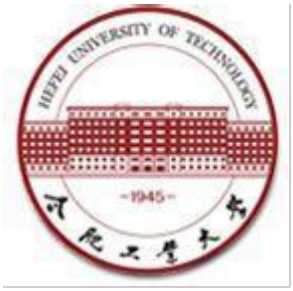
学　　号：2015216972　　密　级：公开

合肥工業大學

Hefei University of Technology

本科毕业设计（论文）

UNDERGRADUATE THESIS



类　　型：论文

题　　目：基于深度学习对澳大利亚森林覆盖的预测分析

专业名称：物联网工程

入校年份：2015 级

学生姓名：王 林 钊

指导教师：刘建（讲师）

学院名称：计算机与信息学院

完成时间：2019 年 6 月

合　肥　工　业　大　学

本科毕业设计（论文）

基于深度学习对澳大利亚森林覆盖的预测分析

学生姓名：王 林 钊

学生学号：2015216972

指导教师：刘建 （讲师）

专业名称：物联网工程

学院名称：计算机与信息学院

2019 年 06 月

A Dissertation Submitted for the Degree of Bachelor

Research and Application of Forest Cover in Australia Based on Deep Learning Method

By

Wang Lin Zhao

Hefei University of Technology

Hefei, Anhui, P.R.China

June 06, 2019

毕业设计（论文）独创性声明

本人郑重声明：所呈交的毕业设计（论文）是本人在指导教师指导下进行独立研究工作所取得的成果。据我所知，除了文中特别加以标注和致谢的内容外，设计（论文）中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 合肥工业大学 或其他教育机构的学位或证书而使用过的材料。对本文成果做出贡献的个人和集体，本人已在设计（论文）中作了明确的说明，并表示谢意。

毕业设计（论文）中表达的观点纯属作者本人观点，与合肥工业大学无关。

毕业设计（论文）作者签名：　　签名日期：　　年　　月　　日

毕业设计（论文）版权使用授权书

本学位论文作者完全了解 合肥工业大学 有关保留、使用毕业设计（论文）的规定，即：除保密期内的涉密设计（论文）外，学校有权保存并向国家有关部门或机构送交设计（论文）的复印件和电子光盘，允许设计（论文）被查阅或借阅。本人授权 合肥工业大学 可以将本毕业设计（论文）的全部或部分内容编入有关数据库，允许采用影印、缩印或扫描等复制手段保存、汇编毕业设计（论文）。

（保密的毕业设计（论文）在解密后适用本授权书）

学位论文作者签名：　　指导教师签名：

签名日期：　　年　月　日　　签名日期：　　年　月　日

摘　要

森林是宝贵的自然资源，但目前森林资源遭受严重损失(1990-2015 年森林占全球陆地面积由 31.6%变为 30.6%，从 41.28 亿公顷减少为 39.99 亿公顷)。森林覆盖损失变化有很多影响因子：除与降雨温度自然因素有关，还与政策条例和经济发展相关。目前的研究主要使用不连续的时间数据集，无法在长期连续时间序列中得到影响因子，使用的主要是统计计量和传统机器学习模型等，其预测评估性能还有提升空间。

本文主要的工作如下：以澳大利亚为例展开使用 1988-2014 年 6 个州的数据集（时间、空间、时空变量），对森林转型的鲁棒性进行研究，并找到可能影响森林转型的特征。首先，对澳大利亚地区使用 GEE 平台进行可视化分析并且进行相关性检测；然后，使用深度学习技术在不同地区抓取长期连续的森林覆盖时空数据集，开发训练 LSTM 模型以用于预测澳大利亚森林覆盖率变化，并将模型运行在高性能计算集群中。在建模过程中，量化森林覆盖数据集中气候，社会经济，自然地理和州管辖等因子。接着，使用 RMSE 和R^2指标与空间计量模型进行比较。最后，分析影响因子时，通过排除每个或每组的解释变量后，再评估模型预测能力，从而达到评估时空动态因素的效果。

关键词：森林覆盖率 ;LSTM ; GEE(Google Earth Engine); 时间序列

ABSTRACT

Forest is a kind of valuable resource, but it has suffered a great loss currently (the rate of forest in mainland globally takes up from 31.6% to 30.6%, which means the land has reduced from 4.128 billion ha to 3.9999 billion ha). The change in forest cover loss enjoys many influential factors: aside from relevant to natural elements such as rainfall as well as temperature, it is still combined with policy regulation and economic development. The present scholarship takes discontinuous time datasets which cannot get influential factors in long-term continuous time series. They mainly employed statistics, econometrics and traditional machine learning model, the assessment performance has much space to enhance.

The main tasks of this article incorporate that: take Australia as an example, we used the datasets in 6 continents from 1988 to 2014(comprising of temporal, spatial and spatiotemporal variables).made research on the robust of forest transformation, and found possible influence on variables of forest transformation. To begin with, we use Google Earth Engine to visualize and analyze the Australia region; Then, we took deep learning tech to capture long -term continuous forest cover spatiotemporal dataset. We developed and trained LSTM(Long Short Term Memory) in order to project changes of Australian forest cover rate and implemented it on the high performance computing clusters. During the modelling process, quantifying factors such as climate, social-economic, natural of forest cover datasets. Next, we used the index RMSE and R^2 to compare it with the special econometrics model. Eventually, when analyzing influential factors, we excluded each variable or group explanatory variables to assess the projection ability of models, so that we could reach the effect of assessing spatiotemporal dynamic factors.

KEYWORDS: Forest Cover; LSTM(Long-Short-Term Memory) ; Deep Learning ; GEE(Google Earth Engine) ; Time Series

目录

插图清单

表格清单

1 绪论

1.1 研究背景及意义

森林作为不可或缺的资源，为人类生产木质品也提供必要的生态服务，包括生态多样性，气候管控，减少土壤流失的危害等。然而近年来，全球的森林局势不容乐观，近几十年的农业扩张导致热带和亚热带地区严重的森林覆盖损失。因为缺少可持续解决森林恶化的森林管理策略以及平衡经济社会和环境的方法。同时，最近几十年北美，欧洲和澳大利亚部分地区与热带森林中的发展中国家相比经历严重的净森林覆盖损失[1]。调查证明，净森林损失主要发生在森林覆盖保持稳定或增长的地区[2]。森林损失主要是全球人口数量增多，并且变得富裕后，他们的日常饮食习惯随之改变，导致对食物和生物燃料产量的需求增加[3](Laurance et al., 2014)。而在北方，温带森林，热带大草原和灌木丛地区，森林覆盖却在增长，这些区域由于 CO2 和气候变化，以及市场和政府的激励使得森林能够增长。

影响森林覆盖因素是复杂的，在不同时期的地区和社会经济条件下有许多不同的驱动因素。需要模型来帮助森林的相关受益者理解森林覆盖的时空动态，识别不同的驱动因素，同时也能够更好的预测森林未来的走势。进一步促进实施森林管理措施，设计森林保护法，以及分配稀有保护资源。

1.2 研究现状

虽然已有许多模型被开发出来用作森林覆盖变化分析。但是这些模型的数据主要是分离且粗糙的时间步长数据[4](Matthews et al., 2007)。Huang 就基于 1990 到 2000 全球 19 个地区的的地球资源卫星图像使用支持向量机模型生成一个森林覆盖变化地图。Freitas (2010)[5]使用广义最小二乘回归模型来评估道路密度，土地使用和森林覆盖变化，数据集来自靠近巴西圣保罗市 3 年(1962，1981 和 2000 年)的热带森林的航空照片。Kamusoko (2013)[6]建立马尔可夫单元自动机模型来模拟老挝地区未来的森林覆盖变化，并使用卫星获取 1993，1996，2000，2004 年的森林覆盖地图。McRoberts (2014)基于 2002 和 2007 年的地球资源卫星图像使用非线性逻辑回归模型来评估 USA 明尼苏达州的东北地

区的森林地区的变化。Kumar (2014)[7]提出用逻辑回归模型通过地球资源卫星图像来分析 1990 到 2000 年森林覆盖变化，同时预测 2010 年的森林覆盖。这些现存的森林覆盖动态模型并没有采集长期且时间连续的森林覆盖动态。因此由于采集到的是这些数据，可能忽视相关驱动因素。一般的，以上模型会被这样构建：假设在时间 t 被观察的土地覆盖的状态被当前所有的解释变量的状态影响着。但是，这个方法忽视了土地覆盖变化会发生在使用数据没有采集的时候，由此就可能会产生有偏差的边际效应评估。最近在处理这个缺陷时又有新的发现。Wheeler 使用了月度数据库来评估印度尼西亚的森林覆盖变化，Marcos-Martinez 在澳大利亚密集型农业区域使用 15 年的步长数据(1988-2010 年)开发空间计量经济模型来评估森林覆盖变化动力，然后使用 2011-2014 四年的时空数据集来验证预测效能。

与此同时，深度学习[8]是传统机器学习家族的一员，该技术在很多方面取得了突破性的进展，例如语音识别[9]，图像识别，自然语言处理[10]，推荐系统[11],以及预测天气，交通和其他方面[12]。但是，深度学习方法目前还没有被应用到森林覆盖模型中。作为新一类的神经网络，它的性能通常超过传统的神经网络[13]，作为四大基础网络架构，深度神经网络有大量的参数和处理层(超过两个处理层)。其他三个架构分别是非监督预训练[14]，卷积[15]和递归[16]神经网络。它们中，有类网络称为长短期记忆 LSTM [17],用来解决梯度消失问题(随着层数增加，网络最终将无法训练)。LSTM 增加携带信息功能可以跨越多个时间步长进行信息交流：能保存时间信息以便后面运用，避免前期的信号在处理过程中逐渐消失。而且应用也十分广泛，例如交通流量预测，电力负载预测[19]，人类活动识别[20],无约束手写识别[21]。通过考虑土地覆盖变化过程中不同时间状态，然后使用 LSTM 网络来提升预测性能。

深度学习模型 LSTM 有预测森林覆盖动态时间序列的能力，将 LSTM 模型与最新的空间计量经济学方法比较评定其预测能力。其中时空数据集是多维高分辨率的，将数据集应用到国家范围的森林覆盖预测集成模型。模型开发中包括使用高性能计算集群给每个 LSTM 网格单元训练。在分析影响因子时，通过排除每个或每组的解释变量后，再评估模型预测能力，从而评估澳大利亚国家森林覆盖的时空动态的因素。

1.3 本论文的选题研究内容

本论文的选题是来源于对澳大利亚森林覆盖的分析和预测。

本论文的研究主题内容有：1.使用 GEE 平台截取 Hahsen 等人绘制的卫星图像集并且分析大致的森林覆盖图表，将得到的卫星图使用 arcGIS 平台进行处理，得到地图分析；2.通过 1988-2014 年澳大利亚森林覆盖的相关数据集，对其先进性相关性分析；3.使用深度学习 LSTM 模型进行预测，同时比较计量空间模型的指标 RMSE 和伪R^2来评定性能优劣；4.分析影响因子过程中，排除各个变量，再运行预测模型将得到的结果进行排序比较，得到影响因子重要性的结果。

1.4 论文组织结构

本文的研究主题是使用深度学习 LSTM 和空间计量这两个模型，在森林覆盖率进行分析预测并将两项进行比较效果，接着分析影响森林覆盖的重要因素。全文一共分为五章。

第一章绪论部分，大致叙述森林覆盖的课题研究背景，研究现状以及研究意义，介绍本论文的选题研究内容，包括本论文的组织架构；

第二章是对相关平台介绍，主要介绍初期分析使用的平台 Google Earth Engine 以及处理地图的工具，将卫星影像处理得到地图，这是量化分析的前提必备步骤；

第三章是对空间计量模型的介绍。首先介绍澳大利亚研究区域森林覆盖状况以及森林的定义，森林覆盖指数以及相关的变量(影响森林覆盖率)，接着具体介绍模型算法的实现，最后总结得到影响森林覆盖的相关因素；

第四章介绍使用深度学习 LSTM 模型对森林覆盖进行预测分析并将结果与上一章的空间模型进行对比，包括相关性分析，模型介绍，训练预测模型的处理流程输入输出数据，最后通过排除各个因素进行性能排序，从而分析影响森林覆盖的因子重要程度；

第五章描述对使用上述工具的总结，基于目前的模型结果，对未来的改进提出了方向建议，为后续研究提供方向。

2 相关平台和数据集介绍

2.1 Google Earth Engine 简介

Google Earth Engine 是 Google 对全球范围提供大量的地球科学资料，能够进行在线可视化分析处理的云平台。它给广大的学术，非盈利，商业和政府组织使用者提供服务。该平台能够存储卫星图像以及其他地球观测数据并将其放入地球观测数据数据库中，其中包括过去 40 年的历史地球图像(已超过 200 个公共数据集和 500 万张影像)。而且图像数据资料库每天都在更新(每天的数据量增加大约 4000 张影像 存储超过 5PB)，并且提供足够的运算能力支持计算处理，使得全球范围的数据挖掘成为可能。与传统影像处理工具如 ENVI 比较起来，GEE 的优势在于快速批量处理大量影像，快速计算 NDVI 降雨量(旱情监测) 农作物产量预测等值。GEE 同时提供了 APIs 能够支持大数据集的分析。不仅有在线平台 JavaScript API 还有离线的 Python API，非常方便用户进行操作。

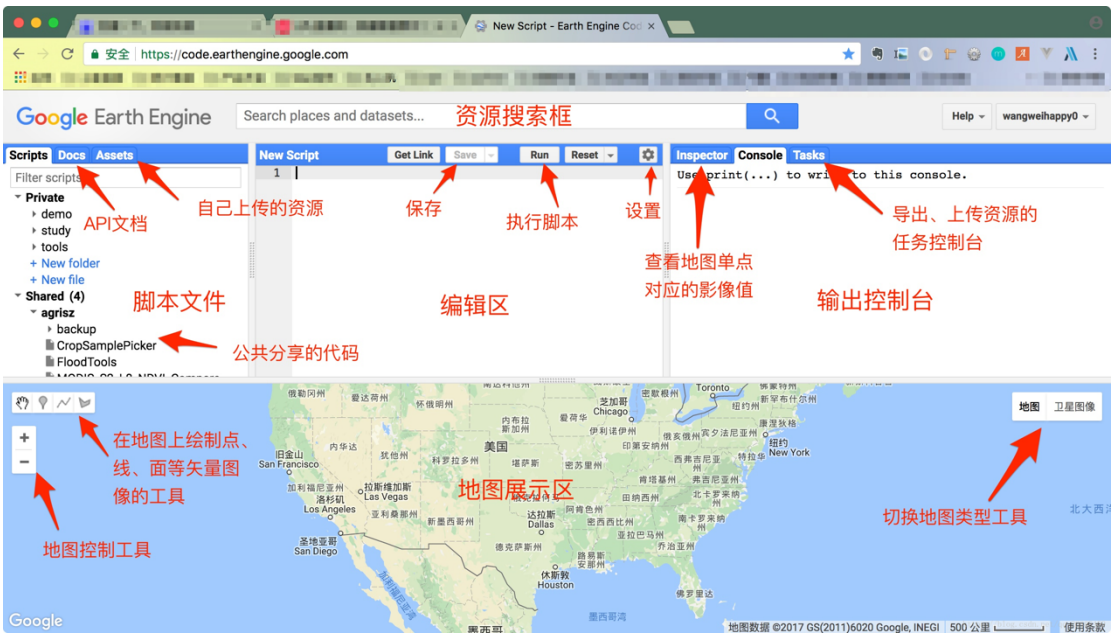


图 2. 1 Google Earth Engine 平台线上操作介绍

2.1.1 Google Earth Engine 平台的森林变化影像

GEE 云平台中收录包括 Hansen 教授绘制的全球森林覆盖的数据影像图和 Global Forest Watch 中的森林监测行为 (FORMA, Hammer et al. 2009) 的数据。可以调用该平台资源进行可视化计算结果，按照年份时间量化森林变化或者其

他地区的统计甚至可以下载数据和图像分析的结果。

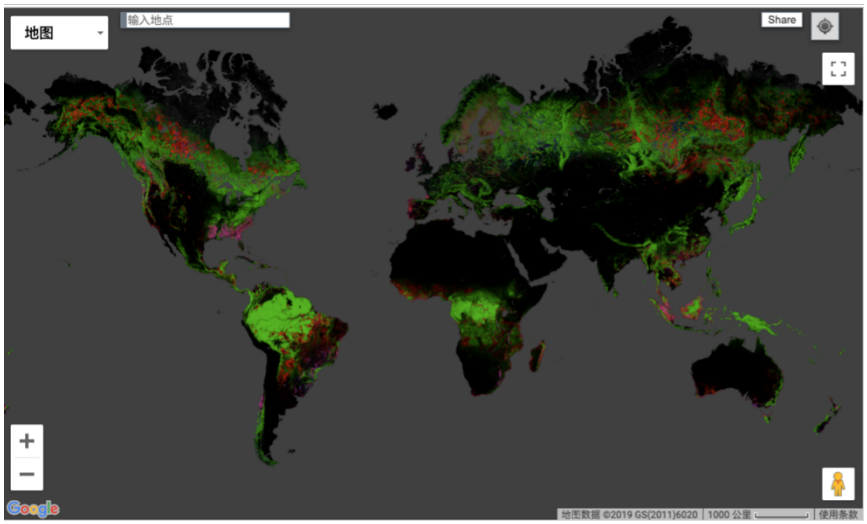


图 2. 2 hansen 绘制世界地图展示

Hansen(2013)在该平台上绘制 2000-2014 年以 30m 为分辨率的全球森林变化数据集，其中该地图集前 3 个波段:treecover2000,loss 和 gain 分别对应影像波段的红色，绿色和蓝色。分别表示 2000 年的森林覆盖，2000 年以后损失的森林面积变化以及 2000 年以后增加的森林面积变化，后两者是在 2000 年后的叠加。在 Google 云平台中也可以预处理影像，对其进行影像色彩处理。针对表 2.1 的各个波段，由于每个波段中每个部分都有不同的值(对应不同颜色)，可以使用色板参数展现不同位置的颜色遮掩；在影像中会有大块黑色区域，为了让该区域透明，可以遮掩这些区域的像素值。

表 2. 1 卫星图像对应图层

波段名	描述	范围
treeover2000	2000 年森林覆盖，显示的是>5m 高度植被范围	0-100
loss	在研究区域 2000-2012 发生森林损失显示	0 或 1
gain	在研究区域 2000-2012 发生森林增长	0 或 1

	显示	
lossyear	森林损失事件的时间，2001 年开始，0 表示没有出现	0-12
first_b30	卫星 7 号波段 3 无云复合图像	0-255
first_b40	卫星 7 号波段 4 无云复合图像	0-255
first_b50	卫星 7 号波段 5 无云复合图像	0-255
first_b70	卫星 7 号波段 7 无云复合图像	0-255
last_b30	卫星 7 号波段 3 无云复合图像	0-255
last_b40	卫星 7 号波段 4 无云复合图像	0-255
last_b50	卫星 7 号波段 5 无云复合图像	0-255
last_b70	卫星 7 号波段 7 无云复合图像	0-255
datamask	无数据(0) 土地结构(1) 永久水位置(2)	0,1,2

2.1.2 分析图像数据

该平台还提供接口函数计算从 2000 开始每年的森林损失面积，以 2000 年为衡量标准，依次将每年收集到的数据与前一年的数据求差即可得到值，并且可以将数据存储到本地，利用平台可视化年森林损失面积图表。由于 hansen 绘制图集一直在更新，图 2.3 使用更新版本得到 2001-2017 年的测量值，并计算出总森林损失数量。

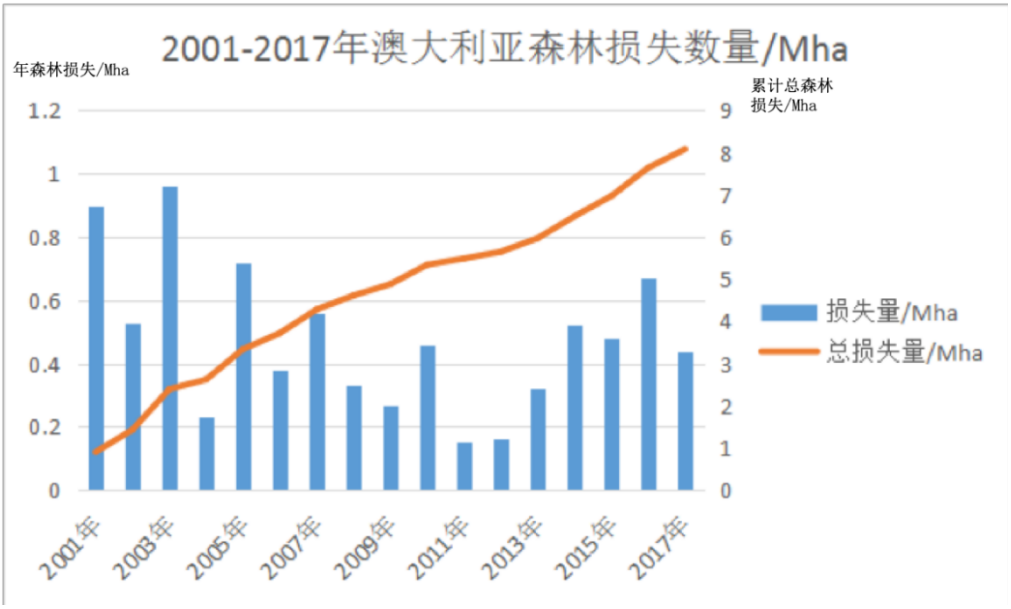


图 2. 3 2001-2017 年澳大利亚森林覆盖损失

2.2 arcGIS 地理信息系统

2.2.1 arcGIS 简介

arcGIS 是一款全面的地理信息系统。该产品已全面整合 GIS 与数据库，软件工程，人工智能，网络技术及其他多方面主流技术，目前在地理信息系统(GIS)构建和应用平台中处于世界领先地位，arcGIS 为全世界的人类提供将地理知识应用到政府，企业，科技，教育和媒体领域的机会。arcGIS 也有全面的，可伸缩的特性，其为用户提供收集，组织，管理，分析地理信息的解决方案。同时可以通过浏览器和移动设备使用本系统。

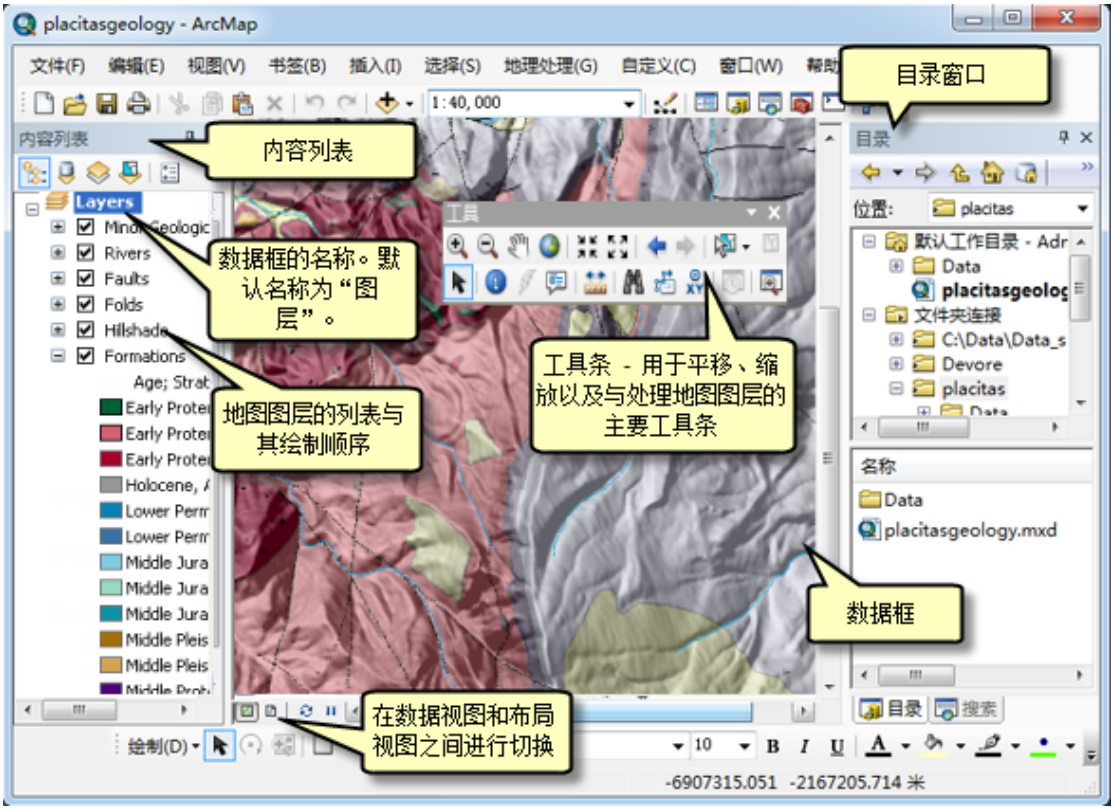


图 2. 4 arcMap 功能页面介绍

2.2.2 地图处理流程

第一，使用 javascript 语言调用相关函数从 GEE 线上平台下载 3 张澳大利亚研究区域的影像图层 treeover2000,loss 和 gain；

第二，提前下载好澳大利亚对应的国界线和地图集，打开 arcMap 软件并对澳大利亚国界进行剪裁，分别对三张图层进行处理；

第三，将裁剪得到的 treeover2000，loss 和 gain 图层分别进行叠加等处理并且需要处理森林损失和增加地区的色调。 由于卫星影像红绿蓝三个波段决定，需要图层无像素值的地方显示为黑色，注意图层的顺序放置，不一样顺序地图效果不同。

第四，将得到的图片，进行图例封装等操作，制作成完整的地图。

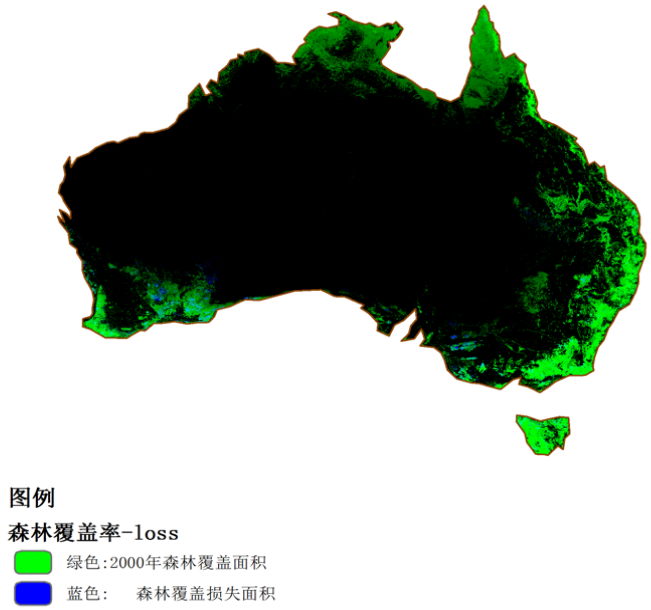


图 2. 5 森林覆盖-损失率展现图

2.3 森林覆盖数据集和解释变量

数据集是从澳大利亚国家碳核算体系-土地覆盖变化计划(NCAS-LCCP)得到，其中共包含 19 段时期(1988,1989,1991, 1992, 1995, 1998, 2000, 2002, 2004,以及 2005-2014 年)，主要以 25m 分辨率的森林确定变量和非确定变量数据集。共有 137 万个观测点，包含澳大利亚 6 个州数据（新南威尔士、昆士兰、南澳大利亚、维多利亚、塔斯马尼亚西澳大利亚州），参数变量多达 309 列（时间变量空间变量和时空变量的时间序列数据集）。

将分析重点放在澳大利亚的集约农业区域，该区域分别占全国农作物和牲畜总价值的 99%和 92%[23]。其中这些区域包括保护区，水文地质（比如水，湖泊）以及其他非农业或森林土地(比如城市地区)。将以上的区域从森林覆盖数据集移除。使用以 25m 为分辨率的森林和非森林数据集来计算以 1.1km 为分辨率的网格单元的森林覆盖指数，这些指数代表每个观测年森林覆盖的比例。使用这些森林覆盖值来研究净森林覆盖变化和森林转型的变量因子。

时空数据汇集在生物物理，社会经济和制度因素上(有文献表明这些因素会影响森林覆盖变化)。因此先将变量分为 4 组：1)时间上不变的，空间改变的(例如，土壤，土地使用权); 2)空间和时间上都变化的(比如气候) ;3)空间不变的 时间变化的(比如，市场价格);4)空间和时间相互作用变量(比如，每个州农场产出

价格指数)

表 2. 2 1988-2014 年澳大利亚农业区环境，经济等变量与净森林覆盖

变量	描述（来源）	单元	分辨率
因变量，空间和时间会变化			
森林覆盖指数	1km 范围内森林占比	%	1.1km
随时间不变，但随空间改变的变量			
自然地理变量			
坡度	地形渐变程度	度数	95m
海拔	高于海平面/m	m	95m
pH	上部 30cm 土壤层 pH 水平	-	250m
粘土含量	上部 30cm 土壤层%粘土含量	%	250m
体积密度	上部 30cm 土壤层体积密度	Mg/m^3	250m
保有期变量			
本土土地	本土拥有土地的指标变量	binary	100m
私有土地	私有土地的指标变量	binary	100m
租赁土地	私有管理租赁土地的指标变量	binary	100m
多用途土地	州土地公共森林的指标变量多用途管理（比如收割木材，水供给）	binary	100m
欧洲前植被(Geoscience Australia, 2004)			
(TSGF)低树	<10m 的平均树高度指标	binary	100m
TSGF 中等树木	10-30m 的平均树高度指标	binary	100m

TSGF 高灌木	高于 2m 的灌木指标	binary	100m
TSGF 高树	平均树木高度>30m 指标	binary	100m
社会经济(短期不会随时间改变变量)			
澳大利亚可获得性和偏远指数	基于从人口密集区域到提供公共和私人服务的市中心的道路距离指数	score	1km
农业利益	2005-06 年观察的农业用地每公顷的利润	2013AU$/ha	1.1km
随着时空变化的变量：时空数据			
降雨量	年降雨量的 5 年移动平均值	mm	0.05°
降雨量变化	年降雨量的 5 年移动平均差	mm	0.05°
最高温度	年最高温度 5 年移动平均值	摄氏度	0.05°
最高温度变化	年最高温度的 5 年移动平均值	摄氏度	0.05°
到保护区的欧几里得距离	每年更新从每个像素点到最近的受保护状态像素的欧几里得距离	km	1.1km
随空间不变，但随时间改变的因素			
木材价格指数	5 年移动平均权重的国际软硬木价格指数	2013 US$	
农场产出价格指标	农业产出价格加权指数的五年移动平均值	score	
农场产出价格指数差异	农业产出价格指数的 5 年差异	score	
二项指标			

州	该指标用来描述未观察的州特定因素	binary	1.1km
州-农产出价格指标	农业价格指数和州指标的相互作用	2013 US$	1.1km
空间数据大部分或最近邻值作确定变量都被转成 1.1km 分辨率			

首先在变量中用特定木材和农产品价格指数的 5 年移动平均值来表示影响净森林覆盖的产出价格预期。将之前的 5 年平均农业价格变量也纳入其中来研究价格不确定性的影响。

气候变化会影响土地利用决策和土地价值(Mukherjee and Schwabe, 2015)[24]。同时计算降雨的 5 年移动平均值和年总降雨量的标准差以及日平均最高温度。将参考使用 2013 年的数据来作为森林所有权和管理标准，虽然在研究期间土地所有权发生变化，但该段时期内土地使用权类型的变化是不明显或不存在的。

如今多个空间领域正发生森林转型，与此同时国家范围的分析可能忽视各州管辖范围的不同森林覆盖模式。这项提议与我们的研究有关，因为在澳大利亚土地使用和森林政策上主要责任方在地方政府（州或者国家政府）。因此，需要用未观测的指标变量来检验影响森林覆盖动态的州特定的变量(比如当地对农业或森林土地的态度，或者州级别的条例)。将上述观测的指标与土地拥有者的价格指数相互作用，来捕获州特定因素对市场信号的限制。

2.4 本章小结

本章主要针对 Google Earth Engine 平台和 arcGIS 软件两个使用工具以及对相关数据集的介绍。通过这两款平台，进行初步的森林覆盖分析，得到卫星影像地图和森林损失图表，进而展现森林覆盖的状态。地图分析，同时数据集也分为 4 种变量，是进一步量化分析森林覆盖的前提步骤。

3 基于空间计量模型 SE 对森林覆盖预测分析

3.1 研究区域森林介绍

根据 National Forest Inventory(1998)对森林的规定：至少有 20%的树冠覆盖，并且初代和二代植被能够达到至少 2m 高可以称为森林覆盖。在该定义下，那些暂时失去森林覆盖的土地仍然被纳入森林植被区域中，即使在数据收集期间内植被没有达到门槛高度(Montreal Process Implementation Group for Australia and National Forest Inventory Steering Committee, 2013)。据统计，澳洲约有 125Mha 公顷的森林，组成 3%的世界森林并在全世界森林覆盖中排名第 7(Australian National Greenhouse Accounts, 2013)。其中森林面积大部分是本土的，只有约 2Mha 的产业种植园(Montreal Process Implementation Group for Australia and National Forest Inventory Steering Committee, 2013)。约有 55%的澳大利亚本土森林在长期租赁或私有土地的管理下，其中 35%是公有管理的，10%由本土组拥有并管理[22]。约 1/3 的私人管理森林位于密集型农业区。澳洲的森林损失和破碎化对鸟，小型哺乳动物，爬行动物以及植物分布产生巨大的负面影响。当然，毁林也是温室气体排放增长的重要来源之一。

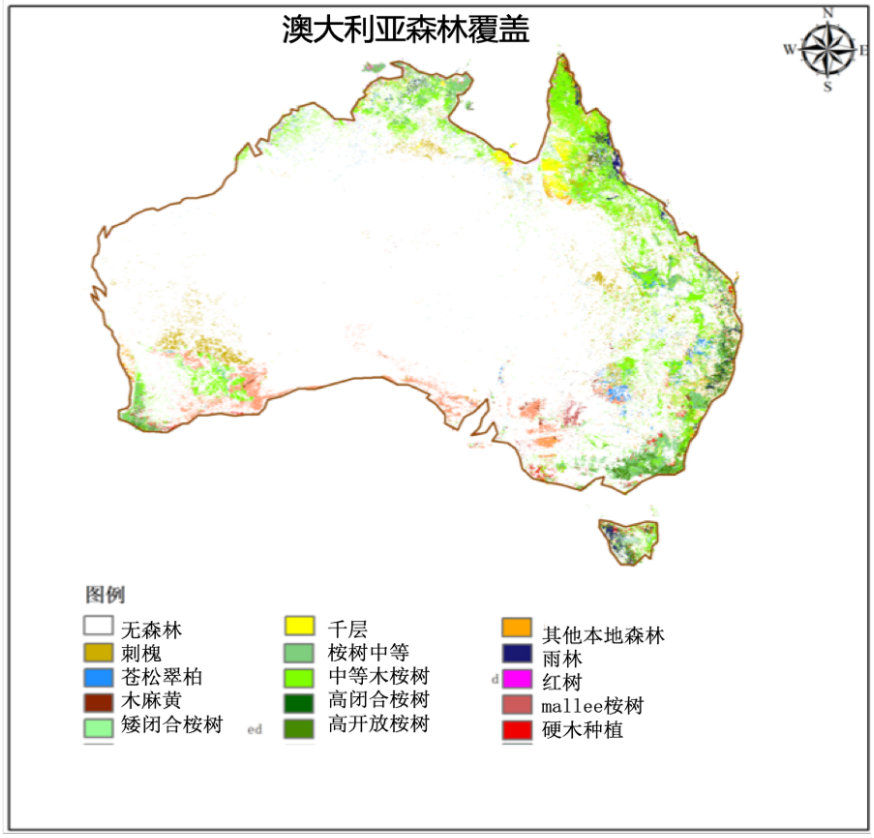


图 3. 1 澳大利亚森林覆盖现状图

3.2 空间计量模型算法实现

3.2.1 计量经济模型

为评估森林覆盖变化的决定因素，我们将采用模型。假设风险中性的土地所有者现在估算森林与非森林土地的净现值，并分配土地以最大化其预期金额。[25]预期金额取决于决策期间生物物理，社会经济和制度等因素的状态。在所有观测中的多种因素里，这些因素会影响森林覆盖，我们模拟在 T 时期 N 个空间单元的模型，包括通过观测值来建模的系统组成部分，误差项(包括在相邻土地中未观察到的空间交互性)，在观测中未观察到的异质性和随机误差。基于迭代空间 hausman 检验的结果，采用随机效果模型的空间面板计算[26]。模型采用以下形式的线性参数规范表示:

$$\mathrm{f} = \beta' X + u \tag{3.1}$$

其中f表明在观测期间森林覆盖以 1.1km 分辨率下的比例向量,时间段表示为$f_{n,t} \in [0,1]$; X指特定像素特定时间的变量矩阵(比如降雨量和农业价格)；β'是系数向量，近似于协变量和结果之间的真实关联 ；u是空间相关未观测到的变量矩阵，可以用以下公式来表示:

$$u = (I_T \otimes pW)u + w \tag{3.2}$$

在(3.2)式中介绍相关参数的含义： I_T是 T 维度单位矩阵; ⊗表示 Kronecker 内积（两个任意大小矩阵间的乘积）；W是 N 大小的空间权重矩阵 ;p 表示为空间自相关系数; w表示由观测特定效应μ组成的复合误差向量。接下来用公式表示w:

$$w = (l_T \otimes I_N)\mu + v \tag{3.3}$$

在(3.3)式中：v表示遵循标准正态分布[27]； μ是观测特定效应；l_T是大小为 T 的向量；I_N是 N 维度单位矩阵。

最后从迭代 Hausman 测试的结果中选择近似未观察到的异质性w的随机效应。

3.2.2 实验方法

对于大量的观测数据集，空间计量模型属于计算密集型，在很小的时间段内计算也可能变得棘手。为了解决这个问题，我们采用 bootstrapping 方法先产生随机样本(尽管受限于观察的森林和非森林模式，仍然包含部分残差空间相关性)，计算空间面板回归。接着从 1000 个样本的回归中计算出平均参数估计和 95%置信区间，其中每个样本由 15 个周期的 10000 个观测组成，也就是每个样本总共 150,000 个记录。使用基于地理距离的特定样本的空间权重矩阵（采用行标准化）来近似观察到的值域与未观测到的协变量之间的空间相互作用。在结果中空间权重矩阵配置的结果(比如:连续性，反权重距离矩阵)没有明显的效果。非确定数据和因变量被转换成森林覆盖指数评估，指数的区间在[0,1]。接着使用 Halvorsen 和 Palmquist 调整半对数方程来计算确定变量的相对效应，采用样本外估计来量化模型预测时间和空间森林覆盖的程度。在最佳线性无偏预测中使用 bootstrapped 边际效应来评估 2011-2014,并使用 1988-2010 的数据进行

预测计算。

模型的初始版本变量包括木材价格，州指标和时间虚拟变量。使用这几种因素来控制土地清理条例和森林激励上的时间差异。然而对应的参数估计在统计意义上不是很重要，在某些例子中产生收敛问题，因此我们选择更简约的方法并从分析中排除那些参数。

3.3 实验结果与分析

3.3.1 森林覆盖结果分析

从 1988 到 2007 年，在研究区域的净森林覆盖损失为 4.02Mha,然而从 2008-2014 净森林覆盖增长了 2.05Mha。几乎所有的变化发生在昆士兰州，该州尽管从 2008 年开始增长，但是在 1988-2014 年期间经历 2.74Mha 净损失。这表示 1988 年这个州有 19%的森林覆盖减少，剩下的研究区域则有相对较小的变化，同时在 1988 年西奥的净损失森林存储为 0.13Mha，Tasmania 和 New South Wales 适量小幅净增长(0.52Mha),Victoria (0.35 Mha)和南奥(0.15Mha).这些趋势导致在 2008-2009 年研究区域完成森林转型。

3.3.2 影响森林覆盖动态因素

计量经济模型解释净森林变化方差比例之高(R^2 = 0.91)，大部分可解释变量在 95%水平的置信区间内。几乎所有不随时间变化的如地形，土地使用期限和前欧洲地区植被变量在解释净森林覆盖空间差异上具有统计意义。与不确定变量的样本均值相比，(表 3. 1 的非分类变量)样本均值中 1%边际影响变化表明，研究区域内更多的森林覆盖率与更高的坡度，更低的土壤粘土含量和 pH 值以及更大的土壤容量相关。这和过去强调的因素影响一致，这些因素包括土壤结构，管理，产出和农业使用的适应性研究[29]。森林覆盖和海拔之间的负相关可能是研究区域特定地形起伏的结果，因为大部分是低洼地带而且更高海拔地区是高原而不是大山。这些高原通常有非常多的雨季，同时属于草地林地生态系统，目前这些生态系统已被广泛用于放牧。

从土地保有期角度，私有土地的森林覆盖率低于任何其他土地类型。与本土土地管理做法相比，本土森林覆盖率较低可能由于限制森林增长的生物物理性质(比如贫瘠，酸性以及干旱的气候)造成的。

对于森林覆盖与保护区的距离效应刚好相反，可能由于保护区网络的溢出

效应也可能是保护区经常位于低风险的森林砍伐区域[30](该区域会让商家承受高运输成本)与过去的发现一致。虽然木材价格指数在统计中不是很显著，但人们可能会出于多种原因预期这点。本土森林的收获在大部分研究区域是可选择的(比如，只有商业价值的个别树种被砍伐，没有严重影响到森林树冠层)。此外，大部分被国内木质工业使用的木材由种植园森林提供，砍伐一般需要合同承诺对树龄限制也有要求，而不会受到短期价格波动的影响。高利润农业地区有更低的净森林覆盖，虽然农产品价格上涨变化与增加的森林覆盖相关。

我们研究的气候变量表明时空变化的气候条件对森林覆盖有组合影响。增长的降水量导致森林覆盖的增加，也反映绝大多数澳大利亚农业(土地)情况处于旱地地中海到半干旱气候带。森林覆盖率与最高温度之间呈负相关又与澳大利亚观察的空间森林分布相对应，在温暖干燥的气候下平均树木数量少于凉爽潮湿地区。结果也反映类似效应：大的年际温度变化可能对树生长有影响(比如，通过热量/水压力)。

州管辖变量解释了其他未知因素，这些州级别因素的差异可以影响森林覆盖。结果大致与观察到的现有的森林覆盖率与南澳大利亚差异相对应，南澳大利亚由于有效的毁林法案，导致研究期间森林覆盖率的变化相对较低。这就可以量化各州的农产品价格波动对异质森林覆盖反馈。西澳大利亚和 Tasmania 森林覆盖的农产品价格变化的影响在统计上不是很重要，但是在昆士兰州，净森林覆盖率对农产品价格变化的反映相对较强且为负相关。

3.4 本章小结

本章主要介绍空间计量经济模型在澳大利亚各州的预测实验。介绍研究区域森林，森林覆盖指数数据和影响森林覆盖的解释变量。着重阐述空间计量算法的公式和方案，通过实验得到结果，主要分析森林覆盖动态以及影响森林动态分布的因素。

4 基于深度学习 LSTM 模型对森林覆盖预测分析

4.1 森林覆盖动态预测的 LSTM 模型

我们使用的数据集和空间计量模型相同，基于之前描述的森林和非森林数据给每个网格单元建立并训练特定的 LSTM 网络来学习和预测森林覆盖动态。LSTM 被 Hochreiter 和 Schmidhuber 引入来解决基础的坡度消失以及在传统的递归神经网络中无法建模长序列的问题。可以使用 LSTM 来采集长期依赖时间序列关系。

模型结构中采用具有遗忘门的 LSTM 来给 RNN 神经网络进行降噪。LSTM 模型空间网格单元 g 包括 1 个输入层，3 个 LSTM 层和 1 个最高层的输出层(看左手边的图)。

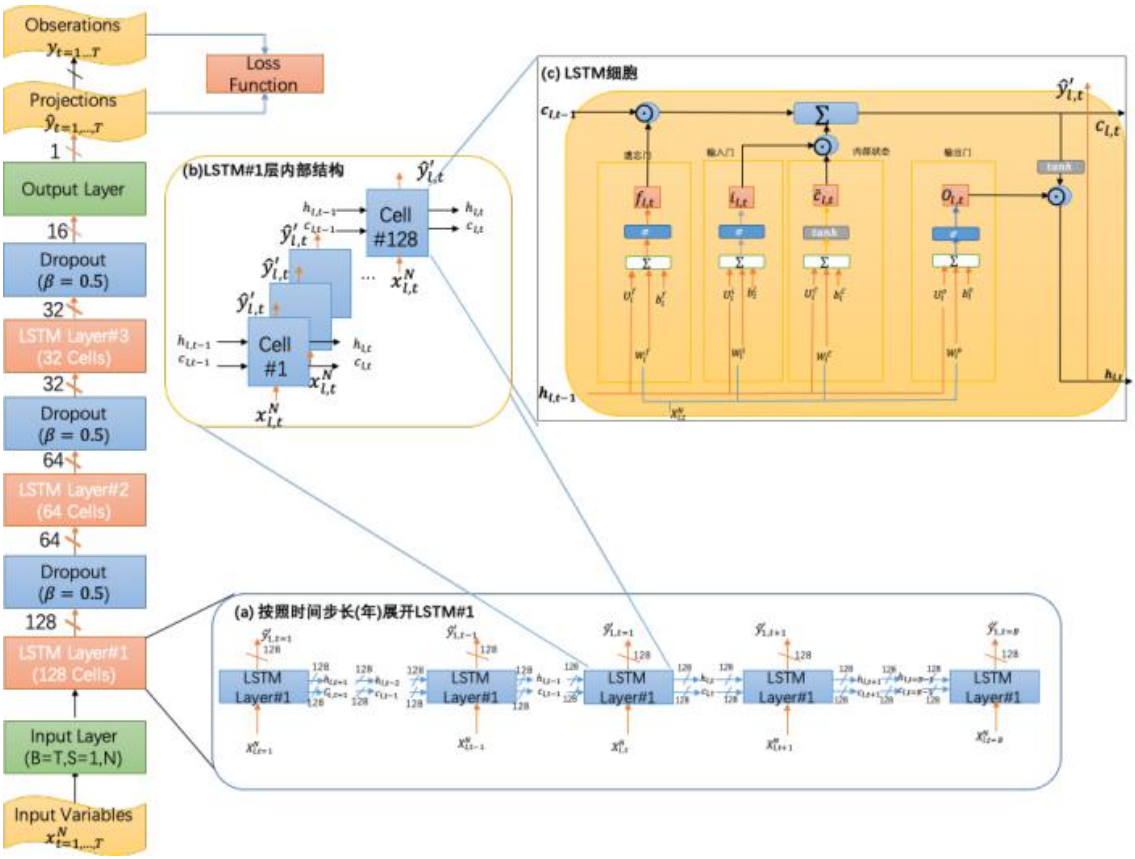


图 4. 1 森林覆盖预测网格单元的单 LSTM 的结构模型

$N * T$个变量向量$X_t = {}^{N}_{1,\dots T}$，表示 T 年 N 个变量(时间，空间，时空变量)，

由此形成输入层。我们使用时间步长为 15 的时间序列数据集，步长横跨共 23 年时间段，1988-2010 作为训练集，2011-2014 作为测试集（由于有些年份数据缺失 1988-2010 筛选 15 个 2011-2014 有 4 个，共 19 个数据集年份）。接着重塑输入层的维度$X_t = {}^{N}_{1,\dots T}$，使得输入变量变为三维模式(B,S,N)。B,S 分别是批组大小(每期迭代训练例子的数量 B = T)以及年数的宽度(S=1,表示每年为一个变化步长)。

使用$\beta = 0.5$的 dropout 操作并将其运用到每层 LSTM 的输出来防止 LSTM 发生过拟合。因此，这三个 LSTM 层各自有 128，64 和 32 个记忆单元。作为典型的递归神经网络，一个 LSTM 网络有循环，其允许信息从网络的一阶段到下阶段传递。

展开的 LSTM 层在图 4.1(a)中展示：$x_{l,t}^{N}$表示 N 维输入变量在年 t 第 l 层 LSTM； $\widehat{y_{l,t}}$是年 t 在 l 层的 LSTM 的输出向量；$c_{l,t}$和$h_{l,t}$是各自表示年 t 在 l 层 LSTM 的单元状态以及隐藏状态。LSTM 层更新$c_{l,t}$和$h_{l,t}$同时计算基于$x_{l,t}^{N}$的输出$\widehat{y_{l,t}}$,前一个记忆单元状态$c_{l,t-1}$和前面的隐藏状态$h_{l,t-1}$。在第一层共有 128 个记忆单元；因此，在每个阶段 t 都产生 128 个记忆单元状态和 128 个隐藏状态。过去的输入信息通过记忆单元状态$c_{l,t}$和隐藏状态$h_{l,t}$被水平传播到每层 LSTM，内部结构显示在图 4.1(b),同层的 128 记忆单元中相互之间没有交流。每个记忆单元执行各自的时间步长(year)迭代。同时每个单元状态的输出$\widehat{y_{l,t}}$在 dropout 之后被一层层垂直传播。

记忆单元主要由以下元素构成：输入门$i_{l,t}$,自循环连接，遗忘门$f_{l,t}$和输出门$o_{l,t}$。W_l, U_l, b_l分别是输入和隐藏层之间连接的权重矩阵，隐藏层之间自循环连接的权重矩阵，以及第 l 层的 LSTM 偏置向量。上标i, f, o各自表示输入门，遗忘门和记忆单元的输出门。

在记忆单元中，输入门$i_{l,t}$被定义为：

$$i_{l,t} = \sigma\left(W_{\mathrm{l}}^{i}.x_{l-1,t}^{N} + U_{l}^{i}.h_{l,t-1} + b_{l}^{i}\right) \tag{4.1}$$

$i_{l,t}$通过门激励函数σ被设置成范围[0,1](使用 logistic sigmoid = $\frac{1}{1+e^{-x}}$)作为记忆单元中的门激励函数)。当数据信号没有通过时，这个值等于 0，当信号完全通过显示为 1。类似的，遗忘门$f_{l,t}$和输出门$o_{l,t}$被计算成以下:

$$f_{l,t} = \sigma\left(W_{l}^{f}.x_{l-1,t}^{N} + U_{l}^{f}.h_{l,t-1} + b_{l}^{f}\right) \tag{4.2}$$

$$o_{l,t} = \sigma\left(W_l^f . x_{l-1,t}^N + U_l^f . h_{l,t-1} + b_l^f\right) \quad (4.3)$$

隐藏状态$h_{l,t}$和记忆单元状态$c_{l,t}$都表示时间 t 经过当前状态到下个时间阶段(t+1)过度状态。隐藏状态$h_{l,t}$也是时间 t 到下层记忆单元的输出$\hat{y}'_{l,t}$(图 4.1(c))。隐藏状态$h_{l,t}$和记忆单元状态$c_{l,t}$被如下公式表示:

$$h_{l,t} = o_{l,t} \odot tanh\left(c_{l,t}\right) \quad (4.4)$$

$$c_{l,t} = f_{l,t} \odot c_{l,t-1} + i_{l,t} \odot \widetilde{c_{l,t}} \quad (4.5)$$

⊙是两个向量的元素乘积；$\widetilde{c_{l,t}}$是记忆单元的内部状态，在(4.6)中表达；$f_{l,t} \odot c_{l,t-1}$表明在记忆单元状态中遗忘旧的或是之前的信息; $i_{l,t} \odot \widetilde{c_{l,t}}$表明记忆单元状态中添加当前或是新的信息。

$$\widetilde{c_{l,t}} = tanh\left(W_l^{\tilde{c}} . x_{l-1,t}^N + U_l^{\tilde{c}} . h_{l,t-1} + b_l^{\tilde{c}}\right) \quad (4.6)$$

在(4.6)中双曲正切(tanh)被用来作为内部状态$\tilde{c}_{l,t}$激励函数。
输出层是全连接神经层，可以被表达成(4.7)

$$\hat{y}_t = \sigma\left(W^D . \hat{y}_{l=3,t} + b^D\right) \quad (4.7)$$

W^D是第三层 LSTM $\hat{y}_l = {}'_{3,t}$的输出权重矩阵, b^D是输出层的偏置。在这里选择 logistic sigmoid $\frac{1}{1+e^{-x}}$作为输出层的激励函数σ,然后得出在 t 年预测的森林覆盖预测指数$\hat{y}_t$。同时使用均方差公式(4.8)作为损失函数来控制训练过程: y_t表示年 t 的观察值。

$$Loss\left(\hat{y}_{t=1,\ldots,T}, y_{t=1,\ldots,T}\right) = \frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2 \quad (4.8)$$

4.2 训练预测模型方法:正则化，dropout 和输入数据

在实际工程应用中，有研究称 Normalization 过程可以极大改善神经网络模型的结果。因此将数据进行统一的归一化处理，使得 N 维输入的每个维度 d 都

在[0,1]的数据比例中，参照公式(4.9):

$$z_t^d = Norm\left(x_t^d, x^d\right) = \frac{x_t^d - Min(x^d)}{Max(x^d) - Min(x^d)} \quad d \in N \qquad (4.9)$$

其中z_t^d是年 t 正则化后的 d 维变量，x^d表示从 1988 年到 2014 年 d 维输入变量的全集。在深度学习网络中，比如 LSTM 中，会导致过拟合问题,就是说在训练时可以呈现很好的性能但是在验证测试集时表现非常差。为了应对这种困难，采用 dropout 的方法来防止 LSTM 模型发生过拟合。该方法的关键思想是在训练期间从 LSTM 网络中丢弃一些单元和连接。为在建模序列中保存 LSTM 网络的能力，采用 Pham 等人的方法并且在每层 LSTM 后采用 dropout($\beta = 0.5$)方法。

在所有空间和时间变量的基础上，需要比较训练过的 LSTM 网络对每个网格单元的预测性能。因为空间变量中每个网格单元随着时间变化保持不变，便将这些变量排除在 LSTM 训练中。主要采用时空序列变量(总共 11 个变量)作为 LSTM 网络的输入。在国家范围上我们将整套 LSTM 网络的性能和空间计量经济模型（该模型使用所有的空间 时空 和时间变量来训练）进行比较进而得到森林覆盖预测的结果。

4.3 预测性能评估标准

我们在不同州的网格单元上评估模型的预测性能。参考指标有两个 root-mean-square error(RMSE)和 pseudo R-squared。RMSE 可以称为标准差，R-square 则是修正的可决系数，越接近 1 越好。指标被用来评估 LSTM 和空间计量经济模型(也叫 SE)预测性能。

我们也使用 RMSE 作为指标来比较 2011-2014 年不同州的预测性能。在网格单元上，我们计算预测和观测值间的差值作为预测误差，同时映射从 2011-2014 年共 1376866 个网格单元的平均预测误差e_g的空间分布。对于网格单元 g, 可以计算得到 2011-2014 年平均预测误差e_g。当网格单元$e_g \geq 0.3$称为高估，相反当网格单元$e_g \leq -0.3$称为低估。

4.4 识别影响输入变量方法

识别重要的变量有助于了解森林损益的原因，丢弃多余的输入变量并改进预测模型，告知相关投资方需要收集和分析额外数据来改进模型预测，并为改

进森林覆盖动态建模机制提供可行性。开发量化变量方法用来得到森林覆盖预测模型。方法的中心思想是:(1)从森林覆盖预测模型输入的全部变量集 N 中移除单个或者组变量；(2)使用新的变量集来训练新的 LSTM 网络；(3)量化变量重要指标；(4)在量化重要指标中排序变量。通过这种方式，在森林预测模型中从 N 个变量中移除某个变量($termed\ x_{N-1}$)，进而评估每个单变量的重要性。

考虑到某些时空变量可能展现很强的相关性，也采取移除整组的 m 个相关变量($termed\ x_{N-m}$)来评估组变量的重要性。评估的组变量名称被显示在表 4. 1 中，其代表气候的主要影响。例如 rain5SD, rain5MA 和 rain 都代表降雨量的影响，同时当我们只移除单变量，降雨量信息可能被纳入预测模型中。因此，我们也训练剩余的(N-1)或(N -m)个变量作为新 LSTM 网络，同时计算在 RMSE 和 pseudo R-squared 方面的变量重要指标。

表 4. 1 相关变量组

组名	组描述	包含变量
相邻效果	包括之前相邻网格单元的平均森林指数，表示网格单元空间持续性	fidx10NN fidx3NN
降雨	包括年降水量信息	rain5SD rain5MA rain
最大温度	包括年最大温度信息	temp5SD temp5MA tempMax
价格	包括经济因素	priceRec timberPI

4.5 实验运行

我们使用 Tensorflow 1.0 版本给所有 1376866 个网格单元建立 LSTM 模型。1988-2010 年和 2011-2014 年的数据各自被用来作为训练和测试集。创建并计算每个 LSTM 模型并且将任务提交给 Linux 高性能计算机集群，让集群并行执行。给每个 LSTM 模型的网格单元训练 100 期，并使用随机梯度下降法来减小损失，同时也由于计算数据量大，使用随机梯度下降来模拟抽取等同数量的计算。

4.6 实验结果与分析

4.6.1 森林覆盖在国家州和网格单元比例预测效能评估

LSTM 模型在 RMSE 和 pseudo R-squared 都优于 SE 模型(表格 3)，相比 SE 模型，由 LSTM 模型产生的 2011-2014 年的 average RMSE 从 0.1248 下降到 0.070(下降了 44%)。同时 average pseudo R-squared 从 0.848 增加到 0.952(增加了 12%)

表 4. 2 LSTM 和 SE 模型的预测效率指标

	RMSE		Pseudo R^2	
Year	SE	LSTM	SE	LSTM
2011	0.1143	**0.0641**	0.8717	**0.9596**
2012	0.1211	0.069	0.8572	0.9536
2013	0.1313	0.0726	0.8327	0.9489
2014	0.1325	0.0759	0.8302	0.9443
Average	0.1248	0.0704	0.8480	0.9516

在 LSTM 模型下，2011 - 14 年间大约 81％网格单元的平均森林覆盖预测误差在±5％之内，并且 90％的网格单元具有±10％的平均预测误差（表 4. 3）。

此外，只有 1.5%的网格单元被过分低估或高估（其平均预测误差大于±30%）。与 LSTM 相比，SE 表现的较糟糕，网格单元中各有 69%(± 5%)和 79%(± 10%)平均预测误差；4. 5%的网格单元被过度低估或高估。

就每个州高估和低估的网格单元的百分比而言，LSTM 的表现优于 SE（表 4）。 然而，除了在昆士兰州，LSTM 在其他任何州都有过分高估的现象，但被低估的网格单元比 SE 低。

表 4. 3 LSTM 和 SE 模型不同范围的平均预测误差网格单元数

state	LSTM								
	e_g								
	(0.3,∞)	(0.1,0.3]	(0.05,0.1]	(0,0.05]	0	[-0.05,0)	[-0.1,-0.05)	[-0.3,-0.1)	(-∞,-0.3)
NSW	1702	11366	16484	217363	47678	98954	22039	17563	2017
VIC	407	1570	1986	60492	13912	37932	5947	4415	1499
QLD	4577	30330	28906	188635	31638	25073	4681	3885	960
SA	309	1790	2308	76623	18918	25073	4681	3885	960
WA	1817	9277	10011	110065	21042	55026	9063	9266	2251
TAS	278	2215	2736	10861	869	8785	1629	1771	419
Total	9090	56548	62431	664039	134057	310339	67806	60841	11715

state	SE								
	e_g								
	(0.3,∞)	(0.1,0.3]	(0.05,0.1]	(0,0.05]	0	[-0.05,0)	[-0.1,-0.05)	[-0.3,-0.1)	(-∞,-0.3)
NSW	811	4938	6742	77202	147514	87853	37809	66325	5972
VIC	56	474	516	8907	55882	27810	8067	17911	8537
QLD	33152	73735	30461	197830	10685	50480	11044	10935	3290
SA	80	792	818	11625	74423	26092	7187	11529	2001
WA	614	3883	4322	46803	56715	52519	21907	35095	5960
TAS	204	1885	2304	9931	1453	8267	2614	2215	690
Total	34917	85707	45163	352298	346672	253021	88628	144010	26450

统计结果显示，LSTM 模型的预测性能在每个州都优于 SE 模型（图 4.2,4.3）。与其他州相比, TAS(Tasmania)的 RMSE 值在 LSTM 模型很显著，然而 RMSE 指标在 QLD（昆士兰州）对于 SE 模型很显著。在不同州范围上使用

LSTM 模型，SA（南澳大利亚），NSW(New South Wales)和 VIC(Vitoria)在预测效能上排在前三，而 WA ，QLD 和 TAS 的预测性能低于国家预测水平。

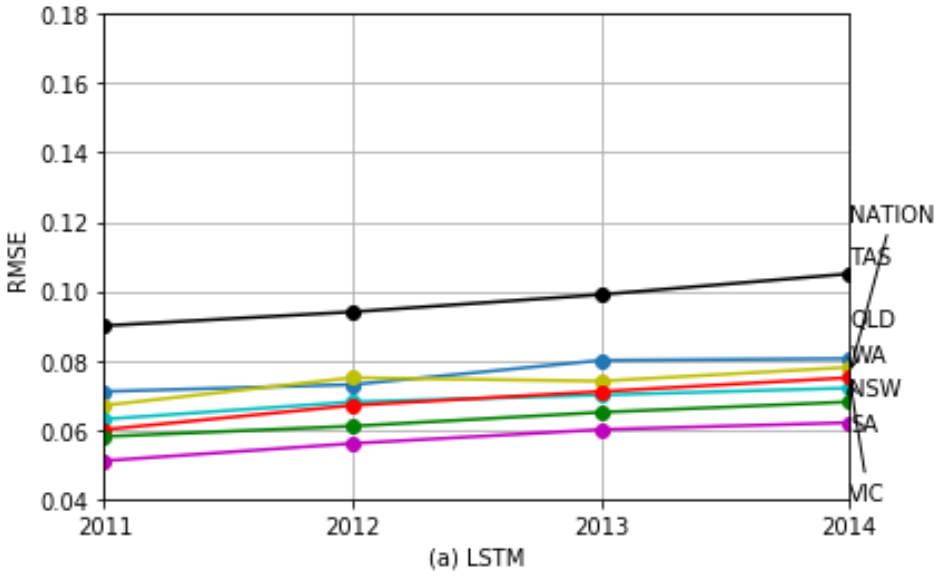


图 4. 2 2011-2014 年 LSTM 森林覆盖预测性能

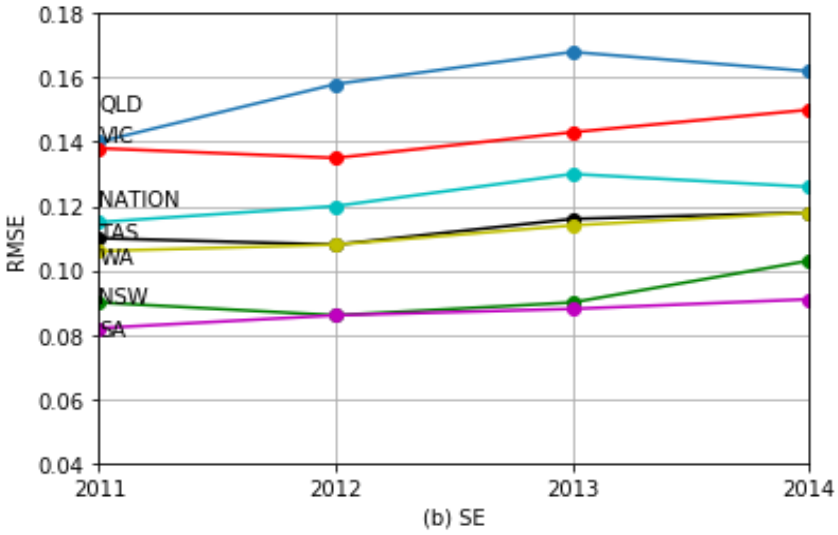


图 4. 3 2011-2014 年 SE 的森林覆盖预测性能

4.6.2 森林覆盖变化的重要影响变量

在单变量和组变量中 RMSE 和 Pseudo R^2都是一致的。排除任何变量或组变量，新的预测性能没有显著减小。平均 RMSE(0.0687,0.0785)和 Pseudo R^2

(0.9399,0.9537)在这之间变化。

排名前五的影响变量和影响组分别是 Neighborhood effects, priceRec, fidx3NN, Prices,和 Maximum temperature。对于变量组 prices,变量成分 priceRec 比组更有影响力。对于其他变量组(Neighborhood effects, Maximum temperature 和 Rainfall),整个组总是比部分变量要更有影响。除了这三个变量(rain, rain5SD, and timberPI)可以稍微提升预测性能。

4.7 讨论与总结

4.7.1 模型预测性能小结

实验结果显示使用深度学习技术处理时间序列变量可以产生比空间计量经济学模型更好的预测效果。结果证实，整套 LSTM 可以采集到长期时间连续的动态，比如森林覆盖率的增长和损失的程度。在预测模型中能采集高分辨率，空间差异化影响森林覆盖的方式，但是预测性能也随着网格单元发生变化，由于 LSTM 网络和空间多元化的解释变量间的复杂作用。总体来说，深度学习方法能展现不同州水平的预测性能，结果反应空间变量在解释变量和模型预测中的森林动态效果。

对大部分网格单元来说, 2011-2014 期间平均预测误差在可接受范围内，很少有网格单元的平均预测误差>30%。例如，在毁林的热点地区-昆士兰州，只有 2.1%的网格单元被过度高估或低估（指平均预测误差>± 30%）。这些有误差的网格单元经历了森林覆盖中的突增或骤减，误差发生在广阔的森林砍伐和再生的网格区域特征中，或由于卫星图像的伪影误差问题。森林增长主要发生在人工种植区域，而减少主要在本土的桉树和金合欢林种植区。这表明在网格单元中需要采集更多有影响力的因素，例如土地清理效应和森林砍伐禁令的积极溢出效应。

4.7.2 深度学习的分析

作为一类学习方法，深度学习可以自动探索原数据中的特征。在实验中也证明这个性质，LSTM 网络可以处理噪声和连续数据。但是，实验里也证明 LSTM 这方面的能力也有限制。比如，训练好的(包括变量 timberPI)LSTM 网络获得的预测性能比那些排除该变量的网络更糟糕。因此需要搜索输入变量来提升预测性能。

该模型由 1.38M 个 LSTM 网络构成，每个网络都有 3 层 LSTM 和一个全连接输出层。训练和诊断每个网格单元模型，同时这些模型是密集计算型的(需要大量浮点运算和高性能并行 I/O 能力)，因此需要高性能计算空间。运行预测模型在高性能并行计算节点的 CSIRO 的 Linux 集群上，要求在规定时间内得到执行结果。而无法接触到这种高性能计算资源的用户可能会遇到计算任务规模的挑战。

4.7.3 实验的优劣势评价

我们为森林覆盖动态建立一个自底向上的预测模型且在不同州进行处理。装配大约 1.38 百万个 LSTM 网格单元。与大部分在空间森林覆盖动态方面的预测模型相比，我们的预测模型提供更多细节：用分辨率为 1.1km 的空间网格单元建立 LSTM 模型网络，该模型能够预测森林覆盖动态，其不单覆盖巨大的地理范围，也随时间在生物物理和社会经济驱动因素中采集重要的变量。这解决重要的限制，即之前使用非时间连续且粗略的时间步长数据来分析土地使用变化。我们的时间分辨率模型采集包括长期且时间连续的动态因素数据集。

另一个优势是“即插即用”功能。不同的机器学习方能在不同的网格单元使用。通过给每个网格单元选择最好的预测模型来提升国家比例的预测性能将成为可能。在影响变量识别实验中也证实了这点，可以得到森林预测模型的健壮性和有效性。当把影响的变量或者变量组从训练集移除时，整套模型仍可以获得更好的预测性能(average RMSE 值(0.0687,0.0785) PseudoR^2(0.9399,0.9537))，相比空间计量经济学模型(average RMSE = 0.1248 and Pseudo R^2= 0.8480)。LSTM 模型的“重新加权”剩余变量可以补偿遗漏变量中信息的丢失。然而 SE 模型不能那样做。

虽然深度学习模型在大部分网格单元中表现很好，但这些单元主要是在森林覆盖变化随时间发生缓慢或保持相对稳定区域，由意外发生的骤增或突减的网格单元中并没有表现很好。因此需要纳入更多的解释变量(例如在土地清理条例的变化)，进而优化模型。

5 展望与总结

本论文首先对分析森林覆盖的工具进行介绍，对澳大利亚进行可视化分析，接着提出空间计量模型和深度学习模型来分析澳大利亚森林覆盖的影响因素和预测结果，对数据集先进行相关性分析预处理，然后采用复杂的深度学习方法成功解决预测长期时间连续的森林覆盖动态的问题。在不同州采用 LSTM 预测模型，其中包括 1.38 百万个网格单元的子模型。每个网格单元模型由 LSTM 模型包括三层 LSTM 和一个全连接输出层建成，同时将项目运行在高性能计算集群上。深度学习模型极大的超越空间计量经济模型，在 RMSE 指示上提升了 44%同时在 pseudo R-squared 提升了 12%。结果也很明显证实：在广泛变量中深度学习预测模型有健壮行和有效性。解释变量的相关重要性与相关观测的森林覆盖动态也被展现出来。

当然该 LSTM 森林覆盖预测模型中也有两个限制和进一步提升的地方：(1)在模型构建中，纳入了训练模型过程中不随时间变化的空间变量，这是没有意义的(也就是表 3.1 空间变量),因为它们对每个网格单元仅以常量出现并且没有提升预测性能。有研究试图证明空间(非时间)变量在森林覆盖预测有使用价值。包括空间变量的深度学习网络方法需要被开发来提升预测性能。训练单个深度学习网络并纳入所有网格单元可能成功，但在并行计算负载方面会成为挑战;(2)深度学习模型在大部分网格单元中表现的很好，这些单元中森林覆盖变化随着时间发生缓慢或者保持相对稳定，但在意外发生的骤增减的网格单元中并没有表现的很好。需要寻找更多的解释变量(例如土地清理条例的变化)。

以上是本次毕业设计的不足之处，对于这些点没有可行的解决方法，在预处理模块，采取统一模式进行数据处理可能不妥，需要多方比较，深度学习模型选用和调参数也是如此。在实际的工程应用中，肯定是要使用多种不同的算法工具来得到准确率最高的方法，而且要与实际情况相结合。由于本人水平有限目前没有掌握多种预处理算法和深度学习框架。

本毕设是深度学习的尝试，本人有兴趣对机器学习进行学习研究，在以后学习过程中，将理论、实践结合起来，希望以后能在这个方向上越走越远。

参考文献

[1] Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Goetz, S.J., Loveland, T.R., Justice, C.O.,Townshend, J.R.G., 2013. High-resolution global maps of 21st-century forest cover change. Science 342(6160) 850-853.

[2] Ceddia, M.G., Sedlacek, S. 2013. Sustainable agricultural intensification or Jevons paradox? The role of public governance in tropical South America. Glob. Environ. Chang. 23 (5) 1052–1063.

[3] Laurance, Cassman, K.G., 2014. Agricultural expansion and its impacts on tropical nature. Trends Ecology and Evolution 29(2) 107-116

[4] Matthews, R.B., Gilbert, N.G., Roach, A., Polhill, J.G., Gotts, N.M., 2007. Agent-based land-use models: a review of applications. Landscape Ecology 22(10) 1447-1459.

[5] Freitas, S.R., Hawbaker, T.J., Metzger, J.P., 2010. Effects of roads, topography, and land use on forest cover dynamics in the Brazilian Atlantic Forest. Forest Ecology and Management 259(3) 410-417.

[6] Kamusoko, C., Wada, Y., Furuya, T., Tomimura, S., Nasu, M., Homsysavath, K., 2013. Simulating Future Forest Cover Changes in Pakxeng District, Lao People's Democratic Republic (PDR): Implications for Sustainable Forest Management. Land 2(1) 1.

[7] Kumar, R., Nandy, S., Agarwal, R., Kushwaha, S.P.S., 2014. Forest cover dynamics analysis and prediction modeling using logistic regression model. Ecological Indicators 45 444-455.

[8] Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural Networks 61 85-117.

[9] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. International Conference on Machine Learning, pp. 173-182.

[10] Hirschberg, J., Manning, C.D., 2015. Advances in natural language processing. Science 349(6245) 261-266.

[11] Elkahky, A.M., Song, Y., He, X., 2015. A Multi-View Deep Learning Approach for Cross Domain User Modeling in Recommendation Systems. International World Wide Web Conferences Steering Committee, pp. 278-288

[12] Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F.Y., 2015. Traffic Flow Prediction With Big Data: A Deep Learning Approach. IEEE Transactions on Intelligent Transportation Systems 16(2) 865-873.

[13] Lipson, H., Kurman, M., 2016. Driverless: intelligent cars and the road ahead. Mit Press.

[14] Schmidhuber, J., 1992. Learning Complex, Extended Sequences Using the Principle of History Compression. Neural Computation 4(2) 234-242.

[15] LeCun, Y., Bengio, Y., 1995. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks 3361(10) 1995.

[16] Medsker, L., Jain, L., 1999. Recurrent neural networks: Design and Applications 5.

[17] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9(8) 1735-1780.

[18] Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks, Advances in neural information processing systems, pp. 3104-3112.

[19] Kong, W., Dong, Z.Y., Hill, D.J., Luo, F., Xu, Y., 2017. Short-Term Residential Load Forecasting based on Resident Behaviour Learning. IEEE Transactions on Power Systems 33(1) 1.

[20] Ordóñez, F.J., Roggen, D., 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors 16(1) 115.

[21] Graves, A., Schmidhuber, J., 2009. Offline handwriting recognition with multidimensional recurrent neural networks, Advances in neural information processing systems, pp. 545-552.

[22] Kanowski, P.J., 2017. Australia's forests: contested past, tenure-driven present, uncertain future. Forest Policy and Economics. 77, 56–68.

[23] Marcos-Martinez, R., Bryan, B.A., Connor, J.D., King, D., 2017. Agricultural land-use dynamics: assessing the relative importance of socioeconomic and biophysical drivers for more targeted policy. Land Use Policy 63, 53–66.

[24] Mukherjee, M., Schwabe, K., 2015. Irrigated agricultural adaptation to water and climate variability: the economic value of a water portfolio. Am. J. Agric. Econ. 97 (3), 809–832.

[25] Mather, A.S., 2007. Recent Asian forest transitions in relation to forest transition theory. International Forest Review 9 (1), 491–502.

[26] Kapoor, M., Kelejian, H.H., Prucha, I.R., 2007. Panel data models with spatially correlated error components. Journal of Econometrics. 140 (1), 97–130.

[27] Elhorst, J.P., 2014. Spatial econometrics: from cross-sectional data to spatial panels. Springer.

[28] Elkahky, A.M., Song, Y., He, X., 2015. A Multi-View Deep Learning Approach for Cross Domain User Modeling in Recommendation Systems. International World Wide Web Conferences Steering Committee

[29] Ferretti-Gallon, K., Busch, J., 2014. What Drives Deforestation and What Stops It? A Meta-Analysis of Spatially Explicit Econometric Studies. Washington, D.C(20036).

[30] Andam, K.S., Ferraro, P.J., Pfaff, A., Sanchez-Azofeifa, G.A., Robalino, J.A., 2008.Measuring the effectiveness of protected area networks in reducing deforestation. Proc. Natl. Acad. Sci. U. S. A. 105 (42), 16089–16094.

致谢

毕业论文的完成也意味着我大学四年的生活接近尾声。很感谢刘建老师给我这个机会能够进一步接触自己感兴趣的领域-数据分析/机器学习深度学习领域，能够进一步的去尝试通过预处理清洗数据，建模型搭模型，再到预测分析数据，也算是了解了其中的流程和规范，当然也有很多缺陷，比如测试相关因子的影响大小等，由于数据量非常庞大，我还没想到优化的方法，当然我相信随着后期的深入学习，我会找到方案去处理类似的大数据集，对于深度学习的认识也可以说算是初步的了解过程，当然，后面还有大量的模型需要学习使用，相信自己会坚持下去。

本次毕业设计是由刘建老师指导完成的，刘老师可以说在毕设中给我了莫大的帮助，从啥也不知道，到帮我捋清思路，慢慢搭建起一系列的流程，给我方向去学习和思考，在提交包含各种报告的时候，老师也会尽全力帮我们修改并给出意见，这种对工作的敬业和对学生的爱护，让我能够顺利完成这篇论文，在此致意谢意。

也感谢这一路走来帮助过我的人，我的同学，辅导员和实习工作时候的同事师兄师姐！

作者：王林钊

2019 年 5 月 30 日