Advances in Social Sciences 社会科学前沿, 2019, 8(7), 1291-1298
Published Online July 2019 in Hans. http://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2019.87177

Hans汉斯

A Comparative Study on Forecasting the Size of Chinese Inbound Tourists Based on ARIMA and LSTM Neural Network

Yunfei Li

Jiangxi University of Finance and Economics, Nanchang Jiangxi
Email: 1764978463@qq.com

Received: Jul. 4th, 2019; accepted: Jul. 18th, 2019; published: Jul. 26th, 2019

Abstract

With the overall improvement of China's comprehensive national strength, China's tourism industry has entered a stage of rapid development. The number of inbound tourists is increasing. It is of great significance to accurately predict the scale of inbound tourists in China. This paper uses LSTM (Long Short-Term Memory) model and ARIMA (Autoregressive Integrated Moving Average Model) model to predict and compare the number of inbound tourists in China, and takes the number of inbound tourists from January 2014 to December 2016 as an example to conduct empirical research. The results show that LSTM neural network is more suitable than ARIMA for predicting the scale of inbound tourists in China, and the prediction accuracy of LSTM model is 22.981% higher than ARIMA. Predicting the number of inbound tourists based on LSTM model has certain guiding significance for relevant departments to optimize the allocation of tourism resources.

Keywords

Inbound Tourist Number, ARIMA Model, LSTM

基于ARIMA和LSTM神经网络对中国入境游客规模预测的比较研究

李云飞

江西财经大学，江西 南昌
Email: 1764978463@qq.com

收稿日期：2019年7月4日；录用日期：2019年7月18日；发布日期：2019年7月26日

文章引用：李云飞. 基于 ARIMA 和 LSTM 神经网络对中国入境游客规模预测的比较研究[J]. 社会科学前沿, 2019, 8(7): 1291-1298. DOI: 10.12677/ass.2019.87177

摘　要

随着我国综合国力的全面提升，我国旅游业也进入高速发展阶段，入境旅游人数日益增多，准确预测我国入境游客规模具有重要意义。本文分别使用LSTM (Long Short-Term Memory)模型和ARIMA (Autoregressive Integrated Moving Average Model)模型对我国入境游客人数进行预测对比，并以2014年1月至2016年12月的我国入境游客人次为例，进行实证研究。结果都表明LSTM神经网络比ARIMA更适合我国入境游客规模预测，LSTM模型预测精度比ARIMA高22.981%。基于LSTM模型预测入境游客人数，对相关部门优化旅游资源配置，具有一定的指导意义。

关键词

入境旅游人数，ARIMA模型，LSTM

1. 引言

近年来，随着我国综合国力的全面提升，经济的快速发展，以及改革开放取得的一系列重大成就，我国全球影响力日益提升，与此同时我国入境旅游人数也大幅增长。据《中华人民共和国文化和旅游部2018 年文化和旅游发展统计公报》所示，2018 年全年我国国内旅游人数 55.39 亿人次，比上年同期增长 10.8%；入境旅游人数 14,120 万人次，比上年同期增长 1.2%；全年实现旅游总收入 5.97 万亿元，占 GDP 比重 6.6%，纳入统计范围的全国各类文化和旅游单位 31.82 万个，从业人员 375.07 万人。由此可见，旅游业的发展不仅对中国经济的发展有重要作用，而且也能很大程度上增加我国的就业人数，对其他行业中也有重要的影响，所以预测我国游客人数在一定程度上具有重要意义。

本文以我国入境旅游人数为例，对其进行时间序列分析并预测，进而分析在我国旅游市场需求情况、市场竞争日趋激烈的环境下，我国入境旅游人数近三年的趋势，以及未来的发展情况，判断我国旅游产业的对外吸引力和影响力，给我国旅游业提供一定的参考。入境游客流量的准确预测有助于我国旅游部门制定相应的政策，资源合理配置，对客流进行合理分散导流，从而促进我国旅游业更好的发展[1]。因此，这在经济全球化进一步加快的大背景下，也有着极大的理论意义和较强的现实意义。

2. 研究方法

2.1. ARIMA 模型

具有如下结构(公式 1)的模型称为差分自回归移动平均模型[2]，简记 ARIMA (p, d, q)，ARIMA 模型根据原序列是否平稳以及回归中所含部分的不同，包括移动平均过程(MA)、自回归过程(AR)、自回归移动平均过程(ARMA)以及 ARIMA 过程，差分次数为 d。ARIMA 模型的公式如下为：

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \ldots + \Phi_p y_{t-p} + \varepsilon_t - \Theta_1 \varepsilon_{t-1} - \Theta_2 \varepsilon_{t-2} - \ldots - \Theta_q \varepsilon_{t-q} \quad (1)$$

上式中自回归系数用 $\Phi 1, \Phi 2, \cdots, \Phi p$ 表示，自回归阶数用 p 表示，$\Theta 1, \Theta 2 \cdots, \Theta q$ 是移动平均系数，q 是移动平均阶数，时间序列 y 当期值用 y_{t-1} 表示，y_{t-1} 表示 y_t 前一期的值，y_{t-2} 则为 y_{t-1} 前一期的值，依次类

推，误差项是当期随机干扰 ε_t，为零均值白噪声序列。建立 ARIMA 模型过程如下：

1) 原始数据预处理

首先对原序列进行平稳性检验，如果远序列不平稳，通过差分或者取对数处理，或者取对数后再差分，而一般差分不超两次。

2) 模型阶数识别

通过自相关图和偏自相关图确定大致 p、q 值，同时比较 AIC 的大小，选择 AIC 最小时的阶数作为模型阶数。

3) 模型参数估计

对确定好阶数的模型进行参数估计。由于游客人次时间序列可能包含季节因素，所有带季节性的 ARIMA 模型可能更能放映我国入境游客人数时间序列的特征。

4) 模型检验

残差序列白噪声检验，如果残差 ADF 检验没通过，或者残差图形不符合正太分布，那么模型是有效的，就进行预测，否则需要考虑对模型进行重新定阶和参数估计。

5) 使用模型预测

经过以上步骤，使用最终得到有效、合理的模型预测。

2.2. LSTM 神经网络

LSTM 网络是一种特殊 RNN (Recurrent Neural Networks)网络(递归神经网络)类型[3]，长期记忆单元(LSTM)于 1997 年被 Sepp Hochreiter 和 Jurgen Schmidhuber 提出，然后被 Alex Graves、Haim Sak 和 Wojciech Zaremba 等人逐步改进。LSTM 公式如下：

$$y_t = \tanh\left(wx_t + uy_{t-1}\right) \tag{2}$$

其中时间序列 y 的当期值为 y_t，时间序列 y 的影响因素 x 的当期值用 x_t 表示，时间序列 y 前一期的值为 y_{t-1}，权重为 w，转换参数为 u。

一般的 RNN 只能与前面若干序列有关，若一过十步，就很容易产生梯度消失或者梯度梯梯问题。产生梯度消失是因为导数的链式法则导致了连乘，造成梯度指数级消失。长短时记忆神经网络正是为了解决局部最优解这个问题而发展出来的，其每一层都设计有多个“GATE”结构的神经元，该结构使得模型得到进一步改善，即误差在整个向后传递的过程中，有一部分可以直接通过“GATE”，而不需要受到当前神经元的影响，在这种情况下，下一层神经元就能完全接收到完整的误差，优化的结果就是，无论梯度的传播路径有多么长，它都不会完全消失或者下降为零，因此它具有良好的收敛性。

LSTM 的基础结构和 RNN 基础结构一样，但是 RNN 与 LSTM 其中不同之处在于，在神经元结构中 RNN 只有一层 tanh 层，而 LSTM 的神经元结构要更复杂。LSTM 在每个神经元结构内部设置了三个门，分别是输入门、输出门和遗忘门。LSTM 结构中的三层门中遗忘门是解决 RNN 存在梯度消失问题的关键。

首先，当前输入向量 $x(t)$和前一个短期状态 $h(t-1)$被输入到四个不同的全连接层。它们都有不同的目的：

主层是输出为 $g(t)$的层。它的基本作用是分析当前输入 $x(t)$和前一个短期状态 $h(t-1)$。基本单元中就只有这一个层，它直接输出 $y(t)$和 $h(t)$。相比之下，LSTM 单元没有直接输出，而是将部分输出存储在长期状态中。其他三个层是门限控制器。因为使用了逻辑激活函数，它们的输出范围在 0 到 1 之间。它们的输出被输入到元素智能乘法操作中。因此如果输出是 0，那么门限关闭；如果输出是 1，那么门限打开。特别是：

遗忘门限(由 $f(t)$控制)控制着哪些长期状态应该被丢弃。

输入门限(由 $i(t)$控制)控制着 $g(t)$的哪些部分会被加入到长期状态(这就是我们说只是“部分存储”的原因)。

最后，输出门限由 $o(t)$控制着哪些长期状态应该在这个时间迭代被读取和输出 $h(t)$和 $y(t)$。

简而言之，LSTM 单元可以学习识别重要输入(这是输入门限的职责)，将其存储到长期状态中，学习需要时保存它(这是忘记门限的职责)，以及学习需要的时候提取它。这就解释了它为什么能够成功捕捉到时间序列中的长期模式、长文字、录音等。

3. 实证研究

本文所采用的数据来于《中国旅游统计年鉴》，2014 年 1 月至 2015 年 12 月共 24 个月的数据作为训练集，2016 年 1 月至 2016 年 12 月份的数据作为测试集，利用 Python 语言构建模型。并利用时间序列分析中的 ARIMA 模型、LSTM 神经网络对我国入境游客人数进行预测。获取数据见图 1：

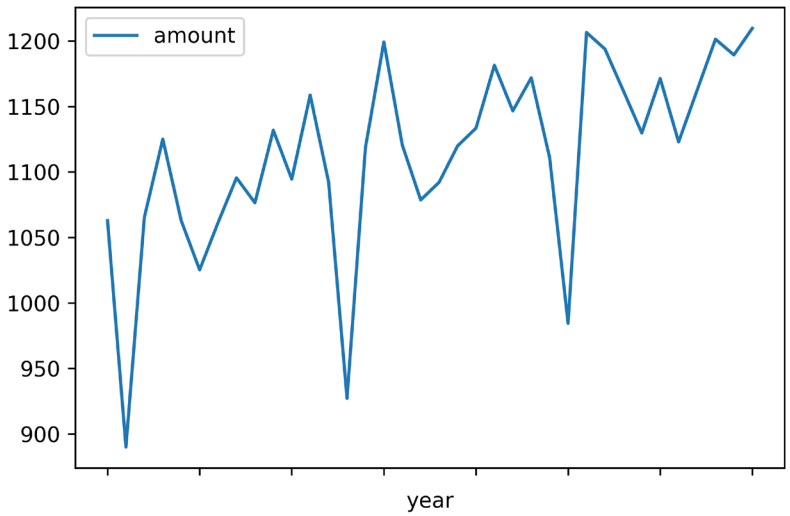

Figure 1. 2014~2016 China inbound number curve
图 1. 2014~2016 我国入境人数曲线

由上图可以看出，进三年我国入境游客人数整体呈上升趋势，部分月份人数比较少，可能有一定的季节性，后续需要对数据进行一定的处理。

3.1. ARIMA 模型的建立过程

首先对入境游客人次时间序列进行 ADF 检验，检验原序列的平稳性[4]，结果(如表 1)所示 $P > 0.05$，在 0.05 的显著性水平下，原序列不平稳；差分后的序列 ADF 检验结果见表 1 所示，差分次数 $d = 1$ 时 ADF 检验 P 值小于 0.05，模型平稳。

Table 1. ARIMA stationarity test results
表 1. ARIMA 平稳性检验结果

序列	ADF 检验 P 值
原序列	0.93481978
差分序列	3.11E−14
取对数后差分	4.63E−12

通过自相关图和偏自相关图初步确定 p、q 值的大小[5]，将初次确定的 p、q 值带入模型进行预测(见图 2)，根据 AIC 最小原则，不断调整 p 和 q 的大小，同时考虑旅游时间存在旺季和淡季，即是说将季节因素考虑到模型中，最终确定的游客人次 ARIAM 模型为 ARIMA (1, 1, 0) (见表 2)。

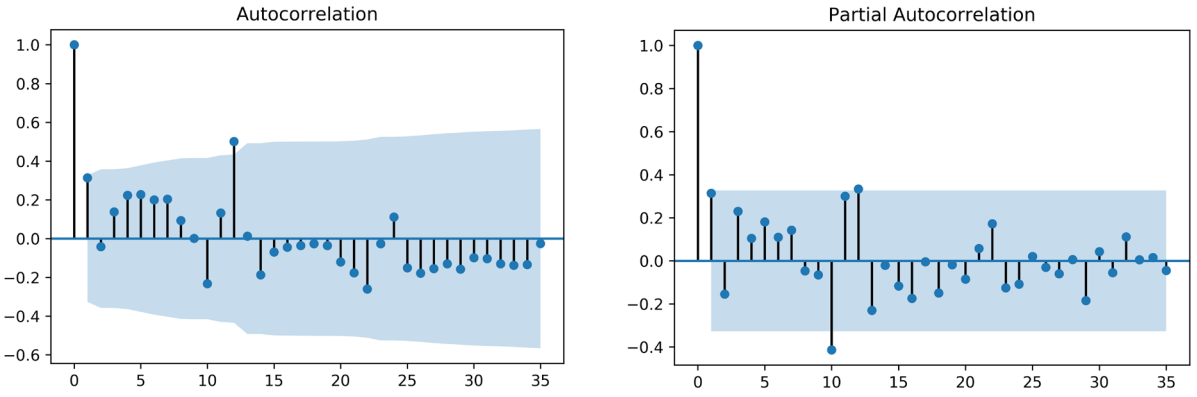

Figure 2. Autocorrelation and partial autocorrelation graphs
图 2. 自相关和偏自相关图

Table 2. ARIMA model results
表 2. ARIMA 模型结果

Results: ARMA			
Model	ARMA	BIC:	417.466
AIC:	412.7155	HQIC:	414.374
	Coef.	t	$P>\|t\|$
const	1113.831	66.6007	0.0000
ar.L1.amount	0.3251	2.0318	0.05
	Real	Imaginary	Modulus
AR.1	3.0763	0.0000	3.0763

ARIMA 模型的残差序列的 Ljung-Box 检验结果的 p 值依次为 0.459 > 0.05，在 0.05 的显著性水平下，残差序列为白噪声，表明所构建的模型是有效的(见图 3)。

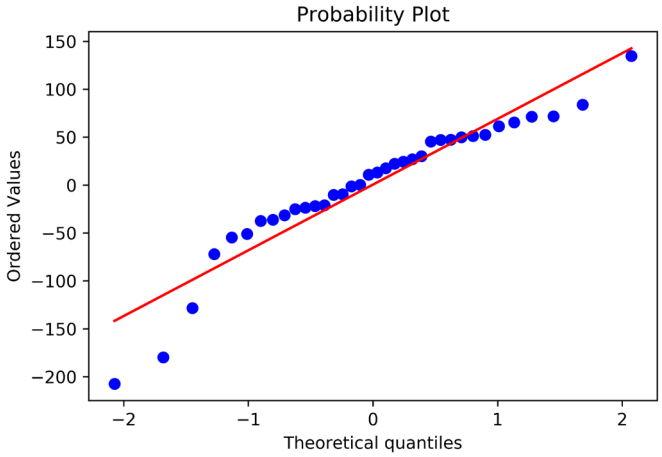

Figure 3. ARIMA residual graph
图 3. ARIMA 残差 QQ 图

从 QQ 图看出残差服从正太分布，残差序列为白噪声，再次表明所构建的模型是有效的。

ARIMA 模型预测(见表 3)我国入境游客数 RMSE (均方误差)为 69.0974129198503，由图 4 也可以直观的看出，ARIMA 模型对于数据的拟合效果并不是很好，只能提前部分信息，预测效果不是很好。

Table 3. ARIMA predict results
表 3. ARIMA 预测结果

日期	预测值
2017/1/1	1144.989
2017/2/1	1123.959
2017/3/1	1117.123
2017/4/1	1114.901
2017/5/1	1114.179
2017/6/1	1113.944

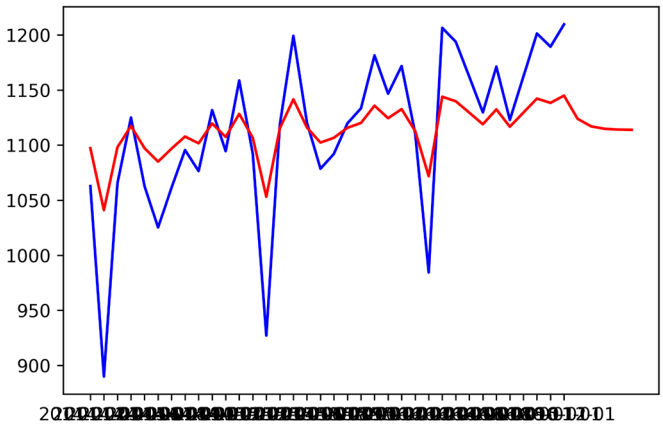

Figure 4. 2014~2016 ARIMA prediction curve
图 4. ARIMA 预测图

3.2. LSTM 神经网络的建立过程

本文基于 Python 的深度学习框架 Keras 来实现 LSTM 算法[6]，选用滚动划窗式的神经网络，输入是序列前 12 个月，例如 2014 年 1 月至 12 月作为输入，2015 年第 1 个月作为输出，2014 年第 2 个月至 2015 年第 2 个月作为输入，2015 年第 2 个月作为输出，后面依次类推。

本文构建的 LSTM 网络为三层的浅层网络[7]，将输入数据进行归一化处理，然后进行训练，最后再将输出结构反归一化，计算评价指标是均方误差(RMSE)，RMSE 越小，说明预测值与真实值越接近，预测的效果越好，训练集和测试集的评价指标分别反映模型的拟合能力和预测能力。

由图 5 可以看出，LSTM 模型对数据的拟合非常好[8]，与真实结果非常接近(见图 4)，由于 LSTM 模型不需要考虑模型的内部结构，学习能力和自适应能力都很强，给定数量的训练样本，反复学习训练样本的规律，学习到输入和输出之间的关系，从而实现相对较高的预测精度(见表 4)，在没有人工干预情况下[9]，具有较强的客观性。

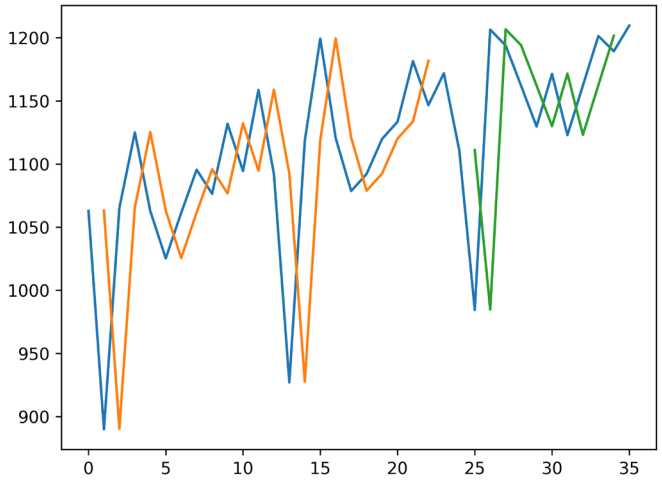

Figure 5. 2014~2016 LSTM prediction curve
图 5. LSTM 预测图

Table 4. LSTM predict results
表 4. LSTM 预测结果

日期	预测值
2017/1/1	1098.766
2017/2/1	1121.702
2017/3/1	1141.959
2017/4/1	1188.439
2017/5/1	1151.136
2017/6/1	1177.614

3.3. 模型对比分析

先利用 ARIMA 模型对序列进行预测，再利用 LSTM 网络对序列进行预测见表 5 可知入境游客人次序列的 LSTM 网络的训练集和测试集的 RMSE 和比 ARIMA 的低，表明了 LSTM 神经网络比神经网络和 ARIMA 预测更准确，LSTM 模型预测精度比 ARIMA 高 22.981%。

Table 5. ARIMA and LSTM comparison of prediction accuracy
表 5. ARIMA 与 LSTM 预测精度对比

模型	RMSE
ARIMA	69.097
LSTM	46.116

4. 结论

本文分别利用 Python 建立 ARIMA 模型和基于 Keras 深度学习框架建立 LSTM 神经网络模型对我国入境游客规模进行预测[10]，并利用入境游客人数序列对模型进行了验证，通过实证研究，总结出以下结论：第一，LSTM 神经网络考虑了时间因素，由于使用深度学习方法预测入境游客人数减少了人为因素的干预，且具有自动融合复杂因素的潜力，对我国入境游客人数预测与真实输出在总体趋势上均能达到

一致，较 ARIMA 对数据的拟合更好[11]，所以动态神经网络 LSTM 比 ARIM 更适合我国入境游客人数预测。第二，基于 LSTM 的模型比 ARIMA 模型对入境游客人次的预测更准确，LSTM 模型预测精度比 ARIMA 高 22.981%，所以和前人用静态神经网络进行时间序列的预测相比，动态神经网络预测我国入境游客规模会相对更合理[12]。本文建立的 LSTM 模型能对未来一段时间我国入境游客人数进行预测，能为相关管理部门进行资源优化配置、提升管理效率提供一定的参考。但如果希望更准确地预测我国入境游人数，只使用历史数据是不够全面的，由于入境游客规模受多种因素的影响[13]，如国家政策、经济发展、突发事件、地理环境等，所以更精确预测我国入境游客人次还应考虑更多其他因素。此外本文只使用了近两年的数据，没有使用更多的数据预测，同时该模型也存在一定的缺点和不足，特别是模型相对较单一，缺乏我国入境游客人数与其他对象之间的内在关系研究[14]，例如入境游客规模与 GDP、第三产业发展、人均收入等的相关性关系研究，这也是下一步研究的重点。

致　谢

感谢本文撰写期间导师对我的辛苦指导，以及同学们的热心帮助。也要感谢参考文献中的作者们，通过他们的研究文章，使我对研究课题有了很好的出发点。再次感谢。

参考文献

[1] 陆利军, 廖小平. 基于 EMD-BP 神经网络的游客量预测研究[J]. 统计与决策, 2019, 27(4): 3-5.

[2] 刘胜. 基于 ARIMA 与 SVM 组合模型的国内旅游市场预测研究[D]: [硕士学位论文]. 南昌: 东华理工大学, 2017.

[3] 徐一轩, 伍卫国, 王思敏. 基于长短期记忆网络(LSTM)的数据中心温度预测算法[J]. 计算机技术与发展, 2019, 30(6): 90-101.

[4] 谢小军, 邱云兰, 时凌. 基于 ARIMA 和 BP 神经网络组合模型的能源消费预测[J]. 数学的实践与认识, 2019, 32(10): 292-298.

[5] 朱家明, 胡玲燕. 基于 ARIMA 和 BP 神经网络对人民币汇率预测的比较分析——以美元人民币汇率为例[J]. 重庆理工大学学报(自然科学), 2019, 35(5): 207-212.

[6] 史亚星. 基于深度学习的车流量预测方法研究[J]. 计算机与数字工程, 2019, 27(5): 1160-1163.

[7] 郑洋洋, 白艳萍, 侯宇超. 基于 Keras 的 LSTM 模型在空气质量指数预测的应用[J]. 数学的实践与认识, 2019(7): 138-143.

[8] 冯宇旭, 李裕梅. 基于 LSTM 神经网络的沪深 300 指数预测模型研究[J]. 数学的实践与认识, 2019, 49(7): 308-315.

[9] 段大高, 赵振东, 梁少虎, 等. 基于 LSTM 的 PM2.5 浓度预测模型[J]. 计算机测量与控制, 2019, 27(3): 221-225.

[10] 杨青, 王晨蔚. 基于深度学习 LSTM 神经网络的全球股票指数预测研究[J]. 统计研究, 2019, 30(3): 65-77.

[11] 罗龙, 李两桓, 王成阳, 等. 基于 ARIMA-LSTM 的绝缘子状态数据挖掘方法[J]. 电力科学与技术学报, 2017(4): 38-43.

[12] 孙烨, 张宏磊, 刘培学, 等. 基于旅游者网络关注度的旅游景区日游客量预测研究——以不同客户端百度指数为例[J]. 人文地理, 2017, 32(3): 158-166.

[13] 王慧, 陈秋华, 修新田, 等. 基于 BP 神经网络的森林旅游景区环境承载力预警系统构建研究——以太岳山国家森林公园石膏山景区为例[J]. 林业经济, 2018, 40(3): 58-64.

[14] Song, Y. and Li, G. (2008) Tourism Demand Modeling and Forecasting—A Review of Recent Research. *Tourism Management*, **29**, 203-220. https://doi.org/10.1016/j.tourman.2007.07.016