

基于Holt-Winters模型的传染病数据分析

杨 斌

淮阴师范学院计算机科学与技术学院, 江苏 淮安

收稿日期: 2024年11月10日; 录用日期: 2024年12月6日; 发布日期: 2024年12月12日

摘 要

如何准确预测传染病趋势对于指导公共卫生政策和社会经济活动至关重要。本文应用一种时间序列分析方法Holt-Winters模型, 对传染病数据进行分析, 以预测其发展和季节性变化, 利用Python进行模型建立和参数优化, 并给出具体的分析结果。通过比较不同预测模型, 进一步探讨了Holt-Winters模型在传染病预测中的优势和局限, 并提出了改进模型预测准确性的策略, 研究结果不仅为制定有效的传染病防控措施提供了科学依据, 也为传染病预测模型的研究提供了新的视角。

关键词

Holt-Winters模型, 传染病预测, 时间序列分析, Python

Infectious Disease Data Analysis Based on Holt-Winters Model

Bin Yang

School of Computer Science and Technology, Huaiyin Normal University, Huaian Jiangsu

Received: Nov. 10th, 2024; accepted: Dec. 6th, 2024; published: Dec. 12th, 2024

Abstract

Accurately predicting infectious disease trends is crucial for guiding public health policies and socio-economic activities. In this paper, a time series analysis method, the Holt-Winters model, is applied to analyze infectious disease data to predict its development and seasonal variations. The study involves model establishment and parameter optimization using Python, and specific analysis results are provided. By comparing different forecasting models, the advantages and limitations of the Holt-Winters model in infectious disease prediction are further discussed. Strategies

文章引用: 杨斌. 基于 Holt-Winters 模型的传染病数据分析[J]. 统计学与应用, 2024, 13(6): 2244-2248.

DOI: 10.12677/sa.2024.136217

to improve the accuracy of model predictions are proposed, offering scientific basis for effective infectious disease prevention and control measures and providing a new perspective for research on infectious disease prediction models.

Keywords

Holt-Winters Model, Infectious Disease Prediction, Time Series Analysis, Python

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

传染病的传播和发展对全球公共卫生和经济社会造成了巨大的冲击。准确地分析传染病数据的趋势和季节性特征，对于传染病防控和决策制定至关重要。时间序列分析模型是一种常用的数据分析工具，可以用于揭示数据的变化规律和趋势。Holt-Winters 模型作为一种经典的时间序列预测模型，能够有效地处理具有趋势和季节性的数据，因此在传染病数据分析中具有广泛的应用前景[1]。

通过对传染病数据的分析和预测，可以为政府和卫生部门提供科学依据，帮助他们制定更加有效的传染病防控策略，合理分配资源，及时采取措施控制传染病的传播。深入了解传染病的发展趋势和季节性特征，有助于优化公共卫生管理体系，提高应对突发公共卫生事件的能力，保障公众的健康和安全。准确的传染病预测可以为企业和社会提供参考，帮助他们做好生产和经营计划的调整，减少传染病对经济社会的负面影响，促进经济社会的稳定发展。

本文有助于丰富传染病数据分析的方法和理论，推动相关领域的学术研究进展，为今后的传染病防控和公共卫生管理提供有益的经验借鉴。基于 Holt-Winters 模型的传染病数据分析具有重要的现实意义和学术价值，对于应对传染病挑战、保障公众健康和促进经济社会发展具有积极的作用。

2. Holt-Winters 模型简介[1] [2]

2.1. 模型原理和基本形式

Holt-Winters 模型是一种基于指数平滑法的时间序列预测模型，它通过对历史数据的平滑处理来预测未来的数据。该模型考虑了数据的趋势和季节性因素，能够较好地捕捉数据的变化规律。

Holt-Winters 模型包括三个部分：水平项(Level)表示数据的基本水平或均值；趋势项(Trend)描述数据的趋势变化，如上升、下降或保持稳定；季节性项(Seasonal)用于捕捉数据的季节性波动。

$$L_t = \alpha(Y_t - S_{t-m}) + (1 - \alpha)(L_{t-1} + B_{t-1})$$

$$B_t = \beta(L_t - L_{t-1}) + (1 - \beta)B_{t-1}$$

$$S_t = \gamma(Y_t - L_t) + (1 - \gamma)S_{t-m}$$

其中， L_t 表示第 t 期的水平项， B_t 表示第 t 期的趋势项， S_t 表示第 t 期的季节性项， Y_t 表示第 t 期的实际观测值， m 表示季节周期， α 、 β 和 γ 分别为水平项、趋势项和季节性项的平滑参数，取值范围在 0 到 1 之间。

2.2. 模型参数估计和预测

- 1) 确定季节周期(m)：根据传染病数据的特点和经验，确定季节周期的长度。

- 2) 初始化水平项(L0)、趋势项(B0)和季节性项(S0): 使用数据的前几个观测值来初始化这些参数。
- 3) 估计平滑参数(α 、 β 、 γ): 使用最小二乘法估计平滑参数的值, 以使模型能够最好地拟合历史数据。这些参数控制着模型对历史数据的平滑程度和对趋势、季节性的响应程度。
- 4) 迭代更新参数: 根据模型的公式, 在每个时间点上迭代更新水平项、趋势项和季节性项的估计值。具体来说, 对于时间点 t :
 - ① 水平项更新: $L_t = \alpha(Y_t - S_{t-m}) + (1-\alpha)(L_{t-1} + B_{t-1})$ 。
 - ② 趋势项更新: $B_t = \beta(L_t - L_{t-1}) + (1-\beta)B_{t-1}$ 。
 - ③ 季节性项更新: $S_t = \gamma(Y_t - L_t) + (1-\gamma)S_{t-m}$ 。
- 5) 使用训练好的模型, 根据当前的水平项、趋势项和季节性项的估计值, 以及未来的时间步数(h), 进行预测。预测公式为: $\hat{Y}_{t+h} = L_t + hB_t + S_{t+h-m}$, 其中 $\hat{Y}_{t+h}$ 表示对未来第 $t+h$ 期的预测值。重复上述步骤, 对未来多个时间点进行预测。

3. Holt-Winters 模型的建立与应用

本文中的传染病数据使用 HIV 数据, 来源于相关政府部门、卫生机构或权威数据平台的发布数据。传染病数据是按照时间顺序排列的, 具有明显的时间序列特征。这使得我们可以运用时间序列分析方法来研究传染病的发展趋势和规律。

3.1. 模型建立与参数选择

1) 确定季节周期

首先, 需要确定时间序列中存在的季节周期长度。这通常取决于数据的采样频率(例如, 每天、每周、每月等)。对于传染病数据, 常见的周期可能是每周或每月, 因为某些传染病可能显示出每周或每月的周期性变化。在本文中, 我们假设数据是以天为单位记录的, 并且可能存在一个月左右的季节性变化, 因此设定季节周期为 30 天。

2) 选择合适的趋势和季节性模型

Holt-Winters 模型有几种不同的形式, 具体取决于数据中存在的趋势和季节性特征[3]:

- ① Additive (加性): 当数据的趋势和季节性是恒定的时候, 即不会随着水平的变化而改变幅度时, 选择加性模型。
- ② Multiplicative (乘性): 当数据的趋势和季节性随水平的增加而成比例增加时, 选择乘性模型。

对于本文数据, 通过观察数据的图形来决定是使用加性还是乘性模型。因数据中没有明显的证据表明季节性随时间序列水平的增加而成比例增长, 故而使用加性模型。

3) 模型参数的估计

确定了模型的形式(趋势 + 季节性), 然后进行模型参数的估计。Holt-Winters 模型的参数包括 α 、 β 、 γ , 参数的值决定了模型对新数据的反应速度以及对旧数据的记忆程度。本文中直接调用使用“stats models”库的“Exponential Smoothing”类进行模型的构建和参数估计。

3.2. 基于 Python 的模型拟合与展示

本文中的国外数据包括: date (日期)、inc (新增病例数)、sum (累计病例数)。为了建立 Holt-Winters 模型, 使用 inc 列(新增病例数)作为我们的时间序列数据。详细数据如见表 1。

为了评估模型的性能, 我们使用一些标准的评估指标: 均方误差(Mean Squared Error, MSE)、均方根误差(Root Mean Squared Error, RMSE), 根据国外数据计算得 MSE: 712914885834.091, RMSE:

844342.8721994939。

得到模型之后，使用拟合好的模型对未来一段时间内的传染病发展趋势进行预测。

最后，我们可以使用 `matplotlib` 库来绘制实际值、拟合值以及预测值，以便更直观地展示模型的效果，结果显示见图 1。

Table 1. Exponential smoothing model results

表 1. 指数平滑模型结果

Dep. Variable:	inc	No. Observations:	144
Model:	Exponential Smoothing	SSE	1111.442
Optimized:	True	AIC	362.278
Trend:	Additive	BIC	463.252
Seasonal:	Additive	AICC	387.176
Seasonal Periods:	30	Box-Cox:	False
Box-Cox Coeff.:	None		

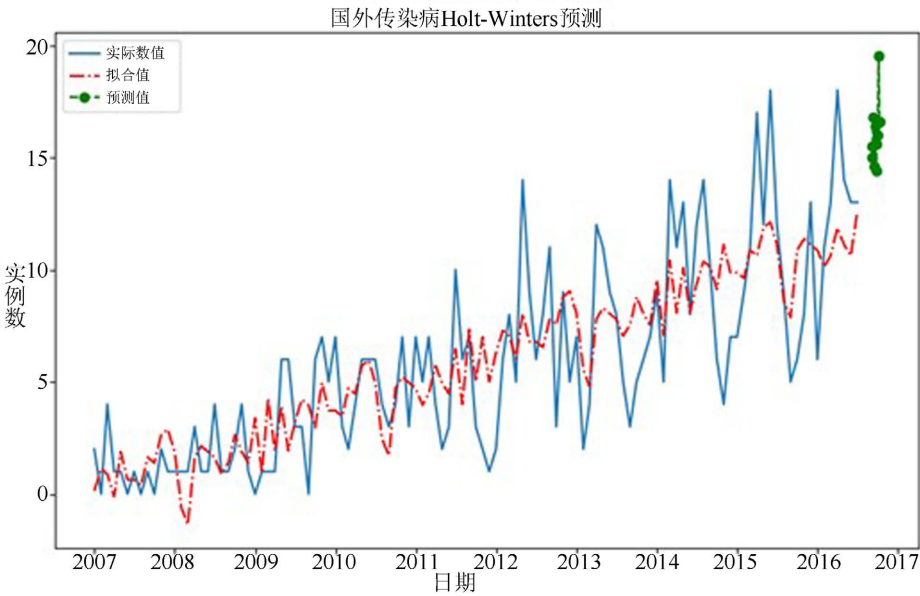


Figure 1. Holt-Winters forecasting plot of foreign infectious diseases

图 1. 国外传染病 Holt-Winters 预测图

4. 结果与讨论

4.1. 与其他模型或方法的比较

Holt-Winters 模型相比传统的简单指数平滑模型，能够更好地处理具有趋势和季节性成分的时间序列数据。例如，ARIMA 模型也可以用于处理具有趋势的时间序列数据，但是它对于季节性的处理不如 Holt-Winters 模型直观和有效。

除了 Holt-Winters 模型之外，还有许多其他模型和技术可以用于传染病预测，例如：

- ① SIR 模型：用于模拟传染病的传播过程[4]。
- ② 机器学习模型：如随机森林[5]、支持向量机[6]等，可以利用历史数据进行复杂的非线性预测。

每种模型都有其适用场景和优缺点。例如，SIR 模型更适合于理解传染病的传播机制，而机器学习模型可能在处理大规模复杂数据时表现更好。

4.2. 模型的局限性和改进方向

Holt-Winters 模型有一些基本的假设，包括数据必须是平稳的(即均值和方差不随时间变化)，并且模型的预测能力受限于数据的历史模式。此外，模型可能无法很好地应对突发事件(如新的病毒变种出现)导致的突然变化。

为了改善模型的预测性能，可以考虑以下几点：

- 1) 引入外部变量：例如天气条件、人口流动等，可以作为输入变量加入模型中，以增强预测能力。
- 2) 组合模型：结合不同类型的模型(如 Holt-Winters 与机器学习模型)，以互补各自的优点。
- 3) 考虑非线性关系：对于某些情况，非线性关系可能更为显著，可以尝试使用非线性模型或深度学习技术。

5. 结论

本文通过建立 Holt-Winters 模型对传染病数据进行了分析和预测，揭示了传染病的发展趋势和季节性特征。模型的建立和应用结果表明，Holt-Winters 模型能够有效地处理具有趋势和季节性的时间序列数据，为传染病防控和决策提供了有价值的参考。

尽管 Holt-Winters 模型在传染病数据分析中显示出了其优势，但也存在一些局限性。为了提高预测的准确性和鲁棒性，未来的研究可以考虑引入外部变量、采用组合模型和考虑非线性关系等方法来改进模型。

此外，随着技术的进步，可以考虑将深度学习等先进的预测技术应用于传染病预测，以进一步提高预测的准确性和实用性。通过不断的研究和实践，我们可以更好地理解 and 应对传染病，为公共卫生安全和社会经济发展做出贡献。在未来，随着深度学习技术的发展，可以考虑将深度学习模型应用于传染病预测。深度学习模型在处理复杂的非线性关系和大规模数据集时通常具有更好的性能，并且可以自动地学习时间序列数据的特征表示。

基金项目

淮安市基础 Research 计划(指导性)项目(HABZ202315)。

参考文献

- [1] 林海, 康宝中. 基于 Holt-Winters 时间序列的图书选题预测模型[J]. 数字技术与应用, 2017(3): 51-53.
- [2] 郭在金, 龚浩, 周罗晶. SARIMA 模型和 Holt-Winters 指数平滑法在江苏省肺结核发病数预测中的应用[J]. 疾病监测, 2022, 37(8): 1042-1047.
- [3] 侯文涛, 杨郁辉, 胡运红. SARIMA 模型与 Holt-Winters 模型在丙型肝炎发病率预测中的应用[J]. 数学的实践与认识, 2021, 51(4): 172-179.
- [4] 张慧敏. 具有垂直传染和脉冲常量接种的 SIR 模型[J]. 中北大学学报(自然科学版), 2007(S1): 1-3.
- [5] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-38.
- [6] 丁世飞, 齐丙娟, 谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报, 2011, 40(1): 2-10.