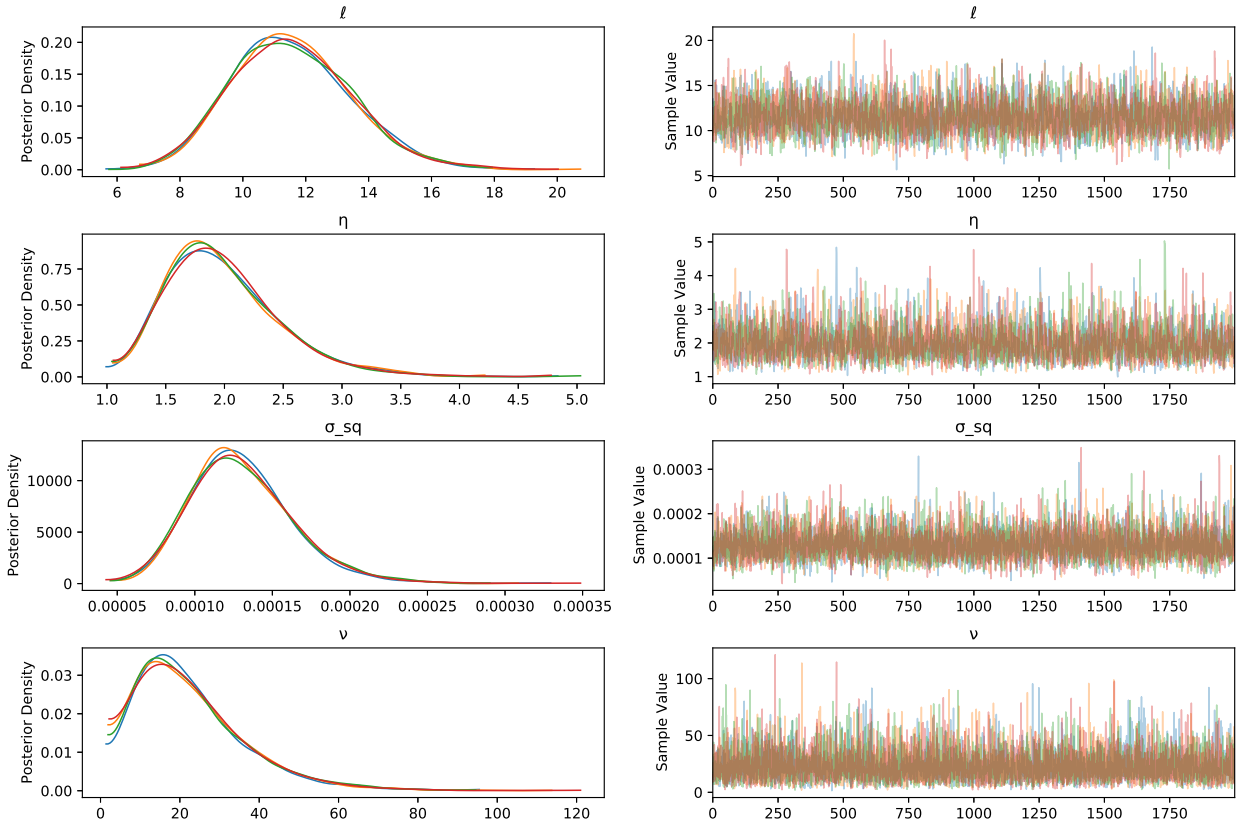


图 5.13：Blue Buttes油田 BOE 放空比例模型的后验分布和轨迹图。已实现良好混合并收敛。

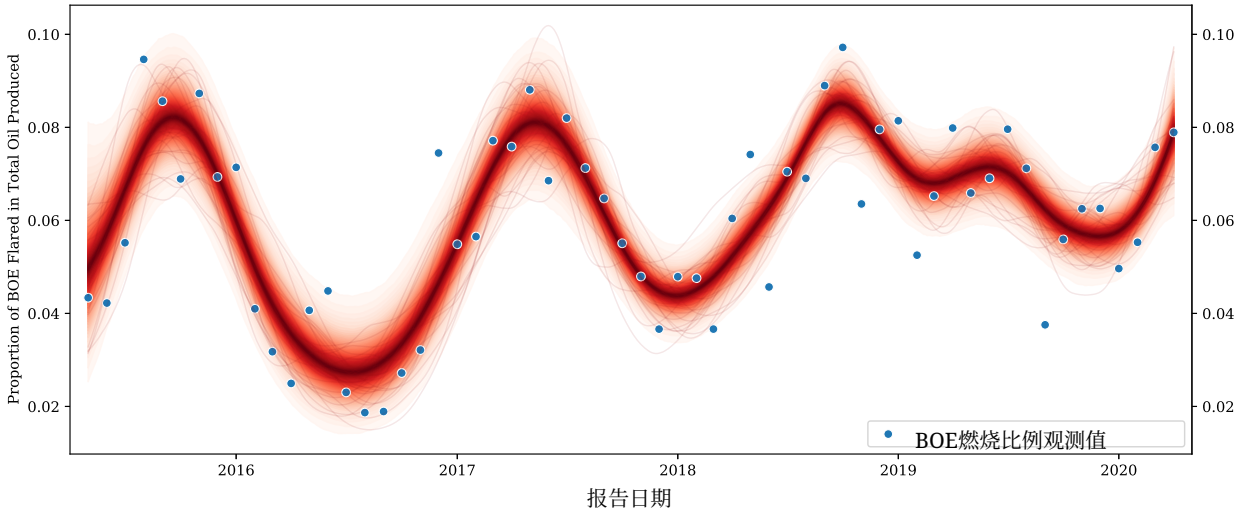


图 5.14：显示 Blue Buttes油田 BOE 火炬燃烧比例变化的后验预测样本。蓝色点表示观测数据，红色线条表示后验样本。

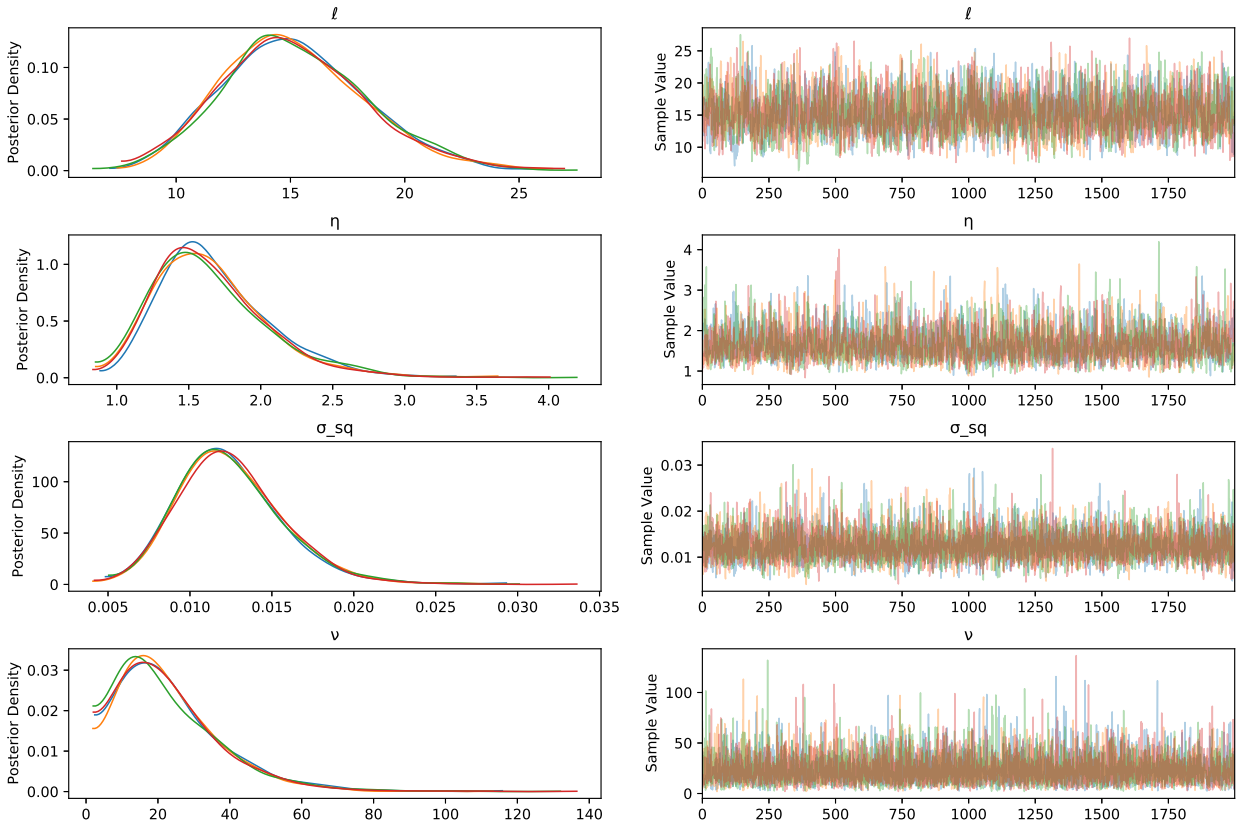


图 5.15：运营商A的 BOE 放空比例模型的后验分布和轨迹图。已实现良好混合和收敛。请注意，这些推断结果与图 5.13 中的结果有所不同，尽管两者完全基于相同的先验分布和似然，这体现了该模型设定的广泛适用性。

1. 对长度尺度 ℓ 采用相同的先验分布，其均值为 2 个月后，两个模型更新后的后验分布都偏离了该均值，反映出长程变化。油田的后验众数约为 1 年，而运营商的众数约为 15 个月。运营商的报告变异性要大得多，这一点由参数 $\hat{\sigma}^2$ 显示。

2. 使用 Student-t 似然时，两种模型都表现出对异常值和过拟合的稳健性。这一点可以从油田 2019年末的观测值以及运营商 2016年初的观测值中看出。对于以细线表示的后验函数样本，其中一些确实会被拉向那些“异常值”。然而，百分位数图

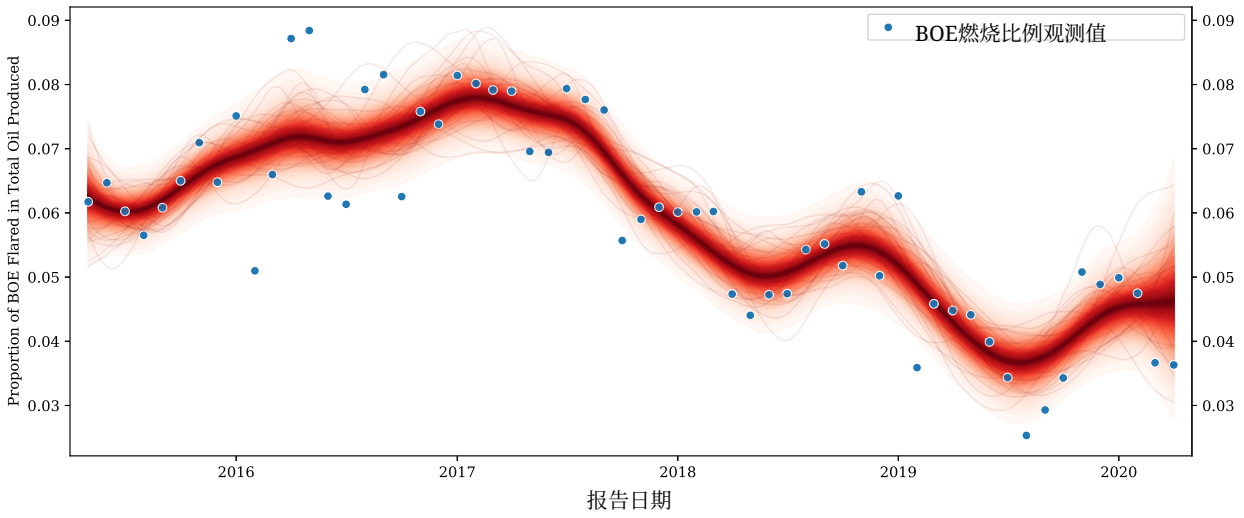


图 5.16：展示运营商A的 BOE 火炬燃烧比例变化的后验预测样本。蓝色点为观测数据，红色线条表示后验样本。

（以彩色带表示）并未受到影响，而且确实可以将其解读为与数据和假设最相符的趋势。贝叶斯方法在选择合适的先验分布时内置的这种奥卡姆剃刀效应令人印象深刻。在许多频率学派机器学习方法中，如果正则化策略实施得不好，尤其是样本量不足以使渐近性质发挥作用时，异常值就会变成“影响点”，对推断结果产生巨大的负面影响。

5.3.5 VIIRS 与 NDIC 之间的比例因子建模

NDIC 和 VIIRS 报告均提供火炬燃烧气体体积（估算值）。这两个来源之间的比例因子有助于了解 NDIC 报告是否一致：

1. 对于不同的实体（例如，一组运营商中的各个运营商），以及

2. 针对某一实体考察特定时间段时。

这是基于以下事实：卫星探测处理算法是无偏且一致的。就时间序列分析而言，第 2 项尤其值得关注。模型通过表达式 5.18a–5.18n 指定：

$$\ell_{\text{MAT}} \sim \text{Gamma}(8, 2) \tag{5.18a}$$
$$\eta_{\text{MAT}} \sim \text{Half-Cauchy}(5) \tag{5.18b}$$
$$T \sim \mathcal{N}(12, 1) \tag{5.18c}$$
$$\ell_{\text{PER}} \sim \text{Gamma}(4, 3) \tag{5.18d}$$
$$\eta_{\text{PER}} \sim \text{Half-Cauchy}(5) \tag{5.18e}$$
$$\nu \sim \text{Gamma}(2, 0.1) \tag{5.18f}$$
$$\hat{\sigma}^2 \sim \text{Half-Cauchy}(5) \tag{5.18g}$$
$$k_{\text{MAT}} = \eta_{\text{MAT}}^2 \times k_{\text{MATÉRN52}}(x, x'; \ell_{\text{MAT}}) \tag{5.18h}$$
$$k_{\text{PER}} = \eta_{\text{PER}}^2 \times k_{\text{PERIODIC}}(x, x'; T, \ell_{\text{PER}}) \tag{5.18i}$$
$$k_{\text{WN}} = k_{\text{WhiteNoise}}(x, x'; \delta = 1\text{e}{-6}) \tag{5.18j}$$
$$f \sim \mathcal{GP}(0, k_{\text{MAT}} + k_{\text{PER}} + k_{\text{WN}}) \tag{5.18k}$$
$$\beta_i = \exp\big(f(x_i)\big) \tag{5.18l}$$
$$\mu_i = \beta_i \times \text{VIIRS}_i \tag{5.18m}$$
$$\text{NDIC}_i \sim \text{Student-}t(\nu, \mu_i, 1/\hat{\sigma}^2) \tag{5.18n}$$

其中：

ℓ_{MAT} 表示 Mat´ern 核的长度尺度；η_{MAT} 表示 Mat´ern 核的边际偏差；T 表示周期核的周期；ℓ_{PER} 表示周期核的长度尺度；η_{PER} 表示周期核的边际偏差；k_{MAT} 表示 Mat´ern 核（分量）；k_{PER} 表示周期核（分量）；k_{WN} 表示白噪声核（分量）；f 表示潜在过程，该过程服从一个高斯过程，其协方差函数为 3 个核函数之和；β_i 表示月份 iVIIRS 与 NDIC 之间的底层比例因子。由于该比例因子被限制为正值，因此将自然指数函数应用于

潜在过程；

VIIRS$_i$ 是第 i月的 VIIRS 报告量；

μ_i 表示第 i月的底层放空气量；

NDIC$_i$ 是第 i月的 NDIC 报告量，该报告量使用 Student-t观测模型进行建模。

其余符号的含义与模型 5.12 中相同。添加周期核的目的是探究是否存在任何季节性模式。坚持规范的贝叶斯工作流程可以让数据自行说明情况，即是否存在季节性行为，如本节的两个案例研究所示。

首先使用州级数据拟合模型，以探究宏观报告一致性。超参数的后验分布和轨迹图见图5.17。比例因子变化潜在过程（β_i）的后验预测样本见图5.18。可视化策略（不同颜色代表不同百分位数等）与第 5.3.1 节相同。从后验时间序列图总体来看，NDIC报告的体积小于VIIRS报告的体积，只有在总燃烧量较小的时期例外（以较小的点表示）。更重要的是，在2015年至2018年的每一年中，比例因子（β_i）的数值都在年中左右呈下降趋势。每年从第二季度到第三季度的潜在过程都可以看作一个“跷跷板”：七月是中间支点，而七月之后的月份数值总是下降。请注意，在每一年中，NDIC报告的燃烧量可能会稳步或大幅增加（事实上，从图3.7的时间序列图可以看出，这种情况确实发生了）；然而，比例因子的下降趋势表明，卫星观测到的燃烧活动远高于公司报告的量！这一发现表明，NDIC报告很可能全年并不一致，州政府应当关注一些公司可能少报其放空气量，尤其是在下半年。

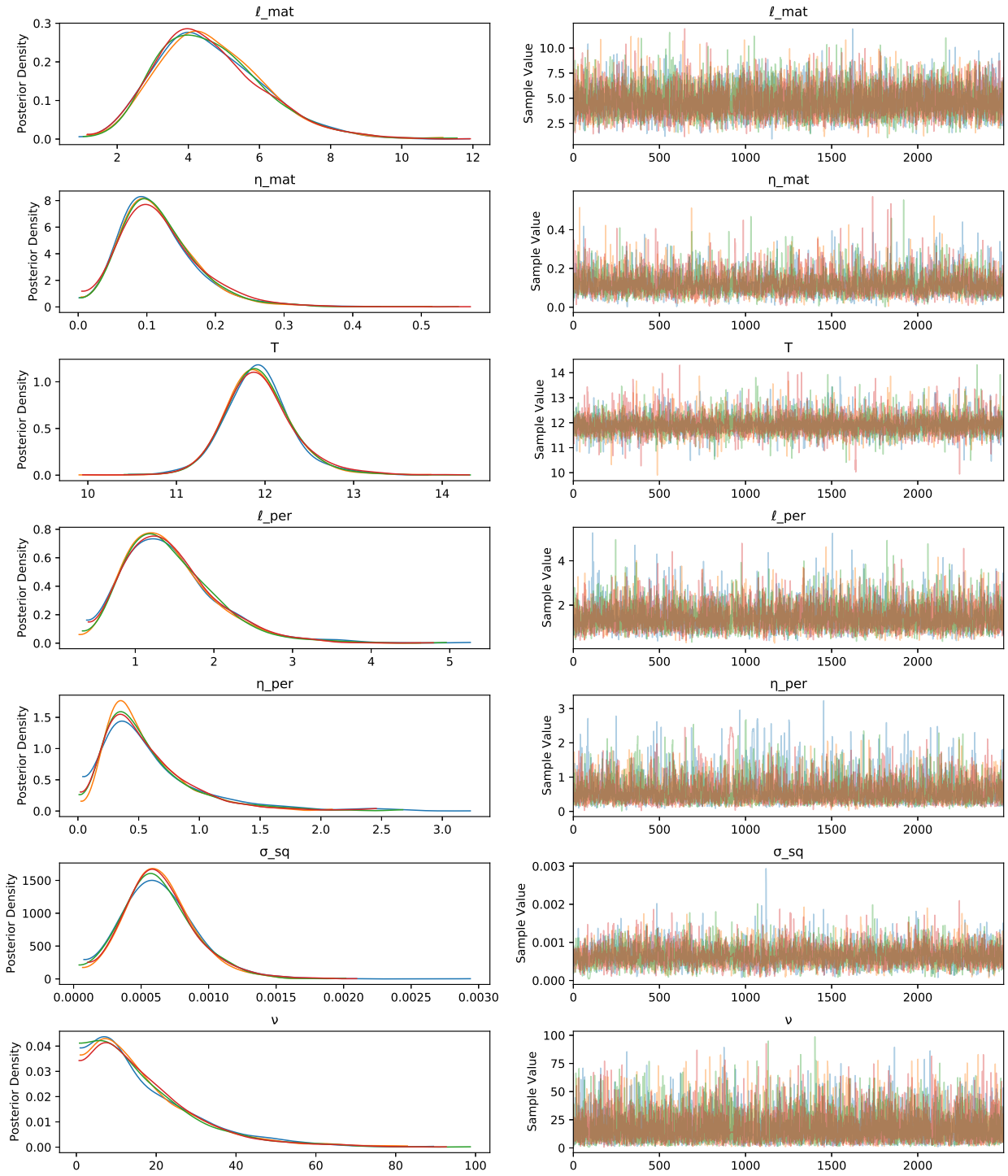


图5.17：北达科他州 VIIRS-NDIC 比例因子模型的后验分布和轨迹图。已实现良好混合并达到收敛。

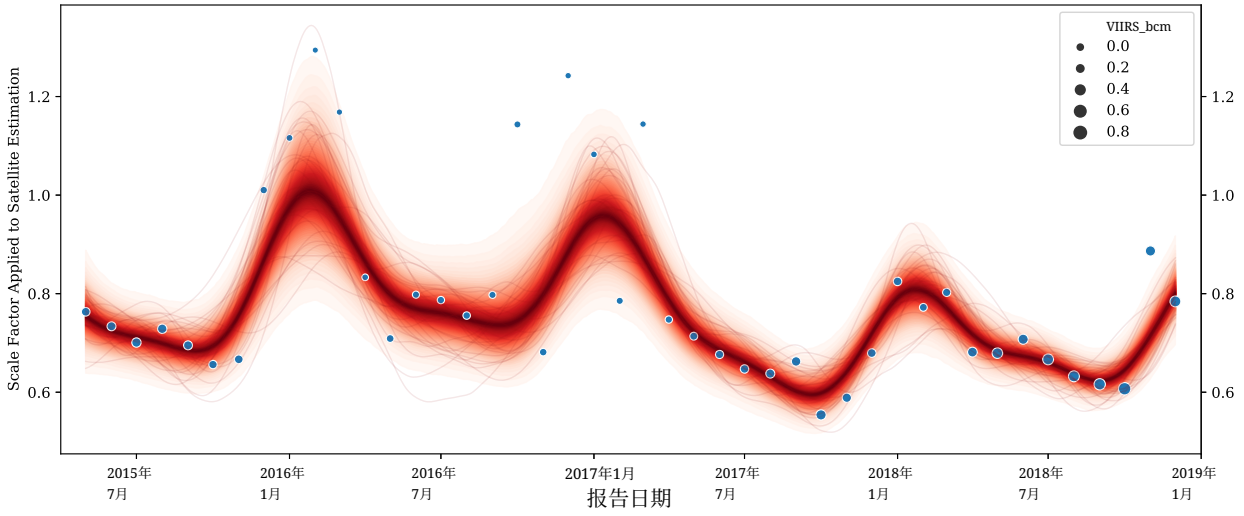


图 5.18：后验预测样本展示了北达科他州比例因子的变化。蓝色点表示观测数据，红色线条表示后验样本。点越大，表示从 VIIRS 观测到的燃烧放空量越大。

一个有趣的问题随之而来：这种季节性行为是否普遍存在于所有实体中？遗憾的是，答案是否定的，这表明一些运营商很可能在全年以不一致的方式报告其燃烧气体体积。事实上，如果使用 Blue Buttes油田的数据来拟合模型，则会观察到相当一致的行为。在这种情况下，超参数的后验分布和轨迹图见图 5.19。潜在过程的比例因子变化（β_i）的后验预测样本如图5.20所示。采用完全相同的模型设定并纳入周期核后，推断过程未发现明显的季节性行为。2016年初前后存在很大的不确定性；点的大小表明，从 VIIRS 观测到的总体火炬燃烧放空量较小，而 NDIC报告量实际上高于 VIIRS 的报告量。这可能是由截断效应而非报告不一致造成的；也就是说，当火炬燃烧零星且较弱时，卫星不容易捕捉到，导致 VIIRS处理流程产生截断样本。将此模型和工作流应用于其他主要产区，很可能会识别出那些报告中呈现“跷跷板”行为的产区。

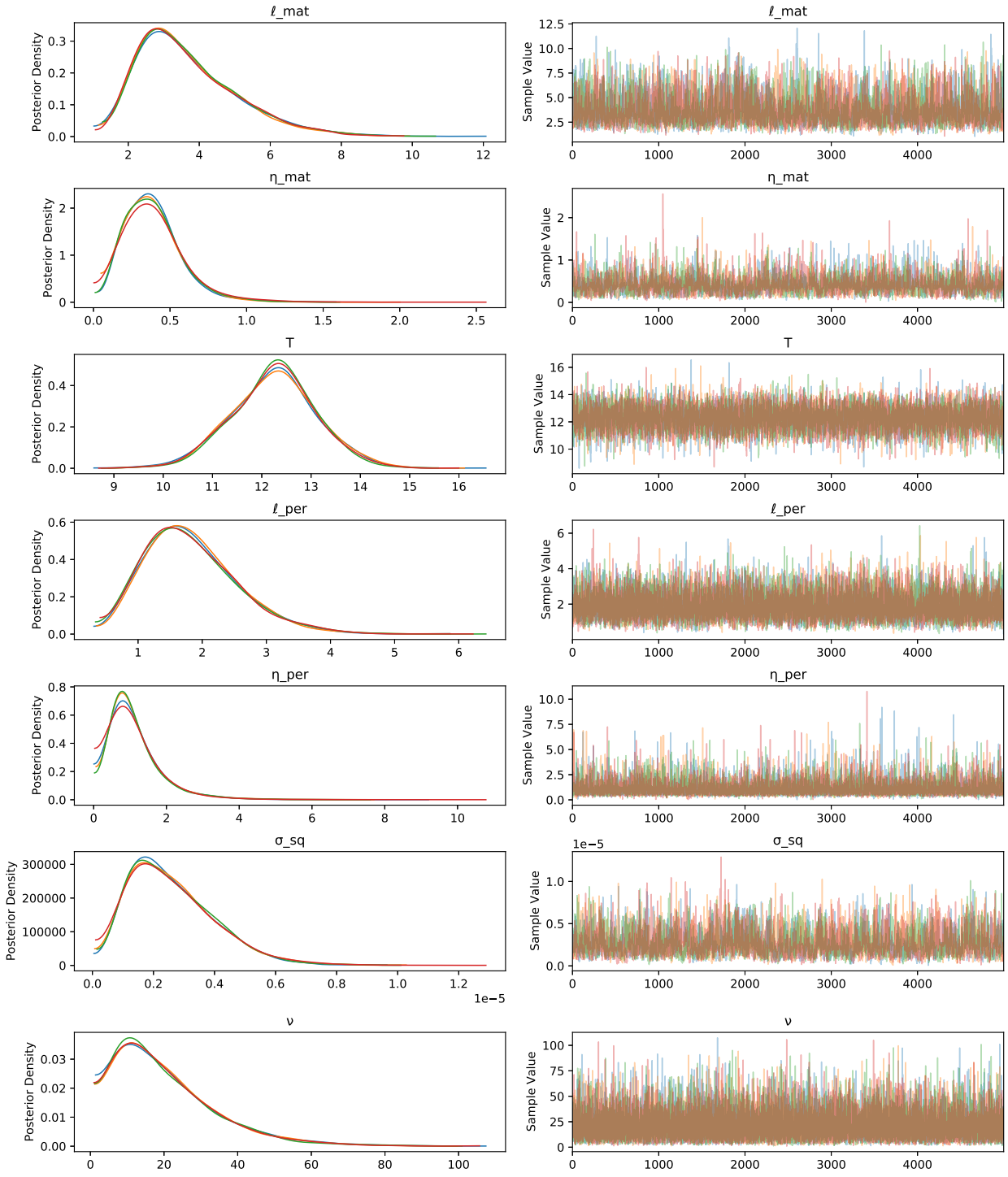


图 5.19：Blue Buttes油田 VIIRS-NDIC比例因子模型的后验分布和轨迹图。模型已实现良好混合并收敛。请注意，这些推断结果与图5.17中的结果有所不同，而两者基于完全相同的先验分布和似然，充分体现了该模型设定的广泛适用性。

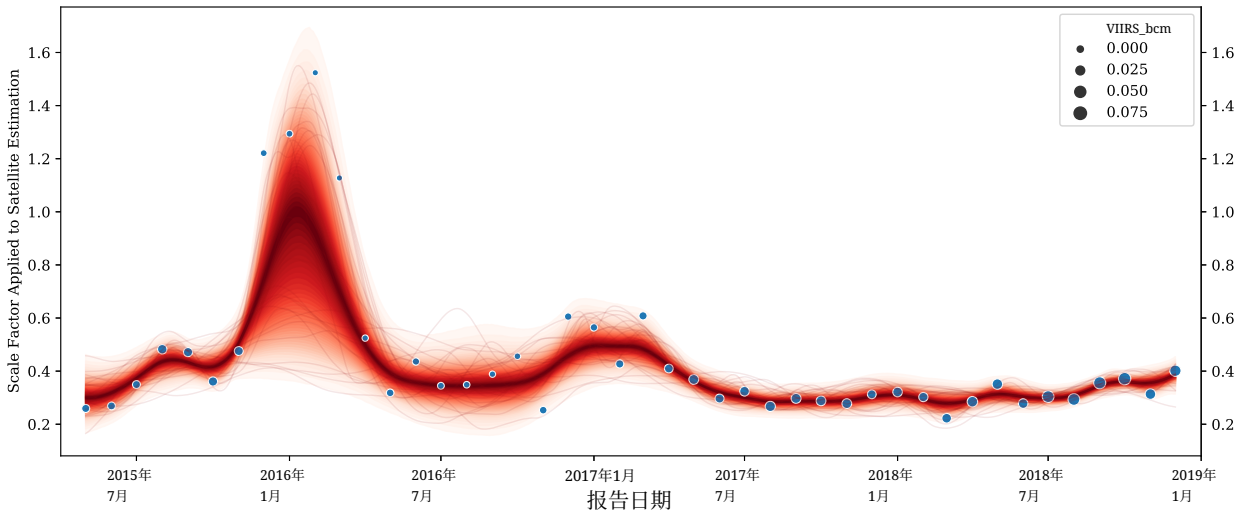


图5.20：展示 Blue Buttes油田比例因子变化的后验预测样本。蓝色点为观测数据，红色线条表示后验样本。点越大，表示 VIIRS 观测到的火炬燃烧量越大。

5.3.6 预测 NDIC 火炬燃烧量

高斯过程不仅在学习到模型超参数后能够充分进行预测，还可以通过表达式5.11严格构建区间，以量化不确定性；许多频率学派机器学习方法则无法做到这一点。作者选择展示一个具体的预测案例研究，即根据 VIIRS 与 NDIC 之间的预期比例因子来预测 NDIC报告量。一旦能够快速进行卫星探测和估计，这将成为一个特别有趣的应用场景，因为其耗时少于等待公司报告并将所有内容整理成适用于分析的格式。

这些预测以后验预测样本的形式生成。图 5.21 展示了历史观测值以及未来六个月比例因子的预测。预测中的百分位带非常宽，这表明季节性行为可能会再次显现，但仍存在很大的不确定性。如果需要点预测（即不包含预测区间），可以使用后验均值、众数等来构建那个“最佳”函数；然而，这也说明了为什么预测

未来通常非常困难，而且应始终恰当地描述不确定性。

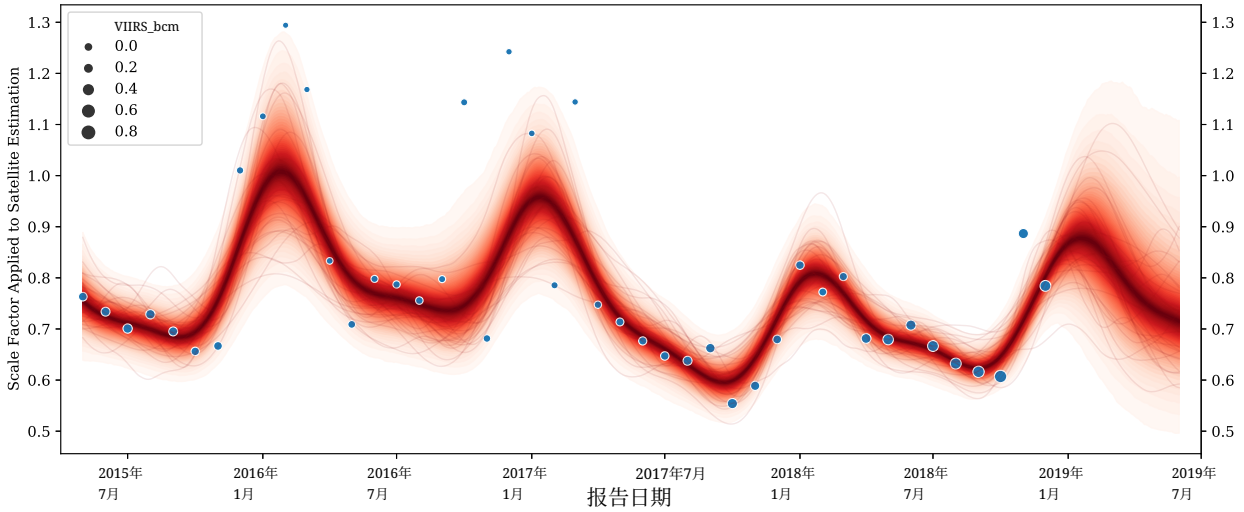


图 5.21：后验预测样本显示了未来六个月比例因子的预测。蓝色点为观测数据，红色线条表示后验样本。点越大，表示从 VIIRS 观测到的火炬燃烧量越大。

5.3.7 回顾先验选择

回顾所构建的模型系列，潜在函数的先验分布集合一直相同（尺度因子模型除外，该模型中加入了周期核）。不过，所有后验分布都会根据每个数据集和建模目标进行更新（即“学习”）。这意味着，下面这组先验分布

$$\ell \sim \text{Gamma}(2, 1) \tag{5.19a}$$
$$\eta \sim \text{Half-Cauchy}(5) \tag{5.19b}$$
$$k = \eta^2 \times k_{\text{MATÉRN52}}(x, x'; \ell) \tag{5.19c}$$
$$f \sim \mathcal{GP}(0, k) \tag{5.19d}$$

可作为通用框架，并可普遍推荐用于高斯过程语境下的燃烧放空时间序列分析。请注意，该先验选择会使无约束空间中的潜在函数值为，即 $f(x) \in \mathbb{R}$。然而，在许多情况下，领域专业知识

表明我们关注的量位于受约束空间中，例如：

- $\mathbf{R}_{>0}$ 用于对计数数据建模时的泊松率参数，且
- $[0, 1]$用于对燃烧井比例建模时的二项成功概率 .

为更好地体现领域专业知识，可以利用链接函数。对于上述情形，分别可以应用对数链接函数和 logit 链接函数。尽管此先验配置经过了多次设计迭代，并使用真实数据进行了测试，但没有理由认为它对每个实体都是最优的。事实上，VIIRS 与 NDIC 之间的比例因子模型在其先验分布中包含定制的成分。Stan 开发团队（2020年）也针对高斯过程给出了一些一般性的先验选择建议。

整套模型充分展现了在不同层面利用不同实体的燃烧放空时间序列中的时间结构的能力。这为从嘈杂的月度数据流中提取见解提供了巨大潜力。对于适合进行横截面数据分析的情况，例如，在最新月份的数据可用时，州政府只需了解该月份的情况（然后再将其追加到全部历史数据中，用于纵向研究），则可以构建其他类型的模型。下一章将对此进行讨论。

第6章

多视角无监督学习

“密度估计是统计学中的一个普遍问题（了解密度便能解决各种问题）。”

— Vapnik（2000）

6.1 学习分布

本章中，作者研究如何在横截面情境下描述北达科他州各油田与火炬燃烧相关量的分布。也就是说，分析在某个时间点或某一段时间（例如某个月份或某个季度）收集的数据。在这种情境下，用于学习的数据没有标签：

$$U = \{x_1, x_2, \ldots, x_N\}, \tag{6.1}$$

其中，x_i、$i = 1$、2、.....、N 是第 i 个油田的观测值。因此，自然会采用无监督学习。要学习的模型采用条件概率分布 $P_{\boldsymbol{\theta}}(x \mid z)$ 的形式，其中 z 是某种潜在结构，而 $\boldsymbol{\theta}$ 表示参数。

这在实践中有许多应用场景。获得最新月份或季度的数据后，北达科他州政府可能需要了解（油田）总体的分布洞见，最好不局限于某些形式的顺序统计量（例如五数概括）。当需要与此前月份/季度进行直接比较（可以是紧邻的前一个月份/季度，也可以是往年同一月份/季度），或需要更深入地了解总体（例如寻找实体之间可能存在的簇）时，这项横截面研究尤其有价值，也值得开展。

6.2 概率模型估计

学习分布是在无监督情境下的概率模型估计问题（Li 2019年）。它有时采用密度估计的形式，一些统计学家认为密度估计是概率机器学习中最基础的主题（Yu 2017）。直方图是一种基础且常用的技术，但很容易被误用，从而导致对数据集产生有偏差的理解（图 6.1）。

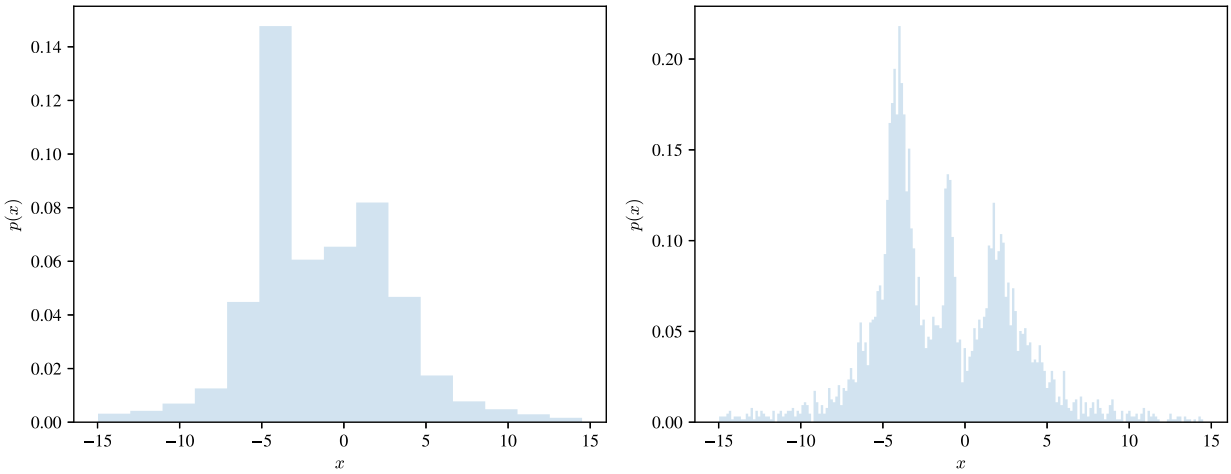


图 6.1：有效运用直方图可能比想象中微妙得多。对于改编自（VanderPlas 等，2012）的完全相同的数据集，组距不同的两个直方图呈现出不同的多峰特征。采用某个软件包的默认配置，只会得到分布的一种视图。

一般而言，假设数据由某个概率模型生成，那么该模型的结构和参数需要从数据中学习。结构的类型，即可能的概率模型集合，通常是给定的（假设的）；而结构的具体形式和参数则需要通过学习确定。目标是找到最有可能生成这些数据的模型结构和参数。

概率模型可以是混合模型，也可以是图模型。在本学位论文中，我们研究混合模型，其假设是数据来自分布混合。从数学上讲，混合模型通过基分布的凸组合来描述分布 $p(x)$：

$$p(x) = \sum_{k=1}^{K} \pi_k p_k(x) \tag{6.2a}$$

$$\sum_{k=1}^{K} \pi_k = 1,\ \pi_k \geq 0, \tag{6.2b}$$

其中， p_k 是混合模型中的成分， π_k 是混合权重。混合模型可以解释为总体由多个不同的子总体组合而成。混合模型也可以泛化到连续情形。例如，负二项分布和 Student t 分布都可以看作是某些连续分布的混合（Martin 2018）。

在模型表示 $P_{\boldsymbol{\theta}}(x \mid z)$ 中， x 表示观测值，可以是离散或连续的量； z 表示潜在结构，它是一个离散随机变量。模型由 $\boldsymbol{\theta}$ 参数化。假设模型属于混合类型时， z表示不同的成分。模型结构和参数的信息是从数据 $U = \{x_1, x_2, \ldots, x_N\}$ 中学习得到的；在本研究中， $x_i \in \mathcal{X} \subseteq \mathbf{R}^1, i = 1, 2, \ldots, N$ 是第 i 个油田的观测值。

6.3 VIIRS检测次数建模

在第 5.3.3 节中，我们提出了分析任意给定油田 VIIRS 检测次数时间序列的方法。本节探讨如何从北达科他州油田任意月份的火炬检测次数中提取洞见。具体而言，通过学习各油田的检测计数，可以概括油田总体的情况，州政府由此能够获得分布洞见。

按照第 6.2 节中的一般形式，这个问题变成了一个特殊情况：潜在结构 z 不存在，即满足 $P_{\boldsymbol{\theta}}(x \mid z) = P_{\boldsymbol{\theta}}(x)$，其中 x 表示检测计数。此时，估计条件概率分布就变成了估计概率分布，因此只需估计 $P_{\boldsymbol{\theta}}(x)$ 的参数即可。经典统计学中的密度估计，例如高斯参数

估计，就是此类情形的一个例子。

由于对计数数据进行了建模，作者将下列四种观测模型与随机选取的多个月份的数据进行比较：

1. 泊松似然

2. 负二项似然

3. 零膨胀泊松（ZIP）似然

4. 零膨胀负二项（ZINB）似然

之所以对上述第 3 和第 4 种模型进行检验，是因为北达科他州的许多油田在某个月份没有 VIIRS 检测记录。因此，也尝试了零膨胀模型。通过后验预测检验发现，负二项观测模型与数据的拟合最为一致，因此本研究采用了该模型。

模型通过表达式 6.3a–6.3c 定义：

$$\mu \sim \mathrm{Gamma}(2, 1) \tag{6.3a}$$

$$\phi \sim \mathrm{Exponential}(1) \tag{6.3b}$$

$$C_i \sim \mathrm{NegBinomial}(\mu, \phi) \tag{6.3c}$$

其中，C_i 表示第 i个油田的检测计数。负二项似然的概率质量函数由位置参数 $\mu \in \mathrm{R}_{>0}$ 和过度离散参数 $\phi \in \mathrm{R}_{>0}$ 参数化，形式如下：

$$P(X = n \mid \mu, \phi) = \frac{\Gamma(\phi + n)}{n!\,\Gamma(\phi)} \left(\frac{\mu}{\mu + \phi}\right)^n \left(\frac{\phi}{\mu + \phi}\right)^\phi \quad \text{for } n \in \mathbb{N}_0\,, \tag{6.4}$$

其中 $\Gamma(\cdot)$ 是伽马函数。通过这种参数化，随机变量 $X \sim P$ 的期望和方差为：

$$\mathbb{E}[X] = \mu \quad \text{and} \quad \mathbb{V}[X] = \mu + \frac{\mu^2}{\phi}\,. \tag{6.5}$$

由于负二项分布描述的是率参数服从伽马分布的泊松随机变量，并且泊松(μ) 的方差为 μ，因此学到的参数对州政府具有很好的解释意义：

- μ 表示从检测计数角度来看，平均强度的含义，这与泊松率参数的解释相同。 μ的值越大，平均而言，油田层面的火炬检测次数就越多。

- ϕ 表示北达科他州各油田之间的异质性。具体而言， μ^2/ϕ 表示率参数为 μ的泊松分布之外的额外方差。 ϕ的值越小，检测计数极端（偏离 μ）的油田就越多。

为展示该模型与观测值的相容性，本文使用了2018年10月的数据。油田总数为506个。所有油田的检测计数分布如图 6.2所示。

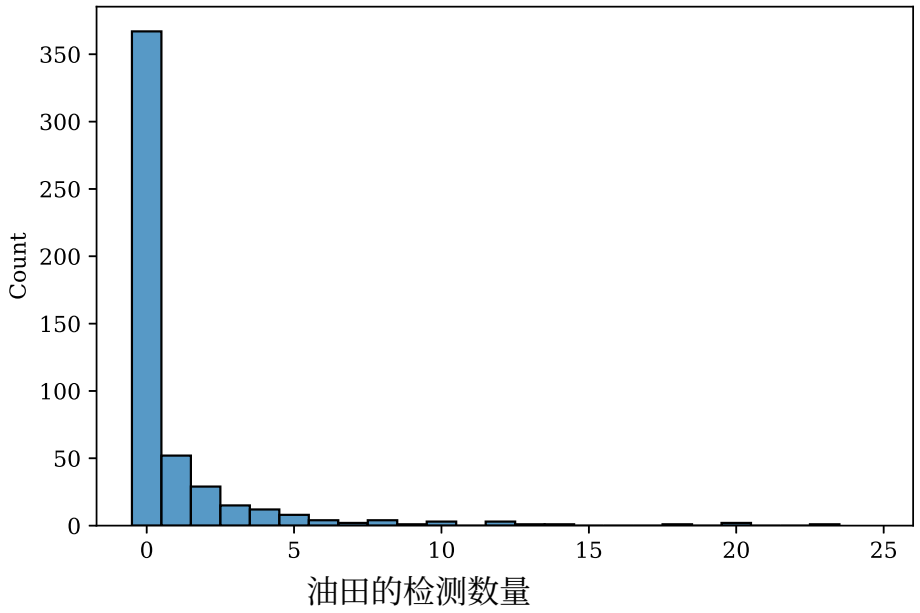


图 6.2：2018年10月油田检测次数分布的直方图。零值很多（超过350个），而且少数油田的检测次数相对较高（例如，大于或等于20）。

拟合模型 6.3后，超参数的后验分布和轨迹图见图 6.3。参数估计结果列于表 6.1。

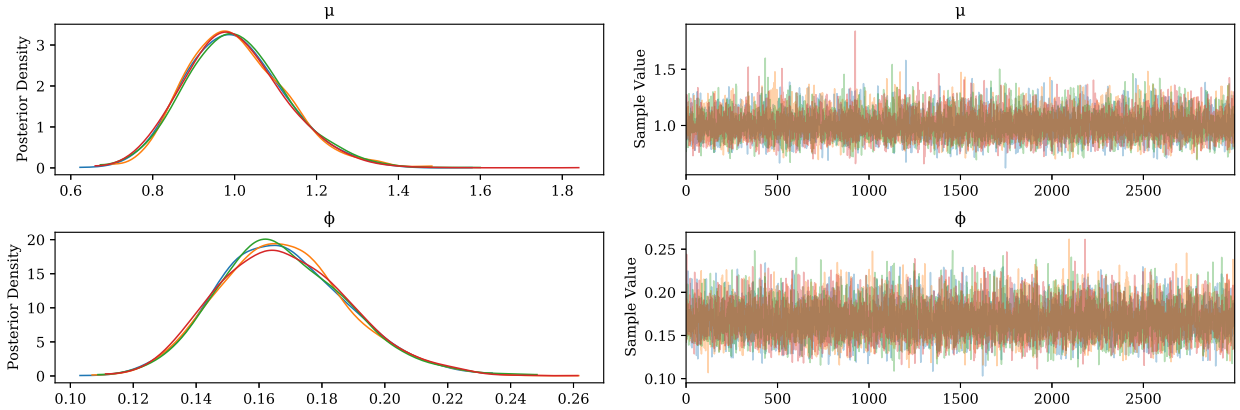


图 6.3：使用2018年10月的数据拟合油田检测计数分布所得的后验分布和轨迹图。已实现良好混合并达到收敛。

表 6.1：油田检测计数分布的参数估计值

参数	变量	点估计	90%置信区间
μ	强度	1.005	(0.814, 1.200)
ϕ	异质性	0.168	(0.135, 0.202)

强度参数 μ 的点估计相对较小（$\hat{\mu} \approx 1$），这可能是因为模型受到大量零计数的影响。然而，通过查看图 6.2 中的直方图，可以确定分布的尾部远远延伸至 $\hat{\mu}$ 之外。因此，我们进行了后验预测检验，以仔细评估模型 6.3 与观测值的兼容性。

这类检查充分利用了从后验分布中抽取的样本所包含的信息。生成模型将后验分布所描述的参数不确定性与似然所描述的结果不确定性结合起来，用于模拟模型所蕴含的观测值。随后生成后验预测图，以展示基于模型的预测以及原始数据。图 6.4 展示了检测计数分布模型的这类图。

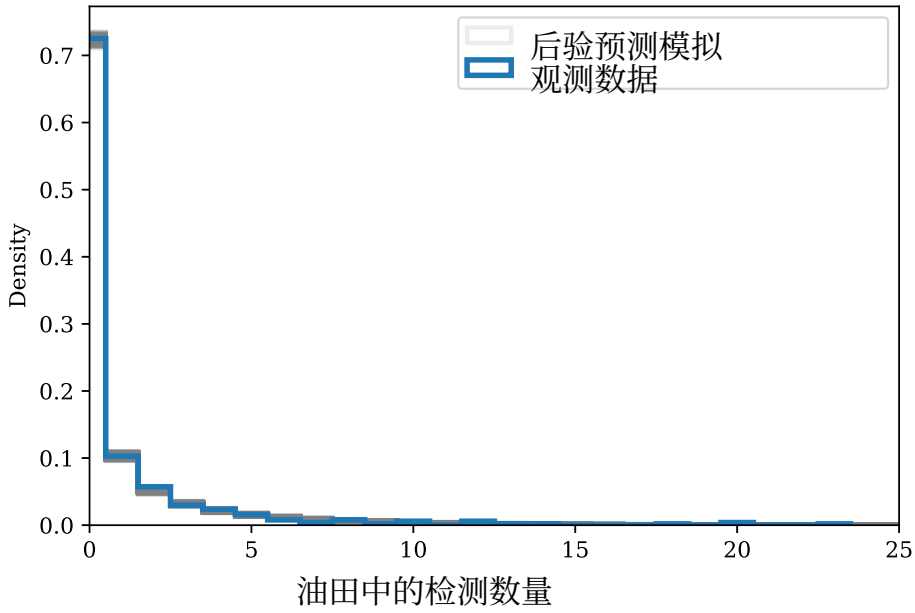


图 6.4：2018年10月油田检测次数分布的直方图。蓝色：VIIRS观测到的原始数据。灰色：由模型 6.3 得出的后验预测模拟结果。每组模拟结果均以带透明度的灰色绘制，采用阿尔法混合（设置 $\alpha = 0.15$），因此直方图中颜色较深的灰色表示与模型期望更吻合的模拟数据。

图 6.4 展示了原始 VIIRS观测以及所有后验预测模拟的直方图。每组参数值（ μ 和 ϕ）都用于模拟一幅北达科他州油田在2018年10月的合成快照，总共生成了 12，000 幅快照（这些快照由四条马尔可夫链的样本构建，每条链均设置为进行 3000 次采样迭代）。每个直方图都以无填充的折线图呈现，即绘制为“阶梯”直方图。

从图 6.4 可以看出，该模型与 2018年10月 的观测值非常吻合，因为观测数据与模拟数据之间没有明显且持续的差异。为了深入考察尾部行为，即零计数之外的部分，图 6.5 展示了放大后的视图。从中可以观察到一些差异，例如计数为 $C_i = 11$ 和 $C_i = 12$ 时。需要注意的是，在均值如此之低（$\hat{\mu} \approx 1$）的情况下，即使过度离散程度相对较高（$\phi \approx 0.2$），模型仍会对较高的检测计数感到意外，例如 $C_i \geq 20$ 时。

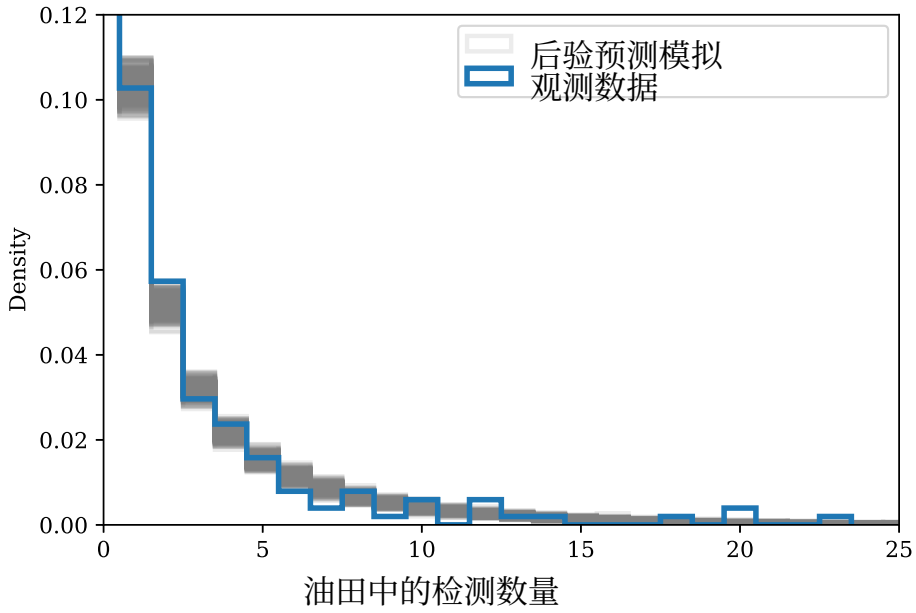


图 6.5：2018年10月油田探测次数分布的直方图，其中截去了 y 轴的部分，以便更清楚地呈现大于零的检测次数。图例及其对应的配色方案与图 6.4 相同。

以负二项似然为特征的模型 6.3 的整体表现，以及真实数据的复杂性，都体现在后验预测检验中。如第 6.3 节前文所述，研究在许多随机选取的月份中，将负二项似然与另外三种似然（泊松、零膨胀泊松和零膨胀负二项）进行了比较，发现总体而言，前者与数据的吻合度更高。事实上，有些月份的数据分布得更为“规整”，即模型 6.3 几乎能够完美描述这些数据。作者选择不挑选这些数据，以免让读者对所开发模型的表现产生误解。

尽管如此，模型 6.3 的简洁性、可解释性和有效性在对检测次数分布进行建模的任务中得到了充分体现。实际上，州政府可以在以下两种用例中受益于该模型：

1. 当最新月份的数据可用时，可拟合模型 6.3，以获得 μ 和 ϕ 的估计值。这些参数估计值及其可信区间可以与较早时期的结果进行比较。对于上述讨论，

学习得到的参数可以与 2018年8月/9月或 2016/2017年10月进行比较。通过比较，可以深入了解平均检测次数是否有所增加（表现为 μ 较大），或者是否发现了更多检测数量异常的油田（表现为 ϕ 较小）。

2. 模型拟合完成后，建议执行图 6.4 和图 6.5 所示的后验预测检验，以识别拟合中的任何问题。与模拟数据存在较大偏差的油田列表，尤其是位于远端尾部的油田（例如，当 $C_i \geq 20$ 时），值得持续跟踪。也就是说，需要调查每个月出现的“异常”是否是从总体中随机抽取的样本，或者是否每月都保持不变。这有助于从检测计数的视角进一步了解油田总体的行为。

检测次数的分布汇总仅展现了燃除景观的一个侧面，而燃除量分布则提供了另一个至关重要的侧面，下面将对此进行讨论。

6.4 燃烧气体体积建模

本节从分布视角研究 VIIRS 估计的不同油田的燃烧量。为便于演示，本文分析了一个三个月期间的数据集。具体而言，按照第3.3节所述进行逆地理编码后，计算并汇总了各油田在2018年第四季度期间的累计燃烧量，以供分析。

在这段时间内，共有 152 个油田有 VIIRS 报告量。数据高度偏斜（图6.6）。因此，在分析中，我们计算每个油田放空气量（以十亿立方米计）的数量级，而不是使用原始绝对体积。

从应用角度来看，对度量取对数可将其转换为量级（McElreath2015），这一做法应用于每个油田的放空气量：

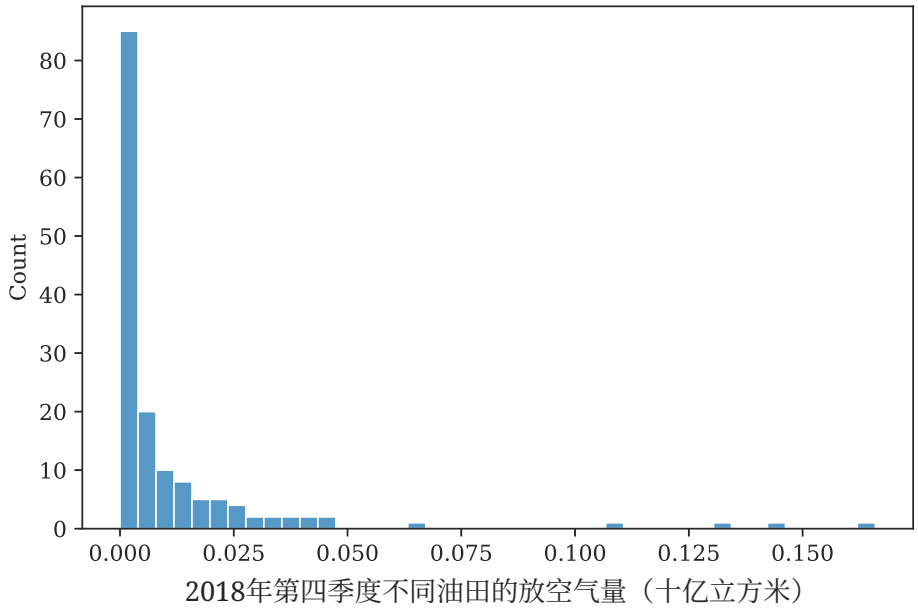


图6.6：2018年第四季度油田燃除气体体积分布的直方图。大多数油田的燃烧量相对较小（例如，小于 0.01 十亿立方米），而少数油田的燃烧量大于 0.1 十亿立方米。

$$L_i = \log(F_i), \tag{6.6}$$

其中， F_i 是以十亿立方米为单位的原始放空气量， L_i 是放空气量的量级，两者均对应第 i个油田。在本学位论文中，对数始终采用底数 e （即自然对数）。图6.7展示了这些量级的单变量分布。

在用于可视化分布的三种方法中，只有地毯图不会因所用超参数而产生细微差异。不过，作为一维散点图，其表现能力自然有限。直方图存在图 6.1 所示的问题。曲线由核密度估计（KDE）生成。对于公式6.1中定义的给定数据集，KDE 将潜在分布表示为：

$$\hat{p}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\Big(\frac{x - x_i}{h}\Big), \tag{6.7}$$

其中， $K(\cdot)$ 是核函数， h 是带宽参数。为生成图6.7，使用了高斯核，其定义如下：

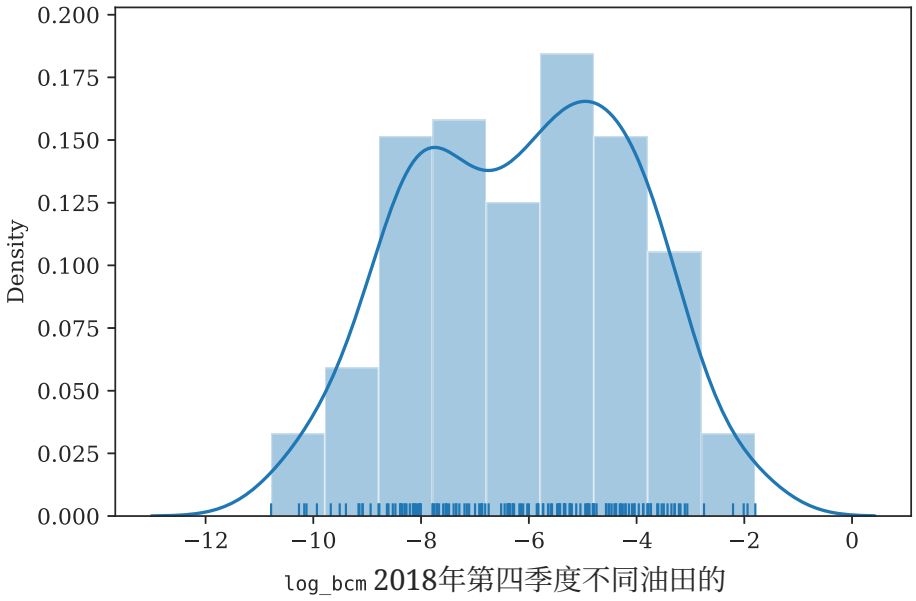


图6.7：2018年第四季度油田放空气量的分布。地毯图标出了每个油田对应的数值。直方图采用九个组距生成。曲线显示了高斯核密度估计。

$$K(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \tag{6.8}$$

并根据斯科特规则选择 h 。

由于带宽与直方图中的组距作用相似，核密度估计也可能遇到与直方图相同的问题。尽管如此，地毯图、直方图和核密度估计这三种方法都表明，用单个高斯近似来表示生成这些数据的密度并不理想。因此，我们采用高斯混合模型（GMM）来表示数据，也就是说，模型6.2中的基分布选用高斯分布。高斯混合模型具有更强的表达能力，也可用于聚类。

6.4.1 模型设定

如前所述，由于放空气量是连续量，因此适用密度估计，并采用高斯混合模型进行处理。首先考虑数据生成过程，这为潜在的聚类应用铺平了道路。也就是说，假设每个数据点 L_i （定义于

公式 6.6）恰好由一个混合成分生成。成分数量 K 未知，因此尝试用最多七个成分拟合图6.7中展示的数据集。尝试的成分数量相对较少，因为随着成分数量增加，建模结果会变得更难解释。模型由表达式 6.9a–6.9i 指定， $\forall K \in \{2, \ldots, 7\}$:

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) = 6 \cdot \mathbb{1}_K \tag{6.9a}$$

$$\mathbf{p} \sim \text{Dirichlet}(\boldsymbol{\alpha}) \tag{6.9b}$$

$$z_i \sim \text{Categorical}(\mathbf{p}) \tag{6.9c}$$

$$l_1 = \min\{L_1, \ldots, L_n\} \tag{6.9d}$$

$$l_2 = \max\{L_1, \ldots, L_n\} \tag{6.9e}$$

$$\widetilde{\mu}_k = l_1 + (k-1)\left(\frac{l_2 - l_1}{K-1}\right), \quad k = 1, \ldots, K \tag{6.9f}$$

$$\mu_k \sim \mathcal{N}(\widetilde{\mu}_k, 2), \qquad k = 1, \ldots, K \tag{6.9g}$$

$$\sigma_k \sim \text{Half-Normal}(2), \qquad k = 1, \ldots, K \tag{6.9h}$$

$$L_i \mid (z_i = j) \sim \mathcal{N}(\mu_j, \sigma_j) \qquad j \in \{1, \ldots, K\} \tag{6.9i}$$

其中:

$\boldsymbol{\alpha}$ 是狄利克雷分布的浓度参数向量；狄利克雷分布是贝塔分布的多元推广。

$\mathbf{p}$ i是混合成分的概率单纯形，其先验为狄利克雷先验。该先验将 $\boldsymbol{\alpha}$ 中的每个值设为 6，是一种弱信息先验，预期 $\mathbf{p}$ 中的任何 p_k 都可能大于或小于其他值。从 Dirichlet($[6, 6, 6, 6, 6, 6, 6]$) 中随机抽取的 10 个样本显示在图 6.8 中；

z_i 是第 i 个油田所属的可能混合成分；l_1 和 l_2 分别是 $\{L_i\}_{i=1}^n$ 的下界和上界；$\widetilde{\mu}_k$ 用于“初始化”第 k 个混合成分的位置，而 $\{\widetilde{\mu}_k\}_{k=1}^K$

本质上表示 $[l_1, l_2]$之间均匀分布的 K 个点；

μ_k 是第 k个高斯分量的均值；

σ_k 是第 k 个高斯分量的标准差；

L_i 是第 i 个油田的放空气量，其由混合成分 z_i 生成。

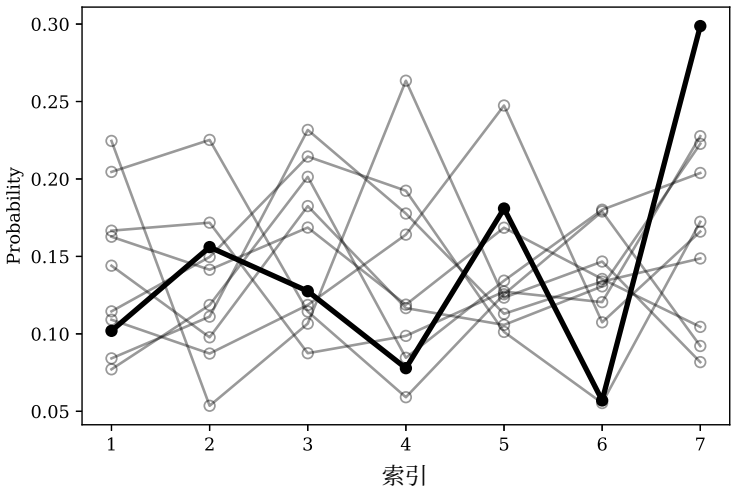


图 6.8：从狄利克雷先验中随机抽取的十个样本，其中 $\boldsymbol{\alpha} = (6, 6, 6, 6, 6, 6, 6)$。突出显示的一个样本表明，该先验较弱，因为它不会强制单次抽样中的所有概率都相等。

模型 6.9 明确表达了假设的生成过程，但依赖于对离散潜变量 z_n 进行采样，而其采样由类别混合分布控制。这种依赖会导致混合缓慢，并且无法有效探索后验分布。解决这些问题的一种等价参数化方法是将 z 参数边缘化。边缘化后的模型由表达式 6.10a–6.10h 指定，$\forall K \in \{2, \ldots, 7\}$:

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) = 6 \cdot \mathbb{1}_K \tag{6.10a}$$

$$\mathbf{w} \sim \text{Dirichlet}(\boldsymbol{\alpha}) \tag{6.10b}$$

$$l_1 = \min\{L_1, \ldots, L_n\} \tag{6.10c}$$

$$l_2 = \max\{L_1, \ldots, L_n\} \tag{6.10d}$$

$$\widetilde{\mu}_k = l_1 + (k-1)\left(\frac{l_2 - l_1}{K-1}\right), \quad k = 1, \ldots, K \tag{6.10e}$$

$$\mu_k \sim \mathcal{N}(\widetilde{\mu}_k, 2), \qquad k = 1, \ldots, K \tag{6.10f}$$

$$\sigma_k \sim \text{Half-Normal}(2), \qquad k = 1, \ldots, K \tag{6.10g}$$

$$L_i \sim \sum_{j=1}^{K} w_j \mathcal{N}(\mu_j, \sigma_j) \tag{6.10h}$$

其中 **w** 为混合权重（即混合比例），其余符号的含义与模型 6.9 中相同。表达式 6.10h 中定义的似然函数，对应于以一般形式表示的混合模型的密度（方程 6.2a）。

模型 6.10 被实现并拟合了六次（$\forall K \in \{2, \ldots, 7\}$），以比较指定不同成分数量时的推断结果。对于每个 K，马尔可夫链均实现了快速混合与快速收敛。建模结果见图 6.9，其中绘出了核密度估计（与图 6.7 相同）、推断得到的高斯分量以及后验样本。

可以观察到，使用高斯混合时，能够相对轻松地表示多峰特征，而且所有均值拟合都与使用核密度估计得到的结果相当接近。随着成分数量增加，例如当$K = 6$ 或 $K = 7$时，使用高斯混合模型得到的平均密度估计会更接近核密度估计，但后验样本表现出更强的随机性，这是潜在过拟合的一个指标。这自然引出了一个问题：如何为这个数据集确定成分数量。

6.4.2 模型比较

选择最佳的 K 是一个模型比较问题，对此并不存在一招制胜的办法。在这篇学位论文中，作者选择采用信息准则方法，具体使用 Watanabe（2010年）提出的广泛适用信息准则（WAIC）。信息准则可对样本外相对 KL 散度提供理论估计（McElreath 2020年），因此数值越低越好。按照 Martin（2018年）和 McElreath（2020年）的方法，WAIC 的计算公式如下：

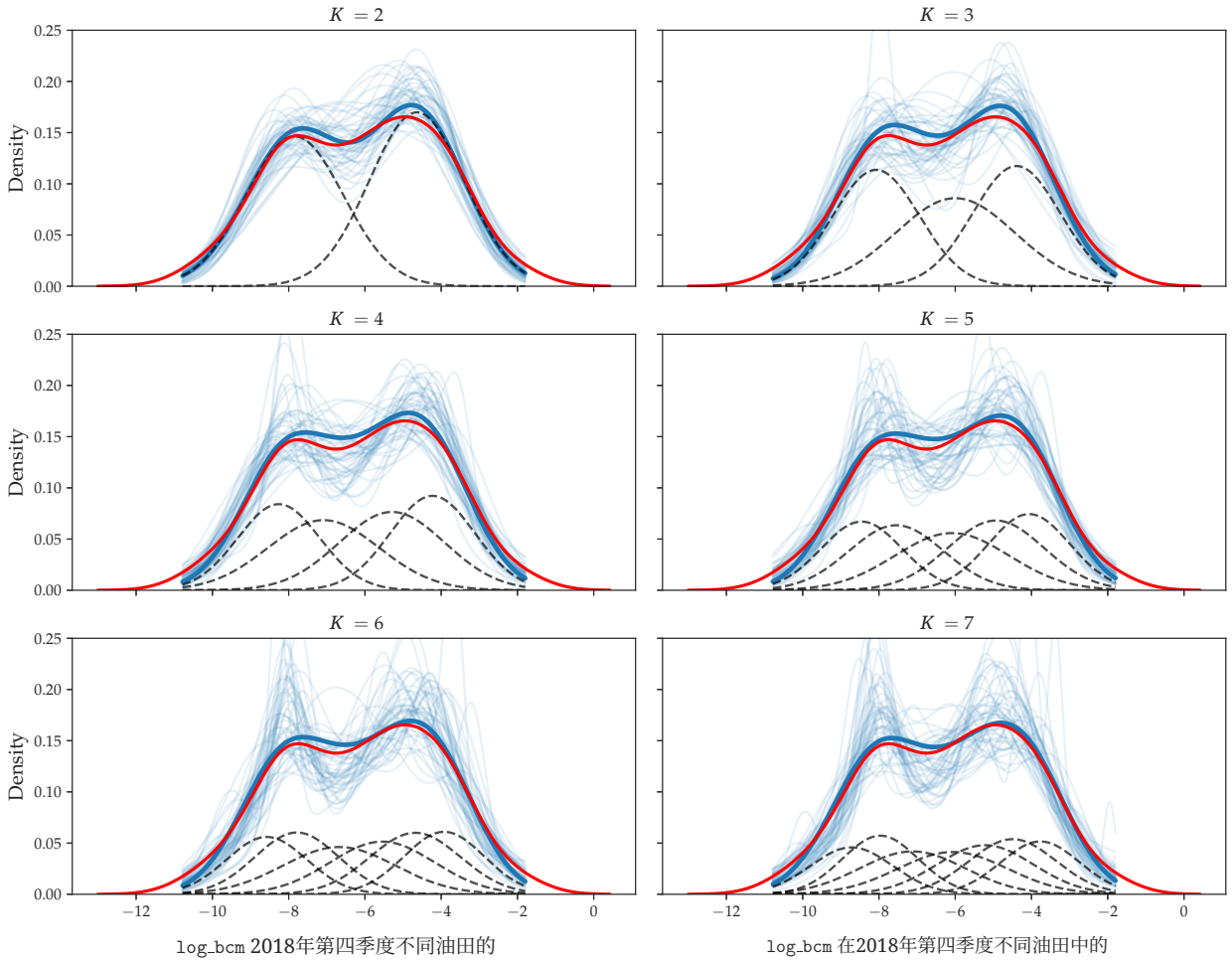


图 6.9：不同 K的高斯混合模型推断结果。粗蓝线表示底层密度的后验均值拟合。浅蓝线表示从后验分布中随机抽取的50个样本。虚线表示后验均值高斯成分。红色曲线表示使用核密度估计得到的拟合结果。

$$\mathrm{WAIC}(\mathbf{y}, \boldsymbol{\Theta}) = -2 \times \mathrm{lppd}(\mathbf{y}, \boldsymbol{\Theta}) + 2p_{\mathrm{WAIC}} \tag{6.11a}$$

$$= -2 \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{j=1}^{S} p(y_i \mid \Theta_j) \right) + 2 \sum_{i=1}^{n} \mathbb{V}_{\boldsymbol{\Theta}} [\log p(y_i \mid \Theta_j)] \,, \tag{6.11b}$$

其中：

$\mathbf{y}$ 表示观测值， y_i 是第 i个观测值；

$\boldsymbol{\Theta}$ 是后验分布， Θ_j 是第 j 组采样参数值；

S 是后验样本的数量；

lppd(·)计算逐点对数预测密度；

p_{WAIC} 是惩罚项，计算方法是对每个观测 i，将 S 后验样本中对数似然的方差求和。

从根本上说，模型比较是通过运用奥卡姆剃刀来进行的，即鉴于预测性能，优先选择简约的模型。模型根据其广泛适用信息准则值进行比较，并汇总于图 6.10。

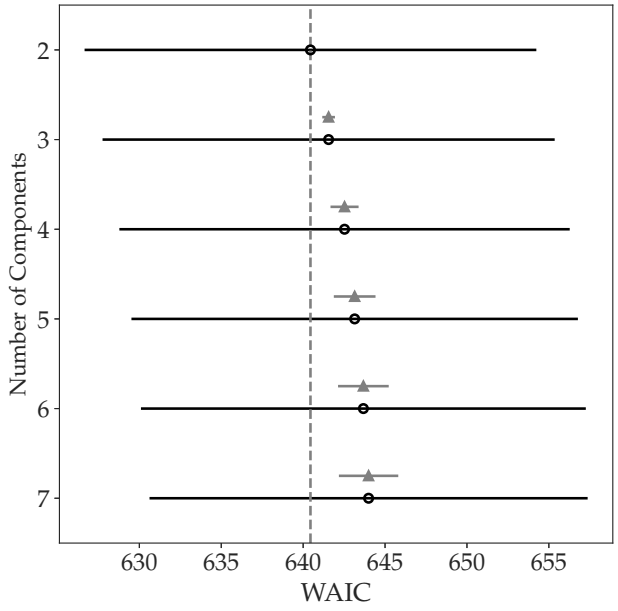


图 6.10：不同 K下的广泛适用信息准则值。空心点表示广泛适用信息准则值。较长的水平线段表示各广泛适用信息准则值的标准误。广泛适用信息准则差值的标准误（各模型与排名第一的模型之间）由带三角形的较浅色线段表示。

可以看出，包含两个高斯成分的模型是最佳模型（广泛适用信息准则值最小）。不过，考虑到估计标准误，所有模型之间存在相当大的重叠。鉴于 $K = 2$ 给出的模型最简单，而且该数据集中只有 152 个观测值（油田），包含两个成分的高斯混合模型（GMM）应是最佳选择。

6.4.3 聚类

从潜变量视角来看所建立的模型（模型 6.9），可以明显看出，混合模型是解决聚类任务的天然候选模型，因为每个观测值（L_i）都可以由其中一个 K 数据生成过程生成，每个过程都有自己的一组参数，$\mathcal{N}(L_i \mid \mu_k,\ \sigma_k)$。既然已经建立了概率模型，出于聚类目的，一个合理的做法是将数据点分配给后验概率最高的混合成分（即簇）（这些概率也称为责任度）。对于前几节中训练的双成分高斯混合模型，给定某个观测值 x，可以使用贝叶斯定理（公式 2.3a）计算该观测值属于簇 1（$z = 1$）的概率：

$$p(z = 1 \mid x) = \frac{p(z = 1)\,\mathcal{N}(x \mid \mu_1, \sigma_1)}{p(z = 1)\,\mathcal{N}(x \mid \mu_1, \sigma_1) + p(z = 2)\,\mathcal{N}(x \mid \mu_2, \sigma_2)}\,, \tag{6.12}$$

公式中的每一部分都可以从后验样本中获得（例如，使用后验均值）。

聚类作为一种无监督方法，可用于揭示观测值中的隐藏群组。就本章的油田燃烧量数据而言，这两个簇可以直接对应大型和小型燃烧油田等概念。然而，州政府通常更关注导致这些簇形成的深层原因，例如以此为决策和政策制定提供依据。如果分析更多月份或季度的数据后，发现属于大型燃烧聚类的油田成员构成似乎变化不定，那么燃烧量的变化可能与公司战略和动态的关联更为密切。另一方面，如果发现有一组油田经常加入大型燃烧聚类，这或许能为下一步天然气处理厂的选址提供参考；也就是说，应根据这些油田的情况，优化新建天然气厂的位置和产能。

本章为无监督学习整理的数据集是单变量的，即 $x_i \in \mathcal{X} \subseteq \mathrm{R}^1$。当数据维度超过 1 时，高斯混合模型同样适用于密度估计和聚类任务。例如，对于研究的同一批油田，若从 NDIC 提取其 2018年第四季度的石油产量，可在图 6.11 中看到天然气燃烧量与石油产量的散点图。底层分布的密度很可能可以用二元正态分布或二维高斯混合模型来建模。在这种情况下，混合成分将成为多元正态分布，而成分协方差矩阵则可借助 LKJ分布（用于模型 4.5 和 4.7）构建。例如，所建立的密度模型可用于异常检测，以寻找那些有趋向左上角趋势的油田（其特征是石油产量极低而燃烧放空量巨大）。与本学位论文中提出的所有推断类似，这种做法的一项优势是，决策可以依据一些一致的指标（例如概率评分），而不是依赖人工目测或临时发挥来确定标准。

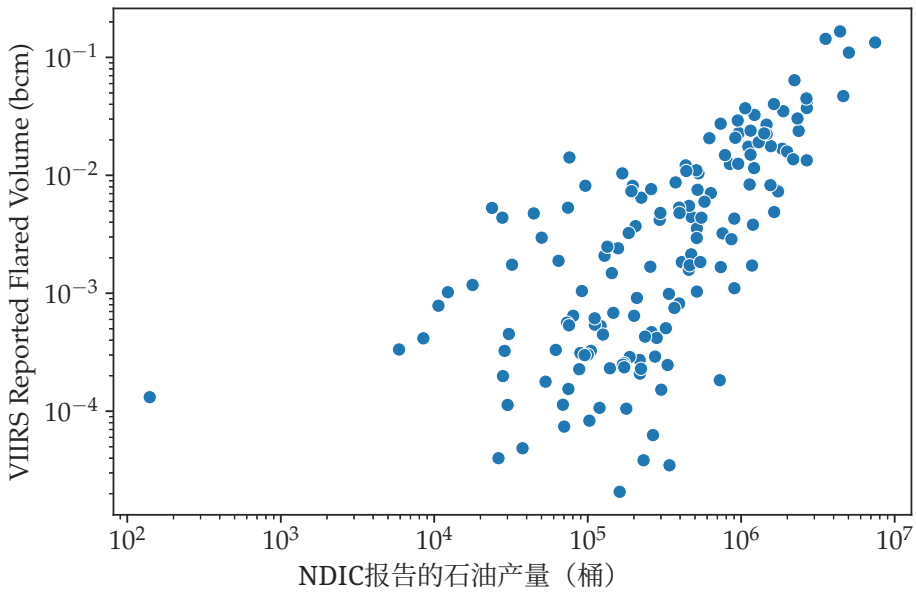


图 6.11：2018年第四季度不同油田的石油产量与燃除气体体积散点图。 x轴和 y轴均采用对数刻度，展示了两者数量级之间的关系。

至此，本学位论文的统计建模之旅告一段落。下一章将讨论将贝叶斯学习应用于燃烧数据的一个扩展情景，以及一个更宏观的视角。

第7章

讨论

本章讨论了运营商级监测与分析的可能性、潜在的结果不一致性，并将从燃烧放空数据中学习这一探索与机器学习在石油工程领域的更广泛应用过程联系起来。

7.1 运营商级监测与分析

到目前为止，卫星探测的燃烧放空统计数据已应用于州、县和油田层面。这得益于第3.3节讨论的逆地理编码。理想的应用场景是利用卫星检测结果中的信息，在运营商层面开展监测与分析。

遗憾的是，将火炬归属到相应公司并非易事。一种可能的解决方案是利用租赁地块的形状文件，而 NDIC 并未提供这些文件。一些数据供应商的数据库中有此类文件。然而，在花费一番功夫调查一家供应商提供的租赁地块形状文件后，作者认为，引入这类信息可能会带来更多问题，而非解决现有问题。具体而言，原因包括：

- 单个租赁地块上存在多家公司。
- 租赁地块形状文件中的公司名称并不总是与 NDIC 月份生产报告中的名称相符。
- 供应商数据库中的一些租赁地块缺少开始日期或结束日期数据。
- 供应商需要时间来汇编和数字化此类信息，因此现有的租赁地块形状文件并非最新。

尽管如此，鉴于这是一个如此重要的应用场景，作者还是设法开发了一种基于最近邻的方法，部分解决了这一问题（算法 7.1）。该方法的核心是谨慎地将距离每个卫星探测到的燃烧最近的油井的运营商分配给该燃烧。最近的油井是根据相应的时间窗口确定的。例如，对于2016年1月探测到的燃烧，只查询同月份 NDIC 生产报告中列出的活跃井。函数 `FindClosestOperator()` 返回每个 VIIRS 探测结果对应的最近运营商（OP$_j$），以及每对燃烧和油井之间计算出的距离（d_j）。该距离基于哈弗辛度量计算，即大圆距离，因此需要地球半径（R_E）。该函数本质上是在为 $k = 1$ 执行 k 近邻（k-NN）搜索。当样本量像本例这样大时，即每个月通常有几百个 VIIRS 检测结果和超过 15，000 口油井，针对每个 VIIRS 检测结果对每口油井的位置进行线性扫描会过于缓慢。因此，在本研究中，该函数内部依赖 Scikit-learn（Pedregosa 等，2011）中的球树实现，以加快 k-NN 搜索。

针对每个 VIIRS 探测结果获得二元组（OP$_j$，d_j）后，会执行一些逻辑来判断是舍弃还是保留运营商分配。思路很简单：当 d_j 非常小时，立即保留该分配；当其非常大时，则立即舍弃。如果 d_j 处于中间范围，即 $d_{\text{SECURE}} \leq d_j \leq d_{\text{CUTOFF}}$，则只有在火炬与运营商被判定位于同一镇区/区段/地段时，该分配才会生效。作为算法 7.1 输入的一部分，镇区/区段/地段形状文件可从 NDIC GIS 地图服务器获取。逆地理编码遵循与第3.3节完全相同的流程。处理完成后，少部分 VIIRS 检测结果不会用于运营商层级分析，因为它们距离报告的井位太远，或镇区/区段/地段无法匹配。需要指出的是，算法 7.1 的伪代码旨在展示数据处理逻辑的具体细节。在本研究的实现中，为提升性能，部分 for 循环被向量化运算取代。

算法 7.1： 基于最近邻的火炬所有者分配

输入： VIIRS 和 NDIC 报告数据的 WGS 84 坐标，以及
北达科他州的乡镇/区段/分区 Shapefile，d_{SECURE}, d_{CUTOFF}, R_E

输出： 分配给大多数 VIIRS 探测结果的运营商

```
n ← 个月份数
对于 i ← 1 到 n 执行
    VIIRS_i ← 来自 VIIRS 的第 i 个月份观测值
    NDIC_i ← 来自 NDIC 的第 i 个月份报告数据
    (OP d) ← FindClosestOperator(VIIRS_i, NDIC_i, R_E)
    m ← OP 中的记录数量或 d
    对 j ← 1 到 m 执行
        OP_j ← 第 j 条记录中找到的最近运营商
        d_j ← 火炬与最近油井之间的距离，对第 j条记录
        如果 d_j > d_CUTOFF 那么
            删除 OPj
        否则如果 d_j < d_SECURE 那么
            保留 OPj
        else
            如果 镇区/区段/地段一致则
                保留 OPj
            else
                丢弃 OPj
            end
        end
    end
结束
```

所开发的方法使用来自北达科他州的真实燃烧数据进行了测试。本节展示的案例中，算法 7.1 采用以下取值：

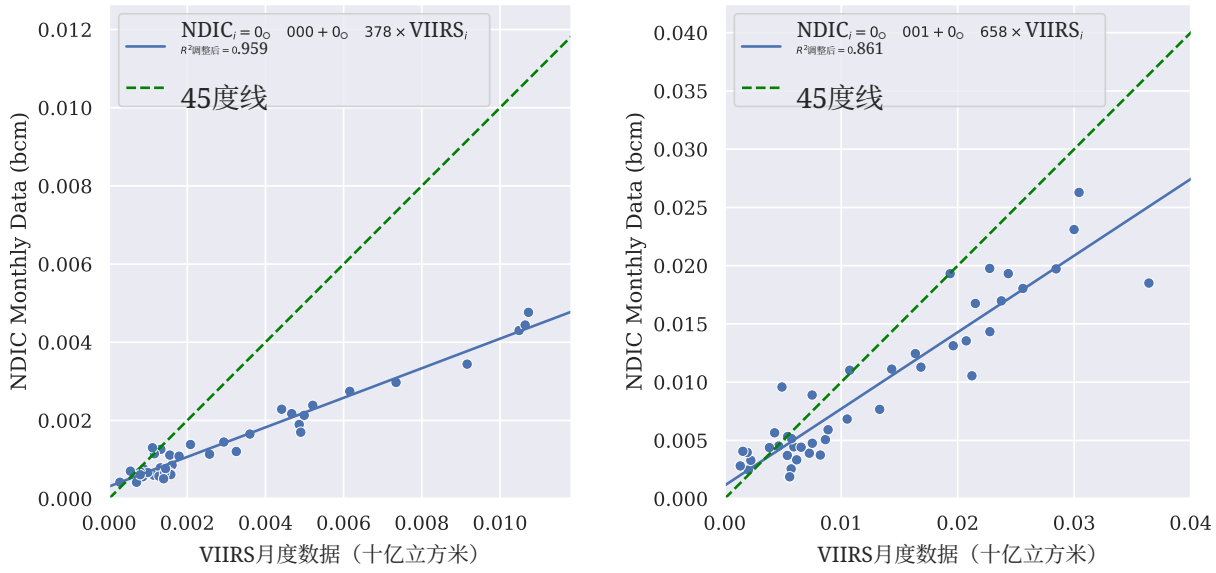
$$d_{\text{SECURE}} = 300\,\text{m} \tag{7.1a}$$

$$d_{\text{CUTOFF}} = 800\,\text{m} \tag{7.1b}$$

$$R_E = 6371\,\text{km} \tag{7.1c}$$

研究发现，部分运营商的 NDIC 与 VIIRS 报告体积呈正相关。图7.1展示了两个运营商的示例，分别记为运营商 B 和运营商 C。坐标轴的含义与图3.7右侧面板中的相同。

图例展示了采用普通最小二乘法（OLS）拟合公式 3.2d 的结果。 R^2_{adj} 表示调整后的 R^2。尽管ˆ β_{OPERATOR} 的差异表明不同公司之间存在异质性，但这些运营商各自的报告方式较为一致，并且与 VIIRS 数据吻合良好，仅相差一个比例因子（因为截距非常接近零）。

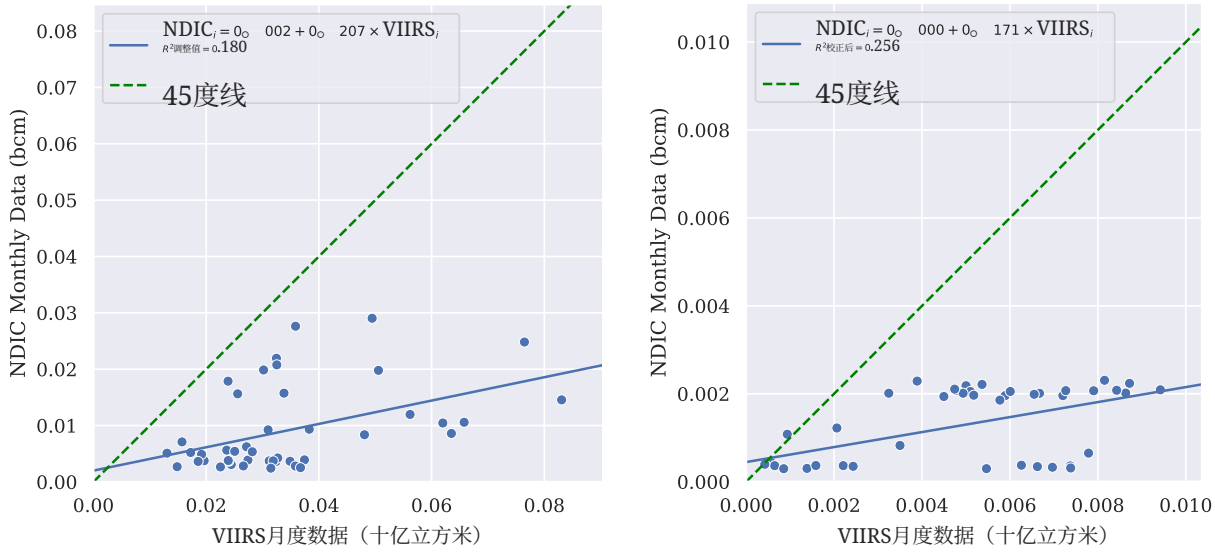


(a) 运营商 B　　(b) 运营商 C

图7.1：NDIC 与 VIIRS 报告量之间拟合良好的示例（运营商层面）。

然而，一些运营商（例如运营商D和运营商E）的报告数据与卫星探测的燃烧放空统计数据存在差异，这体现在拟合效果较差（图7.2）。当然，线性模型拟合效果不佳本身并不能说明太多。尽管如此，两个散点图中都存在一种模式：一些点似乎被“压低”到 x 轴附近。绘制这两家运营商的时间序列后可以看出，这种情况是由于公司报告体积在某一段时间内趋于平稳所致（图7.3）。时间序列中的 VIIRS 曲线表明，这些时段的燃烧强度发生了变化。由算法 7.1 驱动的这一工作流程能够在遇到此类数据集时发出警示，可作为强大的监测

和分析工具，但使用时务必保持高度谨慎。



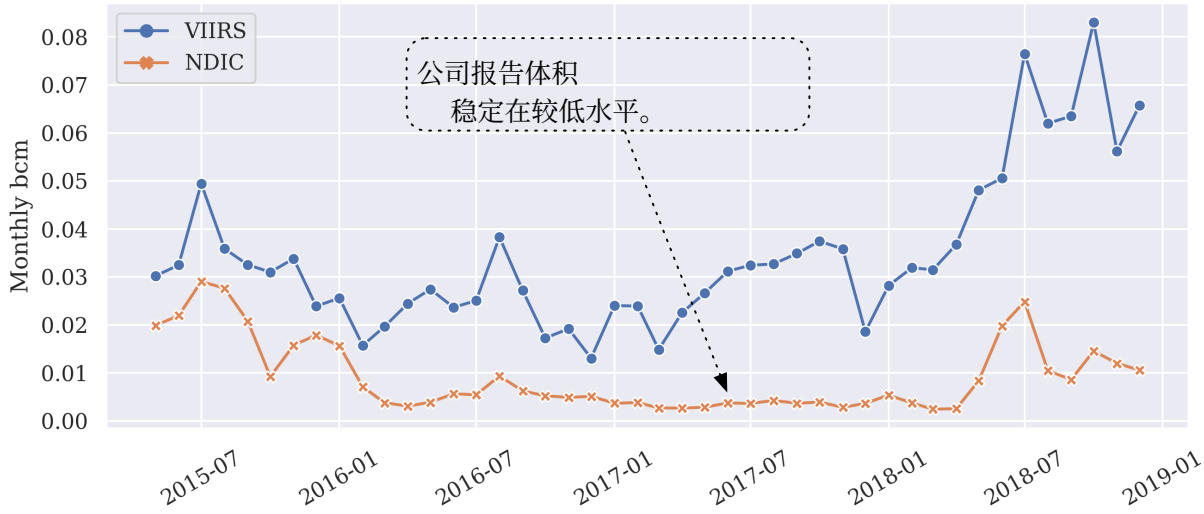
(a) 运营商D (b) 运营商E

图7.2：运营商层面 NDIC 与 VIIRS 报告体积拟合不佳的示例。

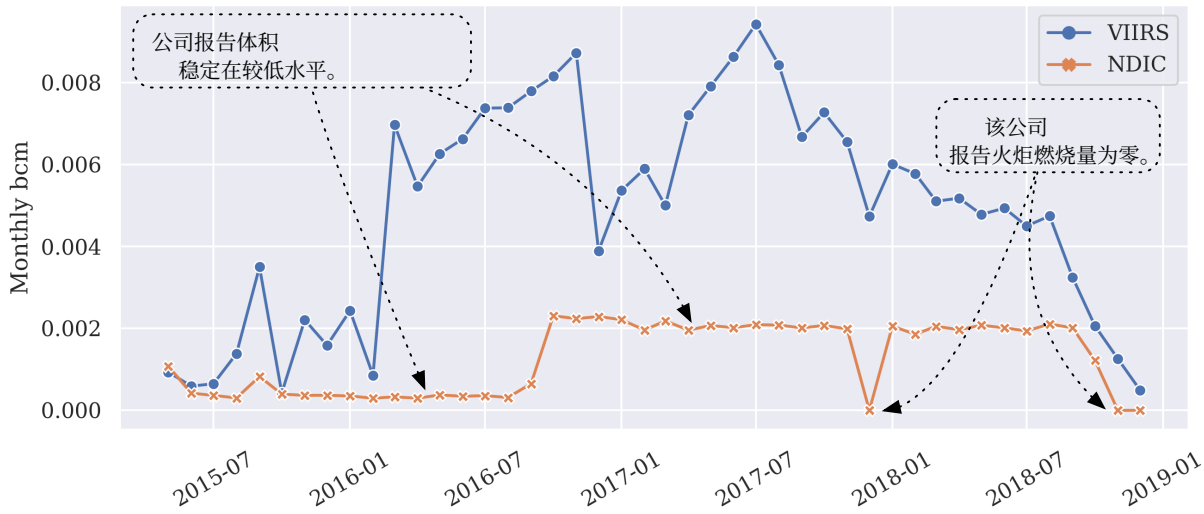
尽管所提出的方法看起来前景可期，但绝非一站式解决方案，而且存在被误用的可能。首先，可能会将卫星探测到的燃烧错误地归属给运营商。每当有人提出这一担忧时，都可以通过查看探测地图以及运营商生产现场的卫星影像开展进一步调查。此外，该方法对产量或火炬燃烧量相对较大的运营商更为有效，因为当一家公司几乎不进行火炬燃烧时，图5.20中针对峰值讨论的截断效应会被放大。

7.2 不一致警告

鉴于卫星影像的分辨率，将特定的火炬燃烧量归属于某个运营商面临诸多挑战。尽管 VIIRS 处理流程能够识别面积约为 1 m^2 的火炬（图 3.2(a)），但像元覆盖范围要大得多（表2.1）。由于每次 VIIRS观测都会记录像元中心的纬度和经度（Elvidge等（2015）），当多个运营商存在亚像元燃烧



(a) 运营商D



(b) 运营商E

图7.3：两个报告与 VIIRS 探测到的趋势/模式并非完全吻合的示例运营商的时间序列。公司报告数据与卫星检测结果显著不同的时间点或周期已作标注。

源时，火炬所有者归属就会极具挑战性。在这种情况下，仅通过对比公司报告与 VIIRS报告得出的结论很可能并不准确。事实上，在 NDIC报告方面，必须针对这些结果中的任何不一致之处发出警示，并从三个方面加以考量。首先，美国能源部（2019a）的报告提供了支持以下观点的数据：北达科他州

与得克萨斯州和新墨西哥州相比，美国国家海洋和大气管理局对火炬燃烧量的估算与各州报告数据更为吻合。第二，与放空相比，火炬燃烧更可取，因为甲烷（天然气的主要成分）的危害性强于二氧化碳，而二氧化碳是火炬燃烧的主要产物（美国能源信息署，2019b）。由于北达科他州禁止放空，巨大的燃烧放空量表明，直接将天然气排放到大气中的情况已降至最低。第三，火炬燃烧量的估算本身就是一项艰巨的任务。在无法实际测量火炬燃烧气体的情况下，加拿大石油生产商协会（2002）提供了可用体积估算方法的指南。各种方法类别，无论是采用经验法则、实验确定的相关性，还是流程模拟器，都各有局限性和精度问题。鉴于本研究采用的 VIIRS 体积数据主要使用 Cedigaz 报告的数据进行校准（第 2.1 节），而该数据本身也带有误差棒（Elvidge 等人，2015），因此，公司报告与 VIIRS 报告之间的差异无法得出明确结论，也不足为奇，尤其是在差异的标准误大于差异本身的情况下。

通过检查更全面的时间序列概况，可以看出上一节中的运营商D和运营商E向NDIC提交的报告数据前后一致。它们的时间序列分别显示在图7.4和图7.5中。变量及其相关标签（见图例）沿用第3.4节中的定义。表7.1列出了所有变量的个体。显然，报告的燃烧量与天然气产量和气油比的变化曲线吻合良好。燃烧量的一些快速变化与天然气价格的波动相符；即天然气价格下跌时，运营商往往会增加火炬燃烧，而天然气价格达到峰值时，火炬燃烧量很少。总之，要对运营商报告质量相关的决策和结论作出明确判断，还需要分辨率更高的卫星数据，并更全面地审查时间序列概况。

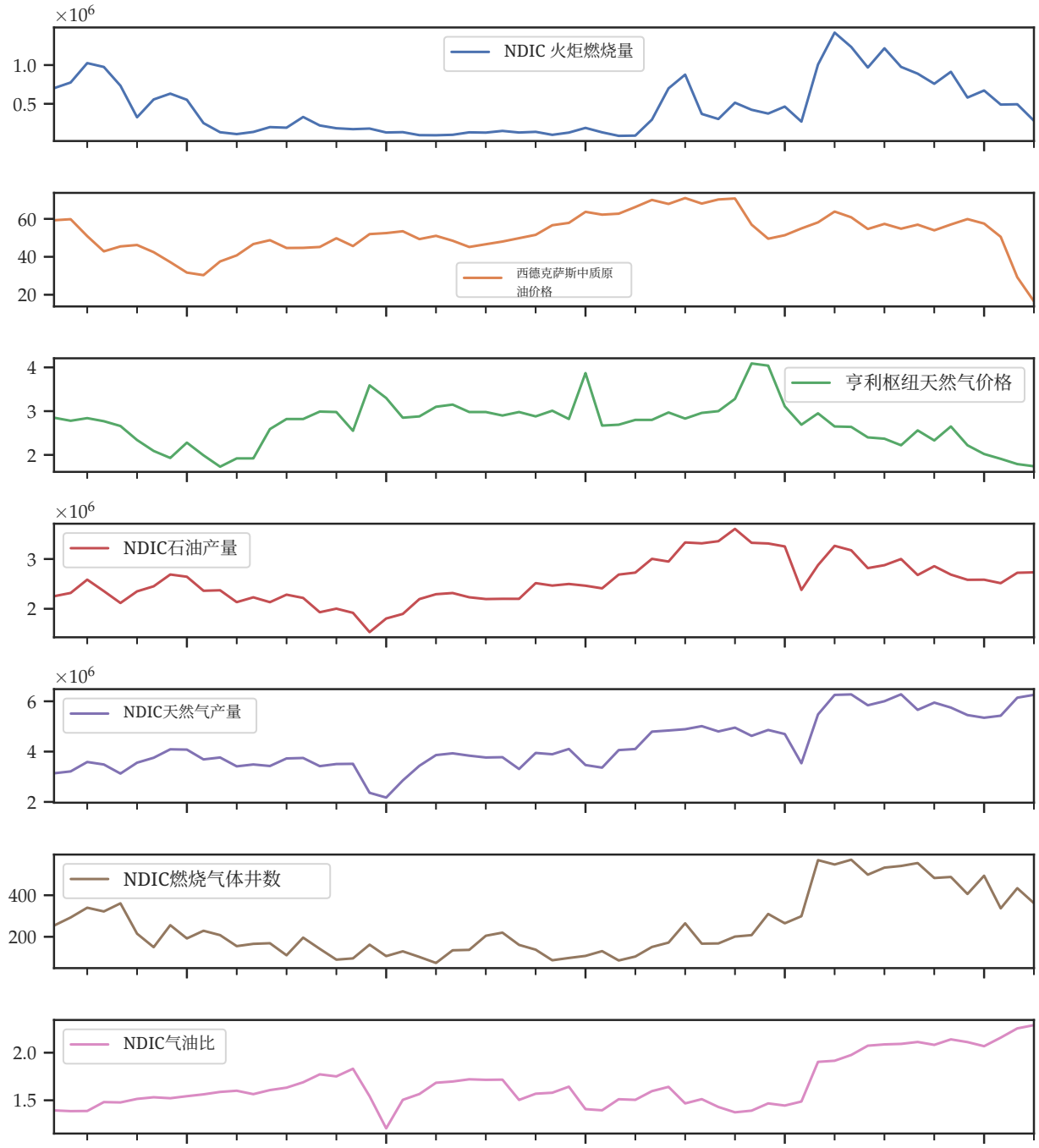


图7.4：运营商D更全面的时间序列图。2019年初报告的放空气量增加，与同期天然气价格下降相对应。

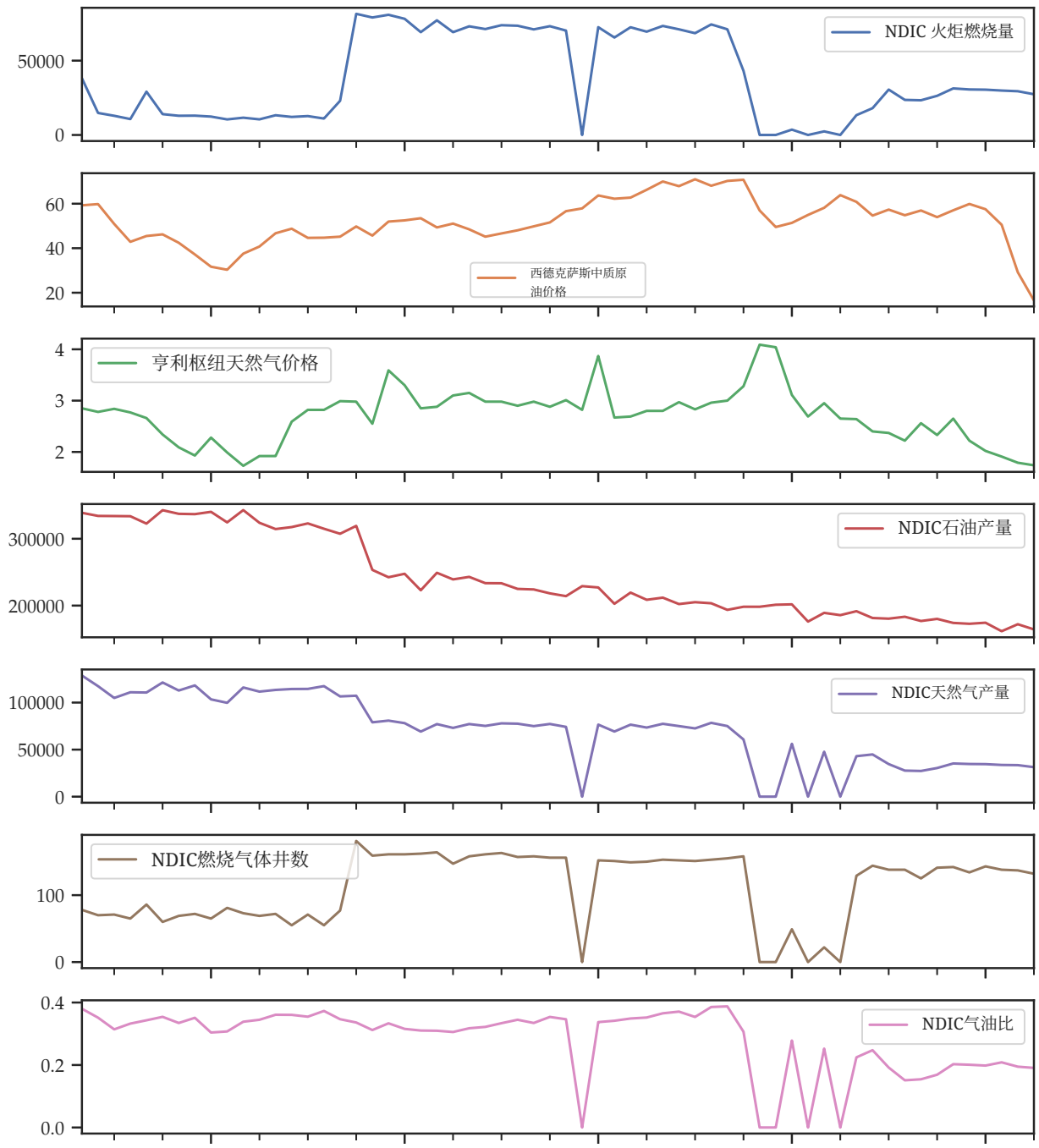


图7.5：运营商E更全面的时间序列图。2018年末报告的放空气量骤降，对应于天然气产量停止。

表 7.1：图 7.4 和图 7.5 中运营商时间序列的个体

变量	Unit
NDIC 火炬燃烧量	Mcf
西德克萨斯中质原油价格	$/桶
亨利枢纽天然气价格	$/百万英热单位
NDIC石油产量	bbl
NDIC天然气产量	Mcf
NDIC燃烧气体井数	1
NDIC气油比	千立方英尺/桶

7.3 石油数据分析中的注意事项

作为一名石油工程师，作者很高兴看到石油和天然气行业及学术界正拥抱数据驱动的思维方式和解决方案，并通过撰写这篇学位论文参与其中。不过，仍有一些方面可以持续改进，本节将讨论其中之一，也就是谨慎地接纳某些黑箱模型。

近年来，某些黑箱模型的广泛影响力，可以通过在 OnePetro 上进行粗略搜索看出（表 7.2）。需要注意的是，从算法的角度来看，这些方法更像是“玻璃箱”，而非“黑箱”；也就是说，其实现层面的所有底层机制都已得到充分理解。例如，作为神经网络训练核心的反向传播，基于链式法则。然而，对于给定任务，网络内部学到的参数对于问题领域几乎没有或完全没有提供任何洞见。因此，它被视为黑箱。

这类模型得到广泛应用，很大程度上得益于开源库的普及。例如，在 Python 生态系统中，借助 TensorFlow 和 PyTorch，神经网络的构建与训练变得简单得多；借助 XGBoost、LightGBM 或 CatBoost，只需几行代码即可构建梯度提升模型。换言之，由于这些统计程序的数学细节已被抽象封装，对从业者而言，实现这些模型几乎就像按下一个学习按钮一样

表 7.2：OnePetro 上的出版物数量增长

精确短语搜索	方法引入年份	出版物数量	
		2010–2014年	2015–2019年
神经网络	1958†	843	2044
梯度提升	2001‡	1	110
随机森林	2001§	9	245

† 基于（Rosenblatt 1958）
‡ 依据（Friedman 2001）§
依据（Breiman 2001）

图形用户
界面。

遗憾的是，易于实现并不意味着适合解决该问题。具体而言，这些黑箱模型面临以下挑战：

1. 如何融入领域专业知识。

   频率学派框架中的许多黑箱模型都假设观测值是条件独立同分布的。人们希望通过向通用逼近器（例如神经网络）输入大量独立同分布的样本，能够以一定的准确度优化出用于预测的某个函数。对于某些应用，领域专业知识通常体现在特征选择过程中。例如，为了训练一个预测石油产量的模型，分析师可能会选择某些完井参数，而不是 API井号或油井名称，作为输入特征。

   然而，在作者看来，这种融入领域专业知识的方式仍然较为浅显，远未达到石油和天然气行业数十年来积累的知识水平。例如，压裂窜扰造成的井间干扰，使得将某些邻井假定为独立同分布这一做法难以成立。另一个例子是，考察某个盆地中由 N 油田服务公司完井的油井总体时，领域专业知识可能表明，每家公司都应有自己的模型，但各家公司在完井技术等方面并非完全

彼此独立。在这种情况下，第4章采用的层次模型可能是更好的选择，这样便可以将大量关于不同服务公司的先验知识纳入总体模型。

2. 如何解读结果。

如前所述，黑箱模型存在可解释性问题。以上文条目1中的页岩气井为例，如果训练的是黑箱模型，那么目前还无法将模型未能捕捉井间干扰效应归因于神经网络的某个部分，或决策树的某个部分（对于梯度提升树或随机森林而言）。Rudin（2019）认为，人们应该“停止解释黑箱机器学习模型”，并在高风险决策中使用可解释模型。在石油工业中，存在许多高风险决策场景，例如实时井完整性异常检测，以及在油井成本高昂的情况下进行产量预测。在这些场景中盲目应用黑箱模型，可能会造成重大损失。在提供可解释性方面，本学位论文始终采用的贝叶斯方法要有效得多。每一项假设都通过先验分布或似然在生成模型中得到明确表达。

3. 如何量化不确定性，尤其是在风险管理和决策的背景下。

与上文第 2 点类似，误差棒至关重要，尤其是在高风险预测应用中。以使用训练好的数据驱动模型预测石油产量为例，像 1000 桶/天这样的点预测结果并没有太多参考价值。事实上，如果 95% 预测区间（PI）为 1000 ± 50桶/天，那么这个点预测就更有参考意义。然而，如果 95% PI 为 1000 ± 1500桶/天，那么同一个点预测就毫无帮助，甚至会造成误导。相反，应该报告的是：所考虑的

模型在这项特定任务中存在很大的不确定性，或者该实体可能完全不会产生任何产量。

需要指出的是，CI/PI 中的“95”并非什么“神奇数字”。州政府或石油公司可能希望基于 73% 或 99.6% 的置信度，或任何其他任意选择来做出决策。真正重要的是，必须以有原则的方式量化机器学习估计/预测中的不确定性，从而能够计算出任意区间。正如本学位论文通篇所述，贝叶斯方法在这方面提供了充分的能力和灵活性。事实上，对于参数估计值，作者选择给出 90%置信区间，而非“惯例上的”95%，以强调这应由领域本身而非统计学来决定。

然而，频率学派框架中的许多黑箱模型无法满足这一要求。最大似然估计（MLE）是某些频率学派学习方法的根本依托，它具有非常理想的性质，也能够量化不确定性，但前提是拥有海量数据，使渐近性质得以发挥作用。遗憾的是，在石油工程领域的许多情形中并非如此，下一节将对此展开讨论。

4. 当数据不够“大”时，如何缓解过拟合。

这里有两个方面值得讨论。首先，大数据并非无处不在。事实上，作者认为，Gelman（2015）提出的“样本量永远不够大”这一观点适用于石油工业中的许多问题。原因在于，如果数据量足够大，分析人员早已转向下一个需要更多数据的问题。例如，巴肯组中500口生产井的样本，就足以开展一些一般性研究。如果分析人员能够获取包含超过 15,000口油井的数据集，那么人们就会希望获得更细致的见解。尤其是在不同服务公司/运营商、不同

地层成员或不同完井技术之间需要进行部分汇聚时，总体中某些个体的数据量可能非常少（第4章的分析便是如此）。另一方面，应结合模型复杂度来考察样本量。参数数量是衡量模型复杂度的一种指标。例如，考虑一个假设的分类问题，其目标是判断给定油井的生产绩效是良好、一般还是较差。有十个完井参数（特征）可用于训练图 7.6 所示的多层感知机。

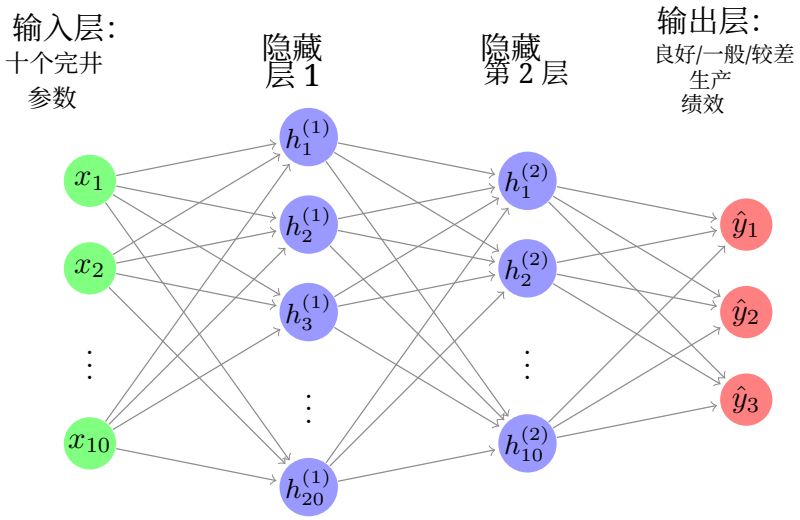


图 7.6：为假设的井性能分类问题设计的神经网络。输入层包含 10 个神经元，用于处理完井参数。第一和第二隐藏层分别包含 20 个和 10 个神经元。输出层包含 3 个神经元，用于多分类。

在这个（小型）神经网络中，参数数量 n_p 为：

$$n_p = 11 \times 20 + 21 \times 10 + 11 \times 3 = 463, \tag{7.2}$$

考虑除最后一层外的每一层都设置一个偏置节点。要训练这个模型，包含500口油井的数据集显然是一个小样本。不过，仍有可能用小样本训练这样的模型，但必须在正则化方面付出巨大努力，希望神经网络能够学到可以泛化的内容，而不是仅仅记住观测到的样本（即过拟合）。

借助正则化先验，贝叶斯方法内置的奥卡姆剃刀大大降低了过拟合的风险。尤其是贝叶斯非参数模型，例如第5章采用的高斯过程，具有很大吸引力，因为模型规模可以随数据规模的增长而扩大（Orbanz and Teh 2010）。这使得所开发的模型既灵活，又能稳健地避免过拟合。

尽管贝叶斯学习模型（例如本研究开发的模型）具有显著优势，值得在石油数据分析中得到更广泛的应用，但它们并非万能药。近年来，研究人员开始强调定制统计模型的必要性（Andorra 2020年；McElreath 2020年）。其观点是，现成模型无论是神经网络还是广义线性模型，都会妨碍领域专业知识的融入。这一点在石油工程领域尤为重要。例如，在针对水力压裂性能开展数据驱动分析时，将裂缝扩展模型引入机器学习流程是合理的。这样，统计模型便有了融入物理信息的模型作为依据。正如本学位论文自始至终所采用的那样，贝叶斯框架能够自然地接纳这一策略，即可以将领域知识（例如以微分方程表示的知识）嵌入生成模型中。其优势之一在于，许多参数将具有直接的科学含义，并且可以根据科学约束、现场经验等设置信息量更丰富的先验分布。最终应能得到更好的推断和预测。

第8章

结论与建议

在本学位论文中，通过天然气火炬燃烧数据学习模型的过程中，可以观察到完整贝叶斯方法的有效性。作者希望本研究能够帮助人们理解将数据驱动方法应用于天然气燃烧时的各种选择与考量。最后，本章提出了主要结论，并给出了未来工作的建议。

8.1 结论

主要结论如下：

1. 采用哈密顿蒙特卡洛实现的贝叶斯学习，可有效应用于天然气火炬燃烧分析中的实际问题，适用于监督式和无监督式场景。贝叶斯方法的优势表明，它值得在整个石油工程领域得到更广泛的应用；这些优势如下：

(a) 可以以有原则的方式融入石油技术领域专业知识。(b) 模型可解释性显著提升，从而便于与石油工程师沟通。(c) 不确定性量化有助于做出更稳健的决策，这对石油勘探与生产公司至关重要。(d) 内置的奥卡姆剃刀使模型在面对带噪声的现场测量时不易过拟合。

2. 构建一系列模型（表 8.1），涵盖参数和非参数技术，为如何从不同

角度提取洞见提供了指导。所提出的模型经过设计和测试，能够泛化到不同层级的各类实体。

3. 为了研究不同实体（如县或油田）之间的异质性，建议采用部分汇聚，因为有些实体的数据非常少。

4. 高斯过程在揭示燃烧放空时间序列中的模式和趋势方面表现出非常出色的特性。一组采用 Mat´ern $5/2$ 核函数的先验分布，在不同的建模目标、观测模型和数据来源下都表现良好。

5. 从分布视角来看，负二项分布和高斯混合模型分别能很好地表示油田燃烧放空次数和燃烧量。学习得到的参数和结构具有很强的可解释性。通过拟合高斯混合模型，可以发现隐含聚类。

6. 本文介绍了一种用于运营商级监测与分析的基于最近邻的方法。该方法在真实数据上进行了测试，并得到了可靠的结果。不过，对于多个运营商的油井彼此距离非常近的情况，需要分辨率更高的卫星数据。

7. 所有学位论文目标（第 1.2 节）均已实现。具体而言，通过拟合截距参数，估算了 VIIRS 漏报的全州及各县燃烧量，并报告于表 3.1 和表 4.2。确定了陆地卫星8号的夜间燃烧源探测限（未校正辉光造成的伪影），并报告于图3.2(b)。利用 Spearman 的 ρ 计算了州级财务因素、生产绩效与燃烧量之间的相关性；原始数据和滞后一期差分的结果分别报告于图 3.5 和图 3.6。大多数变量对在滞后一期差分中均未显示出较强的相关性。稳健高斯过程建模为实现其余目标提供了通用框架，包括展示运营商方法，

评估北达科他州监管政策（Order24665）的目标是否已实现，以及预测 NDIC 火炬燃烧量。

表 8.1：本学位论文开发的模型

编号	建模目标	Page
模型 3.2	州级 VIIRS 与 NDIC 之间的关联	27
模型4.5	县级 VIIRS 与 NDIC 的关联（中心化）	38
模型 4.7	县级 VIIRS 与 NDIC 的关联（非中心化）	41
模型 5.12	火炬燃烧天然气产量比例的时间序列	57
模型 5.14	进行火炬燃烧的油井比例时间序列	63
模型 5.15	VIIRS 检测计数时间序列	66
模型 5.17	石油火炬燃烧比例时间序列	70
模型 5.18	VIIRS 与 NDIC 之间的比例因子的时间序列	74
模型 6.3	油田的 VIIRS 检测次数分布	86
模型 6.9	油田的 VIIRS 体积分布（潜在离散参数化）	94
模型 6.10	油田的 VIIRS 体积分布（边际化）	95

8.2 未来工作

未来研究的若干方向包括：

1. Landsat 8 处理工作流。

   第3.2节的研究表明，纳入 Landsat 8 信息（使用现有的 VIIRS 工作流）会面临辉光导致的处理伪影这一挑战。针对 Landsat 8 定制处理算法将很有意义，这将为 VIIRS 与 Landsat 8 的数据融合铺平道路，从而提供分辨率高得多的解译结果。

2. 按月快速检测火炬。

   开发一种基于卫星影像的快速火炬检测和体积估算方法，将产生连续的月度数据流。由于 NDIC 需要大约两个月的周转时间来汇编并数字化公司报告，本学位论文提出的许多机器学习工作流将能够利用快速检测数据提供预测性见解。

3. 层次高斯过程。

   第5章中的模型是根据每个实体自身的数据学习得到的。部分汇聚（第4章）这一方案能够拓展到何种程度，值得探讨。通过层次高斯过程在不同实体之间进行汇聚，能否改善推断结果?

4. 时空分析。

   在上述第3项的基础上进一步探索，值得研究时空模型（允许跨时间和空间汇聚信息）的有效性。相邻实体的火炬燃烧行为是否相似?

5. 在贝叶斯非参数方法框架下统一一切。

   第6章中对高斯混合模型进行模型比较时，需要事先指定可能的簇数量。事实上，狄利克雷过程是狄利克雷分布的一种无限维推广，属于非参数方法，并能自动选择所需的簇数量。考虑到高斯过程（第5章）的有效性，探究非参数模型在燃烧数据分析中能发挥多大作用，将会很有意思。所有天然气火炬燃烧分析问题能否在贝叶斯非参数方法的统一框架下得到解决?

参考文献

Abu-Mostafa, Y. S., Magdon-Ismail, M. 和 Lin, H.-T. 2012年。《从数据中学习》，第一版。美国加利福尼亚州帕萨迪纳：AMLBook。Adams, R. P., Murray, I. 和 MacKay, D. J. C. 2009年。基于高斯过程强度的泊松过程中可处理的非参数贝叶斯推断。论文发表于第26届国际机器学习年会，美国纽约州纽约市，6月14日至18日。https://doi.org/10.1145/1553374.1553376。

Alexeyev, A., Ostadhassan, M., Mohammed, R. A. 等，2017年。基于测井的巴肯组地质力学与岩石物理分析。论文发表于第51届美国岩石力学/地质力学研讨会，美国加利福尼亚州旧金山，6月25日至28日。ARMA-2017-0942。

Andorra, A.，2020年。与 Michael Betancourt 合作的严谨贝叶斯工作流程。https : / / podcasts . apple . com / us / podcast / 6 - a - principled- bayesian- workflow - with - michael - betancourt/id1483485062?i=1000461447241。（访问日期：2020年10月30日）。

Andrieu, C., de Freitas, N., Doucet, A. 等，2003年。面向机器学习的马尔可夫链蒙特卡罗入门。机器学习50 (1-2):5–43. https://doi.org/10.1023/A:1020281327116。

Bandyopadhyay, S. 2018年。*CSM MATH 532*，讲义：空间统计学。科罗拉多矿业学院。

Berk, A., Anderson, G. P., Acharya, P. K. 等。2006年。MODTRAN5：2006年更新。发表于美国佛罗里达州奥兰多（基西米）国防与安全研讨会，4月17日至21日。https://doi.org/10.1117/12.665077。

Betancourt, M. 2015年。关于 GitHub 问题（Wishart 在有效参数下运行失败）的评论。https://github.com/pymc-devs/pymc3/issues/538#issuecomment-94153586。（访问于2020年10月30日）。Betancourt, M. 2017a。Stan 中的稳健高斯过程。https://betanalpha.github.io/ assets/case studies/gp part1/part1.html。（访问于2020年10月30日）。

Betancourt, M. 2017b。《使用 RStan 的稳健统计工作流程》。https://betanalpha.github.io/assets/case studies/rstan workflow.html。（访问于2020年10月30日）。

Betancourt, M. 2018。《条件概率论（面向科学家和工程师）》。https: // betanalpha.github.io/assets/case studies/conditional probability theory.html。（访问于2020年10月30日）。Betancourt, M. 2019。《概率建模与统计推断》。https://betanalpha. github.io/assets/case studies/modeling and inference.html。（访问于2020年10月30日）。Betancourt, M. 2020。《迈向严谨的贝叶斯工作流程》。https://betanalpha.github. io/assets/case studies/principled bayesian workflow.html。（访问于2020年10月30日）。Bolker, B. M. 2008。*R*语言中的生态模型与数据，第一版。新泽西州普林斯顿：普林斯顿大学出版社。Box, G. E. P., Luceño, A., and Paniagua-Quiñones, M. D. C. 2009。《通过监测与调整进行统计控制》，第二版。新泽西州霍博肯：约翰·威立父子公司。https://doi.org/10.1002/9781118164532。

Breiman，L. 2001年。《随机森林》。《机器学习45 》(1)：5–32。https://doi.org/10.1023/ A:1010933404324。

Burkov, A. 2019. 百页机器学习书，第一版。加拿大不列颠哥伦比亚省维多利亚：Leanpub。Campello, R. J. G. B., Moulavi, D., and Sander, J. 2013. 基于层次密度估计的基于密度的聚类。发表于第17届亚太知识发现与数据挖掘会议，澳大利亚黄金海岸，4月14日至17日。https: //doi.org/10.1007/978-3-642-37456-2 14. 加拿大石油生产商协会。2002. 上游油气设施燃烧与放空量估算。http://www.oilandgasbmps.org/docs/GEN23-techniquestomeasureupstreamflaringandventing.pdf.（访问于2020年10月30日）。Carpenter, B. 2019. 层次建模和多层建模的各种名称。https://statmodeling.stat.columbia.edu/2019/09/18/all-the-names-for-hierarchical-and-multilevel-modeling/.（访问于2020年10月30日）。Carpenter, B. 2020. 概率与统计：基于模拟的入门。https: //github.com/bob-carpenter/prob-stats.（访问于2020年10月30日）。

Carpenter, B., Gelman, A., Hoffman, M. 等。2017. Stan：一种概率编程语言。《统计软件杂志》**76** (1)：1–32。https://doi.org/10.18637/jss.v076. i01. Cedigaz. 2015. 各国年度天然气指标。https://www.cedigaz.org/databases/。（访问于2020年10月30日）。Davidson-Pilon, C. 2015. 《黑客的贝叶斯方法：概率编程与贝叶斯推断》，第一版。美国马萨诸塞州波士顿：Addison-Wesley。de Freitas, N. 2013. UBC CPSC 540，讲座：机器学习——贝叶斯学习。https: //www.youtube.com/watch?v=Fae0j1WN1zA&list=PLE6Wd9FR--EdyJ5lbFl8UuGjec vVw66F6&index=7&t=0s.（访问于2020年10月30日）。Deisenroth, M. P., Turner, R. D., Huber, M. F. 等。2012. 基于高斯过程的鲁棒滤波与平滑。《*IEEE*自动控制汇刊》**57** (7)：1865–1871。https://doi.org/10.1109/TAC.2011.2179426. Duane, S., Kennedy, A. D., Pendleton, B. J. 等。1987. 混合蒙特卡罗。《物理快报*B*》**195** (2)：216–222。https://doi.org/10.1016/0370-2693(87)91197-X. 佩恩研究所地球观测组。2019. EOG海报档案。https://payneinstitute. mines.edu/eog-poster-file/。（访问于2020年10月30日）。美国能源信息署。2019a. 天然气放空与燃烧。https://www.eia.gov/dnav/ng/ng prod sum a EPG0 VGV mmcf a.htm。（访问于2020年10月30日）。美国能源信息署。2019b. 2018年北达科他州和得克萨斯州的天然气放空与燃烧量有所增加。https://www.eia.gov/todayinenergy/detail.php?id=42195#.（访问于2020年10月30日）。美国能源信息署。2020a. 天然气价格。https://www.eia.gov/naturalgas/data.php#prices。（访问于2020年10月30日）。美国能源信息署。2020b. 石油及其他液体燃料价格。https://www.eia.gov/petroleum/data.php# prices。（访问于2020年10月30日）。Elvidge, C. D., Baugh, K., Zhizhin, M. N. 等。2017. 利用VIIRS支持国际社会检测非法捕捞和天然气燃烧。发表于2017年IEEE国际地球科学与遥感研讨会，美国得克萨斯州沃思堡，7月23日至28日。https://doi.org/10.1109/IGARSS.2017.8127580.

Elvidge, C. D., Zhizhin, M. N., Baugh, K. 等，2015年。基于可见光红外成像辐射计套件数据开展全球天然气燃烧调查的方法。《*Energies*》9 (1)。https: // doi.org/10.3390/en9010014。Elvidge, C. D., Zhizhin, M. N., Baugh, K. 等，2019年。利用短波长VIIRS数据扩展夜间燃烧源探测限。《*Remote Sensing*》11 (4)。https://doi.org/10.3390/rs11040395。Elvidge, C. D., Zhizhin, M. N., Hsu, F.-C. 等，2013年。VIIRS Nightfire：夜间卫星测温。《*Remote Sensing*》5 (9)：4423–4449。https://doi.org/10.3390/rs5094423。Ester, M., Kriegel, H.-P., Sander, J. 等，1996年。用于在含噪大型空间数据库中发现簇的基于密度的算法。论文发表于第二届国际知识发现与数据挖掘会议，美国俄勒冈州波特兰，8月2日至4日。Friedman, J. H. 2001年。贪心函数逼近：梯度提升机。《统计年鉴》29 (5)：1189–1232。Gelman, A., Carlin, J. B., Stern, H. S. 等，2013年。贝叶斯数据分析，第三版。佛罗里达州博卡拉顿：Chapman & Hall/CRC。Gelman, A. 和 Hill, J. 2006年。使用回归和多层／层级模型进行数据分析，第一版。纽约州纽约：剑桥大学出版社。Gelman, A. 2015年。N 永远不会很大。https://statmodeling.stat.columbia.edu/2005/07/31/ n is never larg/。（访问于2020年10月30日）。Gelman, A. 和 Rubin, D. B. 1992年。利用多条序列进行迭代模拟的推断。《统计科学》7 (4)：457–472。谷歌地球。2019年。显示北达科他州某油井井位天然气燃烧情况的卫星影像。https://earth.google.com/web。（访问于2020年10月30日）。Hastie, T., Tibshirani, R. 和 Friedman, J. 2009年。统计学习基础，第二版。纽约州纽约：施普林格。https://doi.org/10.1007/978-0-387-84858-7。He, K., Zhang, X.,Ren, S. 等，2015年。深入探究整流器：在 ImageNet 分类中超越人类水平的表现。论文发表于2015年IEEE国际计算机视觉大会，智利圣地亚哥，12月7日至13日。https://doi.org/10.1109/ICCV.2015.123。

Hinton, G. E. 和 Salakhutdinov, R. R. 2006年。利用神经网络降低数据维度。《科学》 313 (5786)：504–507。https://doi.org/10.1126/science.1127647。Hoffman, M. D. 和 Gelman, A. 2014年。无转向采样器：在哈密顿蒙特卡洛中自适应设定路径长度。《机器学习研究期刊》15 (47)：1593–1623。Jordahl, K., den Bossche, J. V., Fleischmann, M. 等，2020年。geopandas/geopandas：v0.8.1。https://github.com/geopandas/geopandas。（访问于2020年10月30日）。Jordan, M. I. 和 Mitchell, T. M. 2015年。机器学习：趋势、展望与前景。《科学》349 (6245)：255–260。https://doi.org/10.1126/science.aaa8415。Jordan, M. I. 和 Broderick, T. 2010年。加州大学伯克利分校统计学260课程，第1讲：历史与德·菲内蒂定理。https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/lectures/lecture1.pdf。（访问于2020年10月30日）。Kazakov, N. 和 Miskimins, J. L. 2011年。多元统计分析在非常规储层系统滑溜水压裂参数中的应用。论文发表于SPE水力压裂技术会议，美国得克萨斯州伍德兰兹，1月24日至26日。SPE-140478-MS。https://doi.org/10.2118/140478-MS。Krizhevsky, A., Sutskever, I. 和 Hinton, G. E. 2012年。使用深度卷积神经网络进行 ImageNet 分类。论文发表于第26届神经信息处理系统大会，美国内华达州太浩湖，12月3日至6日。Kumar, R., Carroll, C., Hartikainen, A. 等，2019年。Arviz：用于在 Python 中探索性分析贝叶斯模型的统一库。《开放源代码软件期刊》4 (33)。https: //doi.org/10.21105/joss.01143。La mbert, B. 2018年。贝叶斯统计学生指南，第一版。英国伦敦：SAGE Publications Ltd. Lewandowski, D., Kurowicka, D. 和 Joe, H. 2009年。基于藤结构和扩展洋葱法生成随机相关矩阵。《多元分析杂志》100 (9)：1989–2001。https://doi.org/10.1016/j.jmva.2009.04.008。Li, H. 2019年。统计学习方法，第二版。北京：清华大学出版社。MacKay, D. J. C. 1998年。高斯过程导论。北约高级研究院系列 *F*：计算机与系统科学168：133–166。

MacKay, D. J. C. 2003年。信息论、推断与学习算法，第一版。英国剑桥：剑桥大学出版社。Martin, O. 2018年。使用 *Python* 进行贝叶斯分析，第二版。英国伯明翰：Packt Publishing Ltd. McElreath, R. 2015年。统计学重新思考：一门以 *R* 和 *Stan* 为例的贝叶斯课程，第一版。佛罗里达州博卡拉顿：Chapman & Hall/CRC。McElreath, R. 2020年。统计学重新思考：一门以 *R* 和 *Stan* 为例的贝叶斯课程，第二版。佛罗里达州博卡拉顿：Chapman & Hall/CRC。McInnes, L., Healy, J. 和 Astels, S. 2017。Hdbscan：基于层次密度的聚类。《开放源代码软件期刊》**2** (11)。https://doi.org/10.21105/joss.00205。Mitchell, T. M. 1997。《机器学习》，第一版。纽约州纽约：McGraw-Hill, Inc. 美国国家航空航天局。2019年。陆地卫星8号。https://landsat.gsfc.nasa.gov/landsat-8/。（访问于2020年10月30日）。美国国家航空航天局。2020年。陆地卫星8号概览。https://landsat.gsfc.nasa.gov/landsat- 8/landsat- 8- overview/。（访问于2020年10月30日）。北达科他州工业委员会。2014年。北达科他州工业委员会第24665号命令政策/指南版本041718。https://www.dmr.nd.gov/oilgas/GuidancePolicy NorthDakotaIndustrialCommissionorder24665.pdf。（访问于2020年10月30日）。Olson, R. S., La Cava, W., Mustahsan, Z. 等。2017。将机器学习应用于生物信息学问题的数据驱动建议。*arXiv*电子预印本。arXiv:1708.05070（即将出版；发布于2017年8月8日）。Orbanz, P. 和 Teh, Y. W. 2010。贝叶斯非参数模型。载于《机器学习百科全书》，C. Sammut 和 G. I. Webb 编。马萨诸塞州波士顿：施普林格。https://doi.org/10.1007/978-0-387-30164-8 66。Papaspiliopoulos, O., Roberts, G. O. 和 Sk¨old, M. 2007。层级模型参数化的一般框架。《统计科学》**22** (1)：59–73。https://doi.org/10.1214/088342307000000014。Pedregosa, F., Varoquaux, G., Gramfort, A. 等。2011。Scikit-learn：Python机器学习。《机器学习研究期刊》**12** (85)：2825–2830。

Rasmussen, C. E. 和 Williams, C. K. I. 2006年。机器学习高斯过程，第一版。马萨诸塞州剑桥：麻省理工学院出版社。Rassenfoss, S. 和 Zborowski, M. 2018年。勘探与生产笔记（2018年11月）。《石油技术期刊》**70** (11)：22–27。https://doi.org/10.2118/1118-0022-JPT。Roberts, S., Osborne, M., Ebden, M. 等。2013。时间序列建模的高斯过程。《皇家学会哲学汇刊*A*辑》**371** (1984)。https://doi.org/10.1098/ rsta.2011.0550。Rosenblatt, F. 1958。感知机：一种用于大脑信息存储与组织的概率模型。《心理学评论》**65** (6)：386–408。https://doi.org/10.1037/h0042519。Rudin, C. 2019。停止为高风险决策解释黑箱机器学习模型，转而使用可解释模型。《自然·机器智能》**1** (5)：206–215。https://doi.org/10.1038/s42256-019-0048-x。Salvatier, J., Wiecki, T. V. 和 Fonnesbeck, C. 2016。使用 PyMC3 在 Python 中进行概率编程。《*PeerJ*计算机科学》**2**。https://doi.org/10.7717/peerj-cs.55。Srivastava, U.,Oonk,D., Lange,I. 等。2019。北达科他州天然气燃烧政策的分配影响。《电力期刊》**32** (8)。https://doi.org/10.1016/j.tej.2019.106630。Stan 开发团队。2020年。Stan建模语言用户指南和参考手册，版本2.24.0。http://mc-stan.org。（访问于2020年9月30日）。美国能源部。2019a。天然气燃烧与放空：州和联邦监管概述、趋势及影响。https://www.energy.gov/sites/prod/files/ 2019/08/f65/Natural%20Gas%20Flaring%20and%20Venting%20Report.pdf。（访问于2020年10月30日）。美国能源部。2019b。北达科他州天然气燃烧与放空法规（情况说明书）。https://www.energy.gov/sites/prod/files/2019/08/f66/North% 20Dakota.pdf。（访问于2020年10月30日）。美国地质调查局。2000。《美国地质调查局2000年世界石油评估：说明与结果。技术报告。美国地质调查局。https://doi.org/ 10.3133/ds60。

Valiant, L. G. 1984年。《可学习理论》。《*ACM*通讯》**27** (11): 1134–1142. https://doi.org/10.1145/1968.1972. ˇ VanderPlas, J., Connolly, A. J., Ivezi´c, Z. 等。2012年。《Astroml简介：天体物理学机器学习》。发表于美国科罗拉多州博尔德举行的2012年智能数据理解会议，10月24日至26日。https://doi.org/10.1109/CIDU. 2012.6382200. Vapnik, V. N. 2000年。《统计学习理论的本质》，第一版。美国纽约州纽约：施普林格。https://doi.org/10.1007/978-1-4757-3264-1. Watanabe, S. 2010年。奇异学习理论中贝叶斯交叉验证与广泛适用信息准则的渐近等价性。《机器学习研究期刊》11: 3571–3594. 维基百科。2019年。可见光红外成像辐射计套件。https://en.wikipedia.org/wiki/ Visible Infrared Imaging Radiometer Suite.（访问于2020年10月30日）。Wolpert, D. H. 1996年。学习算法之间缺乏先验区分。《神经计算》8 (7): 1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341. 世界银行。2019年。全球天然气燃除减排伙伴关系（GGFR）燃烧数据。https: // www.worldbank.org/en/programs/gasflaringreduction#7.（访问于2020年10月30日）。Yu, J. 2017。《机器学习：从公理到算法》，第一版。北京：清华大学出版社。Yu, X., Trainor-Guitton, W., 和 Miskimins, J. L. 2020年。水平井砂堵检测的数据驱动方法。发表于美国得克萨斯州伍德兰兹举行的SPE水力压裂技术会议暨展览会，2月4日至6日。SPE- 199707-MS. https://doi.org/10.2118/199707-MS.