文章编号：1674 - 5566(2014)01 - 0123 - 08

GLM 模型和回归树模型在 CPUE 标准化中的比较分析

官文江[1,2]，陈新军[1,2]，高 峰[1,2]，雷 林[1,2]

(1. 上海海洋大学 海洋科学学院，上海 201306；2. 上海海洋大学 大洋渔业资源可持续开发省部共建教育部重点实验室，上海 201306)

摘 要： 在渔业资源评估中，CPUE(catch per unit effort)标准化是基础性工作。一般线性模型(generalized linear model，GLM)已成为 CPUE 标准化的基本方法，但 GLM 模型在误差结构、自变量的选择、缺失数据、复杂交互效应及异常值处理等方面仍然缺乏灵活性。本文基于模拟数据及我国东、黄海鲐鱼(*Scomber japonicus*)灯光围网渔业数据，比较和分析了基于 GLM 模型与回归树模型在 CPUE 标准化中的效果。研究表明：当渔业数据不存在非线性关系与异常值时，GLM 模型与回归树模型均能较好地对 CPUE 进行标准化，但由于回归树模型具有阶跃函数特征，因而 GLM 模型更具优势；在非线性关系及异常值存在的条件下，回归树模型对 CPUE 的标准化具有相对较小的估计误差，模型更简约、有效。由于回归树模型能可视化显示自变量与应变量间的复杂关系，因此，更有利于探索和分析渔业数据。

研究亮点： GLM 模型为 CPUE 标准化的基本方法，但该方法在误差结构与自变量的选择及处理缺失数据、交互效应、异常值等方面仍然缺乏灵活性。本文讨论了回归树模型在 CPUE 标准化方面的特点、优势及可能存在的问题，有利于丰富渔业数据处理的方法。

关键词： CPUE 标准化；回归树模型；GLM 模型

中图分类号： S 932.2

文献标志码： A

近年来的研究表明，很多渔业资源出现了衰退趋势，如东、黄海鲐鱼资源[1-2]。为使渔业资源可持续利用，必须对渔业资源进行科学管理，而渔业资源评估是科学管理的基础。由于我国科学调查的渔业资源指数等数据相对缺乏，对渔业资源的评估仍依赖于商业性渔业数据[3]。CPUE(catch per unit effort)数据常被假设与资源量成正比关系[4]，并作为资源丰度指数应用于渔业资源评估中[3-5]。但大量研究表明，CPUE 与资源量关系非常复杂，其受资源丰度、资源栖息地面积[6-7]、环境效应[8]，渔民行为或钓具改进[9]、管理规定[10]等影响。因此，在渔业资源评估中，需要对 CPUE 进行标准化，而 CPUE 的标准化也是渔业资源评估的基础性工作[11-12]，其往往直接影响渔业资源评估的结果。

一般线性模型(generalized linear model，GLM)是当前国际上 CPUE 标准化的基本方法[13-16]，如官文江和陈新军[17]利用 GLM 模型估算了鲐鱼灯光围网渔业公司的捕捞效率，李纲等[18]利用 GLM 模型对鲐鱼 CPUE 数据的标准化进行了探讨。但 GLM 模型在误差结构假设、变量选择、缺失数据、复杂交互效应及异常值(outlier)处理等方面仍然缺乏灵活性[17]，而回归树模型则在这些方面具有一定优势[19]。WATTERS 和 DERISO [19]第一次使用了回归树模型对 CPUE 进行标准化，但此后，相关研究较少见于文献。因此，本文尝试使用回归树模型对中国东、黄海鲐鱼灯光围网渔业的模拟 CPUE 数据和实际 CPUE 数据进行标准化，并与 GLM 模型的结果进行比较，以进一步探究回归树模型在 CPUE 标准化中的应用潜力。

收稿日期： 2013-08-08 **修回日期：** 2013-10-15

基金项目： 上海市教委科研创新项目(14ZZ147)；国家发改委产业化专项(2159999)；上海市科技创新行动计划(12231203900)

作者简介： 官文江(1974—)，男，副教授，研究方向为渔业资源评估。E-mail：wjguan@ shou. edu. cn

1 材料与方法

1.1 数据

1.1.1 东、黄海鲐鱼灯光围网渔业数据

1998 至 2010 年我国东、黄海鲐鱼灯光围网渔业数据来自上海海洋大学鱿钓技术组。该数据包含生产日期(年、月)、渔业公司名称、作业位置(空间分辨率为 0.5° × 0.5°)、捕捞网次、捕捞产量等字段[18]。根据鲐鱼灯光围网渔业生产的空间分布及数据特点,将作业海区分为 3 个子区域:32°N 以北称为 1 区,长江口及舟山近海为 2 区,台湾东北部、舟山外海为 3 区[17]。本文仅对 1 区与 3 区的渔业数据进行分析。渔业数据的名义 CPUE(Nominal CPUE)由式(1)计算。

$$O_{y,m,c}=\frac{H_{y,m,c}}{N_{y,m,c}} \tag{1}$$

式中:y 为年;m 为月;c 为渔业公司编号;H 为捕捞产量(箱),一箱为 20 kg;N 为捕捞网次;O 为名义 CPUE。

1.1.2 模拟数据

1998 至 2010 年鲐鱼资源量变化的年效应(Y)、月效应(M)及渔业公司捕捞效应(C)见表 1,则模拟名义 CPUE 可由下式计算:

$$O_{y,m,c}=\beta\times I_0\times Y_y^E\times M_m\times C_c\times e^{\varepsilon} \tag{2}$$

式中:β 为随机数;I_0 为平均 CPUE,设为常数;E 用于描述名义 CPUE 与真实 CPUE 之间关系的量[11];ε 为观察误差项。当 E 为 1.0 时,则名义 CPUE 与真实 CPUE 之间存在线性关系。当 E 不为 1.0 时,则名义 CPUE 与真实 CPUE 之间的关系具有非线性特点。从文献看,E 的取值范围可从 0.5 至 1.5[4,11,20],本文假设 E 服从[0.7, 1.3]的均匀分布。ε 服从 $N(0,\sigma^2)$ 的正态分布,σ 值相对较大(0.4 ~ 0.5)[6],本文 σ 假设为0.25。由于 1998 至 2010 年的东、黄海鲐鱼灯光围网渔业的 CPUE 数据约有 4% 为 0,并且约有 2% 的 CPUE 大于其平均值 8 倍以上或小于其平均值 20 倍以上,这些数据点可能为异常值(outlier)。为在模拟数据中增加异常值,β 分别有 1% 的概率在区间[7.4, 12.2]与区间[0.02, 0.05]中随机取值,94% 的概率为 1.0,4% 的概率取 0。为进行对比分析,本文同时模拟了 E 与 β 分别等于 1.0的情况。综上所述,本文模拟以下两种情况,场景 1:E 与 β 分别等于 1.0。场景 2:E 服从[0.7, 1.3]的均匀分布,β 分别有 1% 的概率在区间[7.4, 12.2]与区间[0.02, 0.05]中随机取值,94% 的概率为 1.0,4% 的概率为 0。

因此,根据实际渔业数据的年、月、渔业公司记录、式(2)及 E 与 β 的取值,可生成与渔业数据相同长度的模拟时间系列。对上述两种情况,各重复 2 000 次,分别产生 2 000 个数据集用于分析。

表 1 年、月、公司效应数据

Tab. 1 The data of year effect, month effect and company effect used in simulation

年效应		月效应		公司效应	
年份	值	月	值	公司编号	值
1998	1.00	1	1.00	1	1.00
1999	0.86	2	1.05	2	0.66
2000	0.62	3	1.09	3	0.34
2001	0.78	4	1.14	4	0.32
2002	0.65	5	1.18	5	0.50
2003	0.64	6	1.23	6	0.49
2004	0.88	7	1.27	7	0.60
2005	0.72	8	1.32		
2006	0.63	9	1.36		
2007	0.79	10	1.41		
2008	1.10	11	1.45		
2009	1.11	12	1.50		
2010	0.92				

注:年效应数据参考由上龍嗣等[21],并以 1998 年为参考年;渔业公司效应根据官文江和陈新军[17];月效应为假设数据。

1.2　方法

1.2.1　回归树模型

回归树的构建共包括两个过程:一是树生长,即不断选择合适自变量将数据分裂为两个子集,以增加子集内部应变量的均匀性及子集间应变量的差异性[19,22];二是剪枝,即确定树的复杂性,以减少树的分裂数或叶子节点数,防止树过拟合。本文树分裂数由 10 倍交叉验证(10-Fold Cross-Validation)确定,即取具有最小相对误差的树作为分析树[22]。

1.2.2　GLM 模型

由模拟数据及李纲等[18]结果可知,CPUE 数据的误差结构可假设为对数正态分布,同时由于存在 CPUE 数据为 0 的点,因此,采用以下方法构造 GLM 模型:

$$g(O_{y,m,c}+\delta)=Y_y+M_m+C_c+\varepsilon \tag{3}$$

式中:g 为连接函数,本文为 log,δ 取名义 CPUE 平均值的 10%[17],ε 为正态分布误差项。

1.2.3　基于模拟数据的方法比较

对 GLM 模型,渔业公司效应与年效应的估计直接来自 GLM 模型的回归系数。对回归树模型,先估算各年、月与渔业公司组合下的预测 CPUE,然后分别选定参考渔业公司与参考年份,并假设:在相同年与月份条件下各渔业公司与参考渔业公司间的 CPUE 之比为公司效应;在相同渔业公司与月份条件下,各年与参考年的 CPUE 之比为年效应[19]。回归树模型估计的渔业公司效应与年效应为其平均值。

为比较回归树与 GLM 模型估算效果,本文比较了参数的平均值(A)与均方根误差(R)及总均方根误差(S),A、R、S 分别由式(4)、式(5)与式(6)计算。

$$A_k=\frac{\sum_{i=1}^{2000}\hat{\theta}_{i,k}}{2000} \tag{4}$$

$$R_k=\sqrt{\frac{\sum_{i=1}^{2000}(\hat{\theta}_{i,k}-\theta)^2}{2000}} \tag{5}$$

$$S=\sqrt{\frac{\sum_{k=1}^{n}\sum_{i=1}^{2000}(\hat{\theta}_{i,k}-\theta)^2}{2000\times n}} \tag{6}$$

式中:$\hat{\theta}$ 为估计值,θ 为真实值(表 1),i 表示第 i 个数据集,k 代表年或公司。当 k 为公司时,n 为 7;当 k 为年时,n 为 13。

1.2.4　基于东、黄海鲐鱼灯光围网渔业数据的实例研究

利用 GLM 模型与回归树模型分别对东、黄海鲐鱼灯光围网渔业名义 CPUE 数据进行标准化。GLM 模型估算年效应的方法同模拟数据。考虑到部分渔业公司的捕捞效应存在年际变化,回归树模型计算年效应的方法分为以下三步:第一步将各年、月的 CPUE 统一到参考公司(参考公司的捕捞效应年际变化应该尽量小)尺度下,以去除公司效应的影响;第二步将每年各月份的 CPUE 除以参考年相同月份的 CPUE,以去除月效应的影响;第三步将第二步获得的比率按年进行平均,该平均值即为标准化的 CPUE[19]。

2　结果

2.1　模拟数据的 CPUE 标准化

GLM 模型总变量数为 32(13 年、12 个月、7 个公司),回归树模型中,场景 1 树分裂平均次数为 10,场景 2 为 13。

利用 GLM 模型与回归树模型,分别估计上述 2 000 个模拟数据集的年效应与渔业公司效应,并计算了各效应的平均值,均方根误差及总均方根误差,其结果见表 2、3。由表 1、2、3 可知,在没有非线性及异常值的影响下(模拟数据的场景 1),GLM 模型与回归树模型均能较好估计年效应与公司效应,但 GLM 模型估计的平均值更接近真实值,均方根误差与总均方根误差更小,效果好于回归树模型(表 1 ~ 3)。

但当数据存在非线性及异常值的条件下(模拟数据的场景 2),GLM 模型与回归树模型估计的各参数的均方根误差均增大,回归树模型估计的公司效应的平均值与 GLM 模型估计的结果相比,回归树模型估计值与真值吻合较好(表 1,2),其总均方根误差相对较小(表 2)。

GLM 模型估计的年效应平均值与回归树模型估计的结果相比,GLM 模型估计值与真值吻合较好(表 1,3),但 GLM 模型的总均方根误差较大(表 3)。这表明 GLM 模型对非线性关系与异常值非常敏感,当数据存在非线性关系与异常值时,结果会出现较大波动,从而产生较大的均方根误差。回归树模型是非参数化方法,具有阶跃函数(Step Function)特征,因此其估计值可能有偏,但当数据存在非线性关系与异常值时,其结

果表现稳健(Robust),估计参数的总均方根误差 较小。

表 2 GLM 模型与回归树模型公司效应估计比较

Tab. 2 The comparison of company effect estimated by GLM model and regression tree model respectively

公司编号	GLM 模型				回归树模型			
	场景 1		场景 2		场景 1		场景 2	
	A_k	R_k	A_k	R_k	A_k	R_k	A_k	R_k
1	1	0	1	0	1	0	1	0
2	0.67	0.01	0.61	0.13	0.65	0.03	0.65	0.12
3	0.38	0.04	0.31	0.07	0.35	0.01	0.35	0.07
4	0.35	0.03	0.29	0.06	0.35	0.03	0.35	0.07
5	0.52	0.02	0.45	0.11	0.54	0.05	0.51	0.13
6	0.52	0.04	0.45	0.10	0.54	0.05	0.51	0.11
7	0.57	0.03	0.51	0.26	0.55	0.06	0.57	0.13
		0.03		0.13		0.04		0.10

注:A_k 为平均值,R_k 为均方根误差,表的最后一行为总均方根误差,即 S;计算方法见式(4)、式(5)与式(6)。

表 3 GLM 模型与回归树模型年效应估计比较

Tab. 3 The comparison of year effect estimated by GLM model and regression tree model respectively

年份	GLM 模型				回归树模型			
	场景 1		场景 2		场景 1		场景 2	
	A_k	R_k	A_k	R_k	A_k	R_k	A_k	R_k
1998	1	0	1	0	1	0	1	0
1999	0.87	0.03	0.91	0.27	0.83	0.08	0.96	0.20
2000	0.63	0.02	0.61	0.17	0.69	0.08	0.84	0.25
2001	0.78	0.03	0.79	0.23	0.81	0.07	0.92	0.21
2002	0.65	0.025	0.64	0.18	0.69	0.06	0.85	0.25
2003	0.66	0.03	0.65	0.21	0.71	0.09	0.87	0.28
2004	0.86	0.04	0.90	0.33	0.96	0.11	1.01	0.25
2005	0.71	0.03	0.72	0.27	0.72	0.05	0.89	0.23
2006	0.52	0.11	0.51	0.22	0.68	0.08	0.84	0.25
2007	0.78	0.04	0.82	0.40	0.87	0.12	0.98	0.30
2008	1.06	0.06	1.16	0.51	1.14	0.11	1.27	0.41
2009	1.04	0.09	1.16	0.75	1.16	0.11	1.27	0.47
2010	0.86	0.07	0.95	0.64	0.94	0.07	1.03	0.31
		0.05		0.38		0.09		0.28

注:A_k 为平均值,R_k 为均方根误差,表的最后一行为总均方根误差,即 S;计算方法见式(4)、式(5)与式(6)。

2.2 我国东、黄海鲐鱼灯光围网渔业 CPUE 标准化

1 区 10 倍交叉验证最小相对误差对应的树分裂次数为 6(图 1A),共 7 个叶子节点(图 2);3 区 10 倍交叉验证最小相对误差对应的树分裂次数为 13(图 1B),共 14 个叶子节点(图 3)。1 区回归树(图 2)能解释 12% 的总离差(deviation),3 区回归树(图 3)能解释总离差的 27%。1 区 GLM 模型能解释的总离差为 11%,3 区 GLM 模型能解释的总离差为 18%。回归树模型中树分裂次数 1 区为 6(可作为回归总变量数指标[19]),3 区为 13。GLM 模型中,参与回归的总变量数,1 区为 27(13 年、8 个月、7 个公司),3 区为 32(13 年、12 个月、7 个公司)。因此回归树能解释更多离差,而且模型更简约,即自由度损失更小。

图 2、3 可视化展示了回归树的构建过程及月、年、渔业公司在构建回归树中的作用。由图 2 可知,渔业公司 1 捕捞效率存在明显的年际变化,如 1998、1999、2002 年具有较大的 CPUE(3 489箱/网次),而 2004 年捕捞效率较差(953.3 箱/网次)。公司 3、4 捕捞效率年际变化不明显,但公司 4 在 2006 年后退出了该渔业,因此可选渔业公司 3 作为参考渔业公司,以减少参考渔业公司捕捞效率的年变化对 CPUE 标准化的影响。图

3 展示了 CPUE 与年、月、渔业公司之间的复杂关系,但依然可以看出,渔业公司 1 的捕捞效率存在年际变化,如 8 月(3 区主要生产月份),1999、2002 年 CPUE 为 4 006 箱/网次,而 1998, 2009 年 CPUE 为 1 793 箱/网次。同样,渔业公司 3 捕捞效率年际变化不明显。

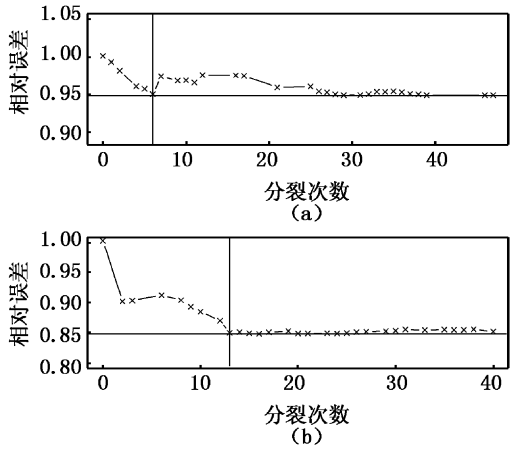

图 1　相对误差随树分裂次数的变化

Fig. 1　The relationship between the relative error and the number of split

(a)为 1 区,(b)为 3 区。

表 4　回归树与 GLM 模型估计的年效应

Tab. 4　The year effect estimated by GLM model and regression tree model

年份	1 区		3 区	
	回归树模型	GLM 模型	回归树模型	GLM 模型
1998	1.00	1.00	1.00	1.00
1999	1.00	1.13	1.15	1.03
2000	0.83	0.76	0.64	0.56
2001	0.83	0.79	0.64	0.47
2002	0.83	0.99	1.26	1.19
2003	0.65	0.47	0.70	0.59
2004	0.95	1.02	0.70	0.66
2005	0.83	0.75	0.70	0.61
2006	0.65	0.38	0.64	0.62
2007	0.65	0.59	0.64	0.42
2008	0.65	0.59	0.70	0.70
2009	0.65	0.60	1.17	0.91
2010	0.65	0.59	0.70	0.57

GLM 模型与回归树模型估计的年效应(表 4)即标准化的 CPUE,在数值上存在差异,但两者均显著相关($R > 0.92$, $P\text{-}VALUE < 0.001$),这表明,两模型估计的年效应变化趋势一致。

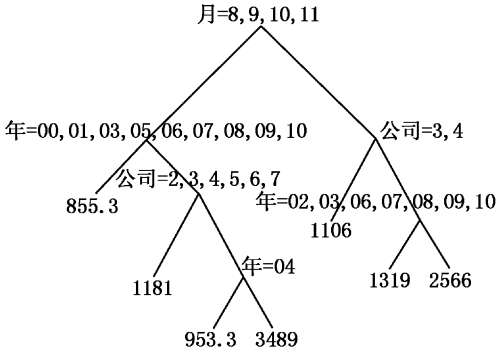

图 2　1 区回归树模型

Fig. 2　Regression tree for area 1

图中 CPUE 单位为箱/网次。

3　讨论

回归树模型是一种非参数化方法,其不需要预先假设应变量和自变量之间的关系,而是通过递归二分方法(binary split),自动选择自变量将应变量定义的空间逐步划分为尽可能同质的类别[23-24],并通过可视化的方式表达应变量与自变量间的复杂关系,这使模型构建非常方便,有利于分析数据,了解数据的特性。同时,当变量间存在非线性关系,具有更高阶的交互效应或数据存在缺失(自变量数据缺失)、不平衡及异常数据(outlier)等问题时,回归树模型具有优势[22,25-26],如 VAYSSIERES 等[27]比较了 GLM、GAM 与 CART(Classification and Regression Tree)方法预测物种分布的效果,发现 CART 具有更好的预测能力;刘洋等[28]和 ZHENG 等[29]也认为回归树模型比 GLM 模型具有更好的预测能力。

在对 CPUE 进行标准化时,GLM 模型或 GAM 模型(本文各变量均为因子变量,GLM 与 GAM 结果一致,因此本文没有与 GAM 模型进行比较。)均需假设误差结构,并需对为 0 的 CPUE 数据点进行处理。误差结构的选择与对为 0 的 CPUE 数据点的处理方式将影响 CPUE 标准化的结果[17]。此外,多种因素均可能对 CPUE 与资源量的关系产生影响如叶绿素浓度、海面高度、海表水温、盐度以及经度、纬度等[18,30-31]。采用 GLM 或 GAM 模型时,变量的选择及交互关系的确定常存在问题[19],而采用回归树模型,则可避免上述问题。

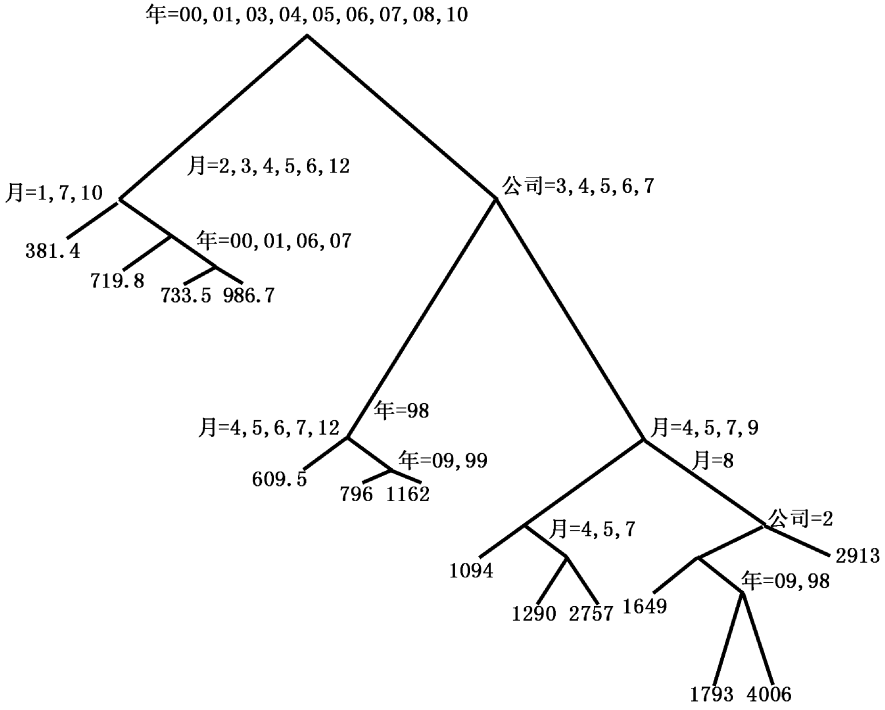

图 3 3 区回归树模型

Fig. 3 Regression tree for area 3

图中 CPUE 单位为箱/网次。

利用模拟数据对两模型的测试表明,在没有非线性关系与异常值存在的情况下,回归树模型与 GLM 模型均能对年效应、公司效应给出合理的估计,但由于回归树模型估计值具有阶跃函数特征,GLM 模型能得到比回归树模型更好的估计效果;若存在非线性关系与异常值时, GLM 模型的总均方根误差比回归树模型大,表明 GLM 模型对非线性关系与异常值敏感。因此,与 GLM 模型相比,回归树模型具有较好的稳健性。

在实际渔业生产中,由于渔业数据并非由随机采样获得,CPUE 与资源量间可能存在复杂的非线性关系[4,11];渔业数据常存在捕捞位置、捕捞时间错误及漏报和错报的情况,渔业数据的质量有时难以得到保证[4,11],数据中常存在异常值。对具体渔业数据,回归树模型的结果会更稳健,更有可能获得相对较小的估计误差。

图 2、3 可视化展示了回归树的构建过程。透过对树构建过程的分析,可发现渔业公司 1 的捕捞效率存在年际变化。官文江和陈新军[17]对 1998 至 2003 年鲐鱼灯光围网渔业数据分析的结果表明渔业公司 1 的捕捞效率最大,但李纲等[18]对 1998 至 2006 年鲐鱼灯光围网渔业数据分析的结果却表明公司 2 的捕捞效率最高,因此,渔业公司 1 的捕捞效率存在年际变化可能是他们结果存在差异的原因。而在 GLM 模型中,很难分析渔业公司捕捞效率的年际变化[17]。同时,图 2、3 展示了 CPUE 与年、月、渔业公司之间的复杂关系,这种关系一般难于用线性关系简单表达。

尽管 GLM 模型与回归树模型估计的年效应存在显著相关性,两者所反映的资源变化趋势一致,但回归树模型用了更少的变量(自由度),获得对离差更多的解释,表明在该渔业数据的处理中,回归树模型比 GLM 模型更简约,有效。

为了避免回归树模型过拟合,通常采用交叉验证(cross validation)方法确定树分裂次数:如 10 倍交叉验证方法,即将数据随机均匀地分为 10 份,依次用其中 9 份数据点拟合模型,另 1 份数据用于测试模型,并估计其预测相对误差。一般随树分裂次数的增加,相对误差随之减少,但当树分裂次数进一步增加,相对误差则可能随之增大(图 1),因此具有最小相对误差的树预测能力最好。一般情况下,在最小相对误差加减一个标准误的区间内,预测能力仅有较少减弱,因此在该区间具有最少分裂次数的树也常被选择作为最

终模型,这就是 1-SE 规则[32]。但由于模型拟合与测试的数据是随机确定的,因此,树的构建具有一定的不确定性。此外,回归树模型存在不能建立应变量与自变量间的定量关系;也不能保证每次分裂均为全局最优;树的结构有可能随采样数据的变化而有很大不同等缺点。由于渔业数据的收集缺少科学设计过程,数据质量常受到影响[19],但回归树模型对此具有较好的稳健性,因此,回归树模型是有效探索和分析该类数据的工具。

参考文献:

[1] 王凯, 严利平, 程家骅, 等. 东海鲳鱼资源合理利用的研究[J]. 海洋渔业,2007,29(4):337 -343.

[2] 张洪亮,周永东,陈斌. 浙江群众传统灯光围网渔业利用资源状况分析[J].海洋渔业, 2007, 29(2): 174 -178.

[3] 李纲,陈新军,官文江. 基于贝叶斯方法的东黄海鲳资源评估及管理策略风险分析[J].水产学报, 2010, 34(5): 740 -750.

[4] HILBORN R, WALTERS C J. Quantitative Fisheries Stock Assessment: choice, dynamics and uncertainty [M]. New York: Chapman and Hall, 1992.

[5] 王继隆, 李继龙,杨文波, 等. 利用气候因子对 Fox 模型计算东海总经济鱼类 CPUE 的优化[J].中国水产科学, 2011, 18(1): 136 -144.

[6] HARLEY S J, MYERS R A, DUNN A. Is catch-per-unit-effort proportional to abundance? [J]. Canadian Journal of Fisheries and Aquatic Sciences, 2001, 58(9): 1760 -1772.

[7] SALTHAUG A, AANES S. Catchability and the spatial distribution of fishing vessels [J]. Canadian Journal of Fisheries and Aquatic Sciences, 2003, 60(3): 259 -268.

[8] ZIEGLER P E, FRUSHER S D, JOHNSON C R. Space-time variation in catchability of southern rock lobster Janus edwardsii in Tasmania explained by environmental, physiological and density-dependent processes [J]. Fisheries Research, 2003, 61(1/3): 107 -123.

[9] 官文江,陈新军. 利用元胞自动机探讨商业性 CPUE 与资源量之间的关系[J]. 中国海洋大学学报:自然科学版, 2008, 38(4):561 -566.

[10] OLIVEIRA M M, GASPAR M B, PAIXAO J P, et al. Productivity change of the artisanal fishing fleet in Portugal: A Malmquist index analysis [J]. Fisheries Research, 2009, 95(2/3): 189 -197.

[11] QUINN T J, DERISO B R. Quantitative Fish Dynamics [M]. New York: Oxford University Press, 1999.

[12] 官文江,高峰, 雷林,等. 渔业资源评估中的回顾性问题[J].上海海洋大学学报, 2012, 21(5):841 -847.

[13] GAVARIS S. Use of a multiplicative model to estimate catch rate and effort from commercial data [J]. Canadian Journal of Fisheries and Sciences, 1980, 37(12):2272 -2275.

[14] PUNT A E, WALKER T I, TAYLOR B L, et al. Standardization of catch and effort data in a spatially-structure shark fishery [J]. Fisheries Research, 2004, 70(2/3):141 -159.

[15] CAMPBELL R A. CPUE standardisation and the construction of indices of stock abundance in a spatially varying fishery using general linear models [J]. Fisheries Research, 2004, 70(2/3):209 -227.

[16] VENABLSE W N, DICHMONT C M. GLMs, GAMs and GLMMs: an overview o f theory for applications in fisheries research [J]. Fisheries Research, 2004, 70(2/3): 319 -337.

[17] 官文江, 陈新军. 应用一般线性模型估算鲳鲹鱼大型灯光围网渔业的捕捞效率[J]. 水产学报,2009,33(2):220 -228.

[18] 李纲,陈新军,田思泉. 我国东、黄海鲳鱼灯光围网渔业 CPUE 标准化研究[J]. 水产学报,2009,33(6):1050 -1059.

[19] WATTERS G, DERISO R. Catch per unit of effort of bigeye tuna: a new analysis with regression trees and simulated annealing [J]. Bulletin of the Inter-American tropical Tuna Commission, 2000, 21(8): 531 -571.

[20] WILSON J R, PRINCE J D, LENIHAN H S. A management strategy for sedentary nearshore species that uses marine protected areas as a reference [J]. Marine and Coastal Fisheries: Dynamics, Management, and Ecosystem Science, 2010, 2(1):14 -27.

[21] 由上龍嗣,依田真里,大下誠二, 等. 平成 23 年マサバ対馬暖流系群の資源評価[EB/OL]. http://abchan.job.affrc.go.jp/digests23/index.html [2012 -4 -16].

[22] DE'ATH G. Multivariate regression trees: a new technique for modeling species-environment relationships [J]. Ecology, 2002, 83(4):1105 -1117.

[23] 赖江山, 米湘成, 任海保, 等. 基于多元回归树的常绿阔叶林群丛数量分类——以古田山 24 公顷森林样地为例[J]. 植物生态学报,2010, 34 (7): 761 -769.

[24] 曹铭昌, 周广胜, 翁恩生. 广义模型及分类回归树在物种分布模拟中的应用与比较[J]. 生态学报,2005,25(8): 2031 -2040.

[25] RIPLEY B D. Pattern recognition and neural networks [M]. Cambridge UK: Cambridge University Press, 1996.

[26] 谢益辉. 基于 R 软件 rpart 包的分类与回归树应用[J]. 统计与信息论坛,2007,22(5): 67 -70.

[27] VAYSSIERES M P, PLANT R E, ALLEN-DIAZ B H. Classification trees: an alternative non-parametric approach for predicting species distributions [J]. Journal of Vegetation Science, 2000, 11(5):679 -694.

[28] 刘洋, 吕一河, 郑海峰, 等. 用回归树模型分析陕北黄土丘陵沟壑区气候因子对 NDVI 变异的影响[J]. 应用生态学报, 2010, 21(5): 1153 -1158.

[29] ZHENG H F, CHEN L D, HAN X Z, et al. Classification and regression tree (CART) for analysis of soybean yield variability among fields in Northeast China: The importance of phosphorus application rates under drought conditions [J]. Agriculture, Ecosystems & Environment, 2009, 132 (1/2): 98 - 105.

[30] 官文江, 陈新军, 高峰, 等. 海洋环境对东、黄海鲐鱼(Scomber japonicus)灯光围网捕捞效率的影响[J]. 中国水产科学,2009, 16(6): 949 - 958.

[31] 官文江,陈新军,李纲. 海表水温和拉尼娜事件对东海鲐鱼资源及时空变动的影响[J]. 上海海洋大学学报,2011, 20(1):102 - 107.

[32] BREIMAN L, FRIEDMAN J H, OLSHEN R A, et al. Classification and regression trees [M]. California, USA: Wadsworth International Group, Belmont, 1984.

Comparisons of regression tree and GLM performance in CPUE standardization

GUAN Wen-jiang[1,2], CHEN Xin-jun[1,2], GAO Feng[1,2], LEI Lin[1,2]

(1. *College of Marine Sciences, Shanghai Ocean University, Shanghai* 201306, *China*; 2. *The Key Laboratory of Sustainable Exploitation of Oceanic Fisheries Resources, Ministry of Education, Shanghai Ocean University, Shanghai* 201306, *China*)

Abstract: CPUE (catch per unit effort) standardization is an essential task in fisheries stock assessment and GLM (generalized linear model) which has been used as a standardized method in the CPUE standardization. Before using GLM, the error structure, independent variables, and interaction between variables in the model had to be assigned and it would cause a great error if the assumption was wrong. Moreover, GLM could not be used to handle missing values automatically and to detect and extract complex interactions from the CPUE data. Outliers also had a great impact on the results estimated by using GLM. In contrast to GLM, regression trees may do a great job to deal with the above situations. In this paper, based on simulation data and chub mackerel (*Scomber japonicus*) catch and effort data from Chinese lighting-purse seine fishery in the East China Sea and Yellow Sea, we compared the performance of the regression tree and GLM in the CPUE standardization and the results showed that both models could do a good job if there were no outliers in the data and nonlinear relationships between nominal CPUE and abundance. Because the regression tree was characterized by a step function, the GLM was better in standardizing CPUE in this situation. However, if there were outliers and nonlinear relationships, the regression tree would harvest less root mean square errors and explain more deviations with fewer variables than GLM. The regression tree also could detect the complex relationships between independent variables and response variables by visualization which was ideally suited to explore and analyze the catch and effort data from fisheries.

Key words: catch per unit effort standardization; regression tree; generalized linear model