# 分布动态法在高校英语四六级通过率演进分析中的应用

夏 艳[1] 张丽娟[1]
XIA Yan ZHANG Lijuan

**摘 要** 针对提高高校各专业英语四六级考试通过率与高校管理部门如何着手英语教学改革的问题，提出利用分布动态法对广州华商学院各个专业四六级通过率数据进行演进分析。首先，采用自适应核密度估计方法刻画了该校各专业四六级通过率发展的动态变化；然后，借助马尔科夫链估计考察了该校四六级通过率发展的内部动态转移过程，分析出该校各专业四六级通过率的宏观发展；分析结果表明，长期趋势下该校各专业英语四六级通过率将持续走低；最后，根据所得结论为该校英语教学改革提供相关建议。

**关键词** 分布动态法；自适应核密度估计；马尔科夫链估计；四六级通过率

doi：10.3969/j.issn.1672-9528.2022.09.032

## 0 引言

大学英语等级考试一直以来都是高校学生参与度最高的考试，四六级通过率不仅反映了英语教师的教学效果，在大学英语教学水平的评估中也起着关键作用；同时，四六级证书也是求职应聘中必备的职业素质，直接影响高校毕业生求职的成败。从 1987 年英语等级考试正式实施至今，有关四六级成绩的研究众多。其中，文秋芳，王海啸 [1]（1996）运用定量研究的方法，客观准确的分析了学习者的相关因素与 CET4 成绩的关系；祯世丽 [2]（2014）利用数据挖掘算法——决策树 ID3 算法分析影响大学英语等级考试成绩的主要因素，进而总结出高校大学生通过大学英语等级考试的相关特征。纵观四六级成绩方面的已有相关研究，大多是以学生个体为研究对象。本文以学科专业为研究对象，分析不同学科专业的英语四六级通过率的发展变化情况，将为高校管理层在教学改革管理上提供有力依据。分布动态法通常用于对经济体的分布动态研究，其实质是对经济变化趋势、区域差异、收敛模式和极化状况的考察。赵黎明，焦珊珊 [3]（2017）运用分布动态法（核密度估计，马尔科夫链）考察了 2000—2015 年中国旅游业发展的分布特征和演进趋势，燕安 [4](2020) 运用分布动态法考察了我国 1978—2017 年 31 个省份各地区人均 GDP 增长分布动态与长期演进趋势。对于分布动态法中的核密度估计方法，其带宽的选取是核心，为了能够准确刻画数据的密度函数，众多学者相继提出并采用最优带宽的确定方法。虽然这些方法可以确定非参数核密度估计的整体固定最优带宽，但在实际应用中由于数据的不确定性与波动性，整体数据密度分布并不均匀，整体最优带宽不能根据局部区间的样本点密度对自身进行调整，导致其局部适应性较差。[7] 本文在核密度估计中，采用自适应宽带核密度估计，即让每个数据点都有自己的带宽，这样避免了常规固定带宽的核密度估计方法估计的不足，使得估计更加灵活与准确。

## 1 数据与估计方法

### 1.1 数据来源与处理

本文采集广州华商学院 2016—2020 年 25 个专业学生的四六级成绩数据，共计 80 451 条有效数据，并对这些数据进行数据预处理，得到该校各专业四六级通过率数据。

### 1.2 分布动态法

#### 1.2.1 自适应核密度估计

核密度估计法能够评估该校各专业通过率的整个横截面分布是如何随着时间的推进而变化的。为进行相关估计，数据中各个观察值都被替换成一个正态分布，该分布的均值与被替换数值相等，该正态分布被称为核函数，正态分布核函数的标准差被称为带宽 [10]。给定一个核 $K$ 和一个成为带宽的正数 $h$，核密度估计定义为：

$$\hat{f}_n(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right) \tag{1}$$

式中：核 $K$ 为任意的光滑函数，它满足 $K(x)\geqslant 0$ 以及：

$$\int K(x)dx=1,\int xK(x)dx=0,\sigma_K^2\equiv\int x^2K(x)dx>0 \tag{2}$$

常见的核函数有 Gaussian 核、Epanechnikov 核、boxcar 核与 Tricube 核，核被用来取局部平均，在许多估计方法中有其重要作用，一般采用较光滑的核会有较光滑的估计。但对于核密度估计，对核 $K$ 的选择不重要，但是对带宽 $h$ 的选

1. 广州华商学院 广东广州 511399

[基金项目]2021 年校级质量工程－数据科学与大数据技术（HS2021ZLGC02）

择则是非常重要的[10]。$\hat{f}_n$ 对 $h$ 的选择非常敏感，带宽的大小直接决定着核密度估计函数的平滑程度；带宽越大，估计越光滑，反之估计越粗糙。在实践中，通常采用交叉验证法或正态参照规则来选择带宽，这里的带宽为一个固定值。适应性核是核方法的一个推广，它对于每个点 $x$, 使用不同的带宽 $h(x)$，用不同的带宽 $h(x_i)$ 于每个数据点，这使得估计更加灵活，而且允许其适应于光滑性变化的区域。本文采用的自适应带宽的核密度估计方法是在固定带宽核密度函数的基础上，通过修正带宽参数位而得到的，其表达形式为：

$$k(x)=\frac{1}{M}\sum_{j=1}^{M}\frac{1}{(\omega h_j)^n}K(\frac{x-x^{(j)}}{wh_j}) \tag{3}$$

式中：$k(x)$ 是带宽为 $h_j$ 的核密度估计函数，$M$ 为样例的个数，每一个点 $j$ 都有一个带宽 $h_j$，$K(x)$ 是核函数，本文选用常用的高斯核函数；$0 \leqslant \alpha \leqslant 1$ 为灵敏因子，通常 $\alpha$ 取 0.5，$\alpha$=0 时，自适应带宽的核密度估计就变成了固定带宽的核密度估计了，$\omega$ 表示带宽的参数。

1.2.2 马尔科夫链

马尔科夫链方法的核心是一个 $m\times m$ 的转移概率矩阵，其中 $m$ 取决于状态空间中的状态个数。本文将每个专业的四六级通过率的等级状态用变量 $X_t$ 表示，$X_t$ 的可能值的集合用状态空间 $S$ 表示，其中 $X_t \in S$。状态空间 $S$ 应取非负整数集 $S$=(1, 2, 3, ⋯, $m$)，本文分析中 $S$=(1, 2, 3)。

（1）马尔科夫转移矩阵

如果 $X_t$=$i$，那么称该过程在时刻 $t$ 的状态为 $i$，并假设 $p_{ij}$ 为处在 $i$ 状态的随机变量下一时刻在 $j$ 状态的概率，马尔科夫转移矩阵形式为：

$$\prod\nolimits_t=\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1} & P_{N2} & \cdots & P_{NN} \end{bmatrix} \quad p_{ij\geq 0}\ ,\ \sum_{j=1}^{N}p_{ij}=1$$

式中：$N$=3。

为了获得在计算上的方便，对于马尔科夫转移矩阵的求解通常加入以下两个假设：第一，对于一阶马尔科夫链：当过程在时刻 $t_0$ 所处的状态为已知的条件下，过程在 $t$ 时刻（$t$>$t_0$）所处的状态仅与时刻 $t_0$ 有关，而与过程在 $t_0$ 之前的时刻无关系；第二，在整个研究期间 $T$，马尔科夫转移矩阵不随时间而变化。

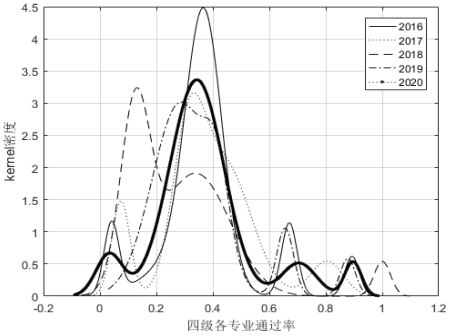
（2）平稳分布

若 $H$=($h_1$, $h_2$, ⋯, $h_N$) 为一状态概率向量，$P$ 为状态转移概率矩阵，若 $HP$=$H$，称 $H$ 为该马尔科夫链的一个平稳分布。本文假定研究的整个过程是时间同质的，计算各专业四六级通过率的转移矩阵，可以用来研究各专业四六级通过率分布的演变。在经历了 $m$ 个时期（从 $t$ 到 $t$+$m$）之后的各专业四六级通过率分布可以通过公式 $h_{t+m}=h_t\prod^m$ 来得到。进一步，当 $t$ 趋于无限大时，将得到稳态分布向量 $H$，这样各专业四六级通过率分布随时间的演化过程就能够得到分析。

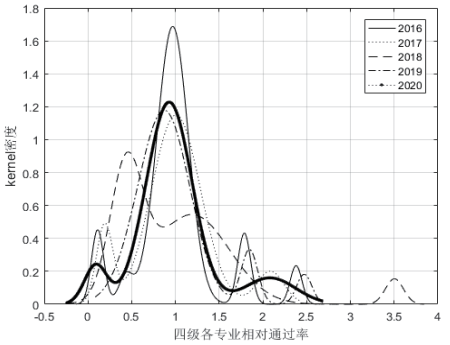
## 2 实证分析

### 2.1 自适应带宽核密度估计

本文采用高斯正态分布作为核密度函数进行自适应带宽核密度估计，得到如图 1 所示和图 2 所示的各专业四六级通过率自适应核密度估计图。对于该校各专业四级通过率的总体趋势，由图 1（a）可知，该校各专业四级总体通过率在 2016、2017、2019 和 2020 年期间，总体上并未发生明显偏移；但在 2018 年期间，该校各专业四级均分发生了明显的“左移”，这意味着在 2018 年期间该校各专业整体四级应试水平出现了共同下降。

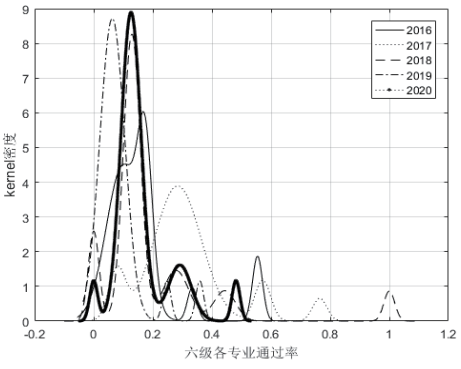



（a）各专业四级通过率

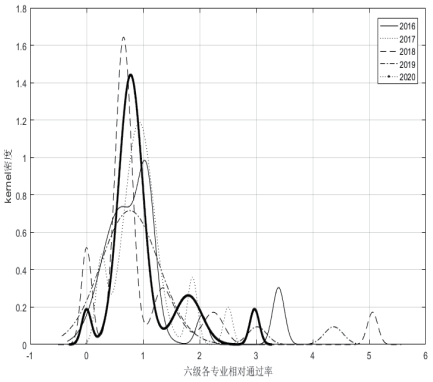



（b）各专业相对四级通过率

图 1 各专业四级通过率自适应核密度估计




（a）各专业六级通过率




（b）各专业相对六级通过率

图 2 各专业六级通过率自适应核密度估计

对于该校各专业六级成绩总体趋势，由图 2（a）可知，该校各专业六级整体应试总体水平在 2017 年有明显的提升趋势，而在 2019 年期间该校各专业六级整体应试总体水平有明显的下降趋势，在其他年份基本保持稳定。

为了得到各专业相对四六级通过率，本文将各专业四六级通过率与全校四六级通过率均值数值相除，这样处理可以更有效的分析“双峰”趋势。通过自适应核密度估计得到各专业相对四六级通过率分布情况图 1（b）与图 2（b）。由图分析可知，该校在 2018 年期间各专业间四级通过率差距明显，大部分专业通过率低于全校平均水平；而该校 2016—2020 年各专业英语六级通过率内部，呈现出动态高斯分布现象。

2.2 马尔科夫链估计

本文采集该校 2016—2020 年 25 个专业四六级 10 次考试的通过率数据，运用马尔科夫链分析各专业四六级通过率的分布演进趋势，将 2016—2020 各专业通过率指标数据按照 0 ～ 1 范围分为三个阶段，即低水平阶段 [0-0.3）、中等水平阶段 [0.3-0.6）、高水平阶段 [0.6-1]；并将收集数据根据阶段范围转化为 1-3 这样的量化值。

该校各专业 2016—2020 年四六级通过率数据被划分为三个通过率状态，计算相关转移概率矩阵，获得各专业四六级通过率分布随时间演进的信息；通过 MATLAB 软件计算稳态分布，将其与初始分布进行比较，通过比对观察两者的不同点。本文采用马尔科夫链估计方法得到表 1 与表 2 该校各专业四六级通过率的马尔科夫转移矩阵。

表 1 各专业四级通过率的马尔科夫转移矩阵

| $t/t$+1 | $L_1$ | $L_2$ | $L_3$ | 初始分布 |
|---|---|---|---|---|
| $L_1$ | 0.729 7 | 0.243 2 | 0.027 0 | 0.37 （83） |
| $L_2$ | 0.259 3 | 0.666 7 | 0.074 1 | 0.54（122） |
| $L_3$ | 0.000 0 | 0.555 6 | 0.444 4 | 0.09 （20） |
| 稳态分布 | 0.448 4 | 0.467 4 | 0.084 1 | |

表 2 各专业六级通过率的马尔科夫转移矩阵

| $t/t$+1 | $L_1$ | $L_2$ | $L_3$ | 初始分布 |
|---|---|---|---|---|
| $L_1$ | 0.875 0 | 0.125 0 | 0 | 0.80（180） |
| $L_2$ | 0.611 1 | 0.277 8 | 0.111 1 | 0.18（40） |
| $L_3$ | 0.5 | 0.5 | 0 | 0.02 （5） |
| 稳态分布 | 0.827 6 | 0.155 2 | 0.017 2 | |

从马尔科夫链估计来看，第一，该校各专业四六级通过率主要分布在低水平状态与中等水平状态。第二，该校各专业英语四级通过率状态在转移上具有一定的稳定性，而各专业英语六级通过率状态在转移上具有一定的“向下转移”趋势。第三，长期稳态分布下，该校各专业英语四六级低水平状态专业通过率均有略微增长，同时中等水平状态专业通过率均略有降低。

## 3 结论

本文采集广州华商学院 25 个专业学生 2016—2020 年英语四六级成绩数据，分析该校各专业四六级通过率的分布情况与演进趋势，分析结果表明，长期趋势下该校各专业英语四六级通过率将持续走低。综合以上分析结论，可对该校提出如下建议：第一，该校应在现有的大学英语教务相关改革工作中加大力度；第二，对于 2016—2020 年大学英语教务改革的经验，该校教务管理部门应重点关注 2017 年的优势与 2018 年的不足，总结出适合该校学生的大学英语等级考试相关教务管理措施，以提高学生英语四六级等级考试的通过率。

**参考文献：**

[1] 文秋芳，王海啸 . 学习者因素与大学英语四级考试成绩的关系 [J]. 外语教学与研究 ,1996(4):33-39+80.

[2] 禤世丽 . 决策树在英语等级考试成绩分析中的应用研究 [J]. 计算机与数字工程 ,2014,42(5):843-845+871.

[3] 赵黎明，焦珊珊，姚治国 . 中国旅游经济发展的分布动态演进 [J]. 干旱区资源与环境 ,2018,32(1):181-188.

[4] 燕安 . 中国区域经济差异的时空演进 [J]. 统计与决策，2020, 36(20):86-90.

[5] 何江，张馨之 . 中国省区收入分布演进的空间 - 时间分析 [J]. 南方经济 ,2007(1):64-77.

[6] 文秋芳，王海啸 . 学习者因素与大学英语四级考试成绩的关系 [J]. 外语教学与研究 ,1996(4):33-39+80.

[7] 朱艳丽 . 基于自适应核密度估计的城管案件时空预测方法 [D]. 北京：北京建筑大学 ,2021.

[8] 黄明月 . 基于自适应核密度估计的交通事故黑点识别及预警模型研究 [D]. 苏州 : 江苏大学 ,2021.

# DevOps 模式下的 CMDB 研究

张亚辉[1] 马海燕[2] 陈 森[1] 李 林[1]
ZHANG Yahui MA Haiyan CHEN Sen LI Lin

**摘 要** 针对配置管理数据库（configuration management database，CMDB）建设过程中经常存在的目标不清晰、盲目建设等问题，提出了一套贴近生产需求的 CMDB 设计方案。首先对 CMDB 的特点及建设误区进行深入分析，然后结合 DevOps 开发运维一体化模式的特点，对该模式下 CMDB 的设计和建设进行研究，充分论证整体架构、模型建设、功能开发方面的宏观要求，最终提出一套 DevOps 模式下的 CMDB 设计方案。该论证方案为开发运维一体化 DevOps 的发展提供了数据分析和应用的基础，以高质量的 CMDB 支撑起面向业务的 DevOps 运营维护体系。

**关键词** CMDB；DevOps；配置管理

doi：10.3969/j.issn.1672-9528.2022.09.033

## 0 引言

在 IT 运维中，配置管理数据库 CMDB 越来越受到运维人员的重视，绝大部分具备规模的 IT 系统都建设有 CMDB 并作为其他运维平台的数据支撑，但实际运维中 CMDB 的成功案例并不多见。大家对 CMDB 的认识、建设目标乃至实际建设成果大多存在一定的误区，或未能充分做好前期规划和设计，盲目进行 CMDB 建设。生产中 CMDB 建设虎头蛇尾，甚至最终弃之不用或重新建设的情况并不少见。

本文作者团队将总结 CMDB 建设维护过程中的经验教训，分析对 CMDB 的认识误区及其特点，结合 DevOps 实践需求，对 CMDB 的架构、对外接口、数据采集、资源模型、功能及流程等方面进行深入研究探讨。

1. 中国移动通信集团设计院有限公司山东分公司 山东济南 250101
2 山东电子职业技术学院 山东济南 250200

## 1 CMDB 的特点和建设误区

### 1.1 CMDB 不仅仅是个 DB

CMDB[1-3] 全称 configuration management database，但不是一个狭义的数据库。一个能发挥最大运维价值的 CMDB 其实是一个在动态变化的配置项（configuration item，CI）数据基础上，通过关联关系，构建分类分层的立体数据架构，再附以相关的流程管理和配置管理而形成的立体数据管理体系。它对外提供的不仅仅是单一数据，而是海量数据的关系及层次模型，以及由此产生的业务信息或数据价值。

### 1.2 CMDB 不等于资产管理

在实际生产过程中配置管理有时也被称为资产管理，资产相关的管理需求常常被作为了 CMDB 的首要任务或急迫需求。资产管理虽是 CMDB 的重要需求，但在建设目的、功能价值、数据颗粒及其关联关系方面均无法达到配置管理的全面性。

配置管理的目的不仅局限在资产、成本、合同方面的

[9] 胡森林，焦世泰，张晓奇．中国城市旅游发展的时空演化及影响因素：基于动态空间马尔科夫链模型的分析 [J]. 自然资源学报，2021,36(4):854-865.
[10] L. 沃塞曼．现代非参数统计 [M]. 吴喜之，译．北京：科学出版社，2008.
[11] 张丽琼，何婷婷 .1997—2018 年中国农业碳排放的时空演进与脱钩效应：基于空间和分布动态法的实证研究 [J]. 云南农业大学学报 ( 社会科学 ), 2022,16(1):78-90.

【作者简介】
夏艳（1996—），女，湖北荆州人，硕士，助教，研究方向：数据挖掘与大数据分析。
张丽娟（1994—），女，湖南株洲人，硕士，助教，研究方向：数据挖掘与大数据分析。

（收稿日期：2022-02-24 修回日期：2022-05-20）