《疾病监测》审稿意见与作者答复

题目：时间序列分析在北京市东城区 HIV/AIDS 发病率预测中的应用

作者：王媛媛 田飞 刘晶磊

————————————————审稿专家意见与答复————————————————

初审专家意见及作者修改说明：

专家一：

1、文中简单称差分后序列更平稳，证据何在？

平稳序列是指不存在趋势的序列，各观察值基本上在某个固定水平上波动，或虽然有波动，但是并不存在某种规律，其波动可以看成是随机的。

本文中 2005-2014 年原始数据的时间序列有明显的上升趋势和季节性的规律变化，通过参考文献得知将原始数据进行自然对数转换再进行一阶差分运算和一次季节差分来消除趋势和季节的影响。平稳序列是结合原序列经过对数转换和两次差分后的序列图来看的，如下图所示。经过处理后的数据在（-2,2）之间波动，不存在明显的规律。

2、文中最后使用的校检是内校检，即用什么数据拟合模型就用什么数据验证模型，一般来说这种校检的说服力不足，一般使用外校检，比如预测一下 2015 的数据。

ARIMA（0,1,1）×（0,1,1）12 模型对 2015 年 1-12 月 HIV/AIDS 月发病率进行了预测，如下表，但是考虑到艾滋病实际发病率比较敏感，单位学术部门不建议将下表列在文章中，因此只是在文章中提及“用该模型预测北京市东城区 2015 年 1-12 月 HIV/AIDS 月发病率，实际月发病率均在预测值 95%可信区间内”。

2015 年 1-12 月 HIV/AIDS 实际月发病率与预测发病率比较（1/10 万）

月份	实际发病率	预测发病率	95%CI 下限	95%CI 上限
1 月	4.27	4.61	1.60	10.52
2 月	1.87	2.85	0.97	6.59
3 月	5.52	5.89	1.96	13.82
4 月	5.61	4.89	1.59	11.63
5 月	5.61	5.13	1.63	12.36
6 月	6.68	6.39	1.99	15.59
7 月	7.93	6.72	2.05	16.61
8 月	4.36	5.80	1.73	14.50
9 月	4.63	5.21	1.52	13.17
10 月	4.63	4.39	1.25	11.22
11 月	4.99	5.05	1.41	13.06
12 月	4.54	7.60	2.09	19.89

3、表头应使用中文。

表头已修改

专家二：

（1)在材料与方法里,“本文数据选取全国传染病网络直报系统报告的 2005-2015 年北京市东城区每月新发 HIV/AIDS 感染者人数”,说明是常住人口、户籍人口还是所有登记的感染者。

2005-2015 年辖区医疗机构、CDC 新报告的病例，包括现住址为本区县、外区县及外省的全部感染者。

（2）1.1 模型介绍，第一句话没有写完整。

已修改：70 年代初，博克思(Box)和詹金斯(Jenkins)提出了著名的时间序列分析中 ARIMA 模型，即自回归移动平均模型（Autoregressive Integrated MovingAverage Model，ARIMA），因此又被称为 Box-Jenkins 模型。

（3）图 2 原序列经过对数转换和两次差分后的序列图，这个图不规范，如横坐标的刻度太少，纵坐标是什么没有说明，请修改。

已修改：横坐标刻度改为 2005-2014 年 HIV/AIDS 月发病率经过对数转换和两次差分后的 120 个数据。纵坐标 lx(1 12)为经过对数转换和两次差分后的数值，在（-2,2）之间波动，用处理后的数据进行 ARIMA 模型识别。

（4)在 2.3 模型的预测，作者指出：预测值与实际值的平均绝对比例误差(MAPE)为 20.18%，根据时间序列模型的预测评价理论[8]，MAPE<20%表示预测模型良好，这提示模型依然有可优化的空间。那么，从几个角度或方面可以考虑提高模型的拟合优度？

1、从数据考虑的话，可以增加观测值，也就是增加年份进行观测 HIV/AIDS 月发病率，但是考虑到全国传染病网络直报系统对 HIV/AIDS 的报卡和监测从 2004 年后逐渐完善，所以增加观测值的观测年份只能后延。

2、可以从统计方法上引入“含自变量的 ARIMA 预测模型”，将每年的宣传干预人数、哨点监测总人数、抗病毒治疗人数、艾滋病知识知晓率等自变量引入 ARIMA 模型，但是同样面临获取历史数据的难题。同时含自变量的 ARIMA 预测模型的研究相对比较少，而该模型的复杂程度会很高，对模型的选择和解释有一定难度。

（5）影响拟合的因素很多，正如作者指出:诸如艾滋病基本知识的宣传力度，高危人群的危险性行为干预的效果、哨点监测任务量的变化、最新的 HIV 检测方法的运用、抗病毒治疗的覆盖率、流动人口的持续增加等,加上数据的漏报问题，AIDS 的趋势有人为的影响因素在里面。

是的，HIV/AIDS 发病率的预测本身有一定的难度，本文研究的主要目的是探讨应用时间序列 ARIMA 模型预测北京市东城区 HIV/AIDS 发病率的可行性，尽管使用 ARIMA 模型预测的结果：2015 年 1-12 月 HIV/AIDS 月发病率，实际月发病率均在预测值 95%可信区间内。但是预测值与实际值的平均绝对比例误差(MAPE)值偏高，说明 ARIMA 模型预测精度不高。我希望通过这篇文章的发表，可以给其他人以帮助，单纯考虑时间的 ARIMA 模型预测 HIV/AIDS 发病率是可行的，但是却不是一个预测准确度很高的模型。因此，后续我计划再研究 BP 神经网络模型预测等其他模型，比较不同模型的预测精度，找出一个更加适合预测 HIV/AIDS 发病率的方法。

专家三：

1. 所用 HIV/AIDS 数据为报告日期，请补充监测能力与报告数据的关系，监测时间、网络报告时间对与数据的影响。

本文所用数据为 2005-2015 年辖区医疗机构、CDC 新报告的病例，包括现住址为本区县、外区县及外省的全部感染者。辖区医疗机构发现 HIV/AIDS 阳性的主要途径包括：术前检测、性病门诊、受血（制品）前检测、其他就诊者检测、孕产期检查 等；东城区 CDC 发现 HIV/AIDS 阳性的主要途径包括：自愿咨询检测（VCT）、阳性者配偶或性伴检测、男男同性恋人群社区小组干预与检测，东城区高危人群哨点监测（社区暗娼、社区吸毒、在押吸毒、在押暗娼、在押嫖客、流动人口等）。

如果辖区医疗机构及 CDC 送检确证的血样诊断为 HIV/AIDS 阳性，24 小时内必须报卡，网络报告时间因为有严格的规定，没有漏报的可能，因此对数据没有影响。

由于东城区流动人口较多，每年年初和年末报告新发 HIV/AIDS 感染者人数较少。哨点监测工作的开展每年集中在 4-7 月份，这段时间积极开展动员监测工作，HIV/AIDS 报告数量会上升，因此数据呈现出季节性。

2.文章用该模型预测北京市东城区 2015 年 1-12 月 HIV/AIDS 月发病率，预测值与实际值的平均绝对比例误差(MAPE)为 20.18%。请提供 2015 年 1-12 月发病与预测值的数据及检验依据。

ARIMA（0,1,1）×（0,1,1）12 模型对 2015 年 1-12 月 HIV/AIDS 月发病率进行了预测，如下表，但是考虑到艾滋病实际发病率比较敏感，单位学术部门不建议将下表列在文章中，因此只是在文章中提及“用该模型预测北京市东城区 2015 年 1-12 月 HIV/AIDS 月发病率，实际月发病率均在预测值 95%可信区间内”。

2015 年 1-12 月 HIV/AIDS 实际月发病率与预测发病率比较（1/10 万）

月份	实际发病率	预测发病率	95%CI 下限	95%CI 上限
1 月	4.27	4.61	1.60	10.52
2 月	1.87	2.85	0.97	6.59
3 月	5.52	5.89	1.96	13.82
4 月	5.61	4.89	1.59	11.63
5 月	5.61	5.13	1.63	12.36
6 月	6.68	6.39	1.99	15.59
7 月	7.93	6.72	2.05	16.61
8 月	4.36	5.80	1.73	14.50
9 月	4.63	5.21	1.52	13.17
10 月	4.63	4.39	1.25	11.22
11 月	4.99	5.05	1.41	13.06
12 月	4.54	7.60	2.09	19.89

复审专家意见及作者修改说明：

1、本文针对的是感染绝对数的预测，而在疾病监测中我们更重视率，请作者就此进行说明；

本文使用时间序列 ARIMA 模型预测北京市东城区 HIV/AIDS 发病率，由（2005-2015 年每月北京市东城区辖区内新报告的 HIV/AIDS 感染者人数/年度辖区总人口数）计算 2005-2015 年月 HIV/AIDS 发病率。

2、使用该模型，作者也称需要序列平稳，但本文缺乏对平稳性的检验；

给出一个随机的时间序列，可以通过序列图来判断是否平稳，平稳的时间序列在图形上往往表现出一种围绕其均值不断波动的过程；而非平稳序列则往往表现出在不同时间段具有不同的均值，例如持续上升、下降或明显的季节波动。

本文对东城区 2005-2014 年 HIV/AIDS 月发病率做序列图，有明显的上升趋势和季节性，是不平稳序列，参考以往研究，不平稳序列可以通过对数转换和差分的方法使其平稳化，因此本文进行自然对数转换、一阶差分运算和一次季节差分来消除趋势和季节的影响，两次差分后的序列图，如下图所示。经过处理后的数据在（-2,2）之间波动，不存在明显的规律。判断处理后的序列图为平稳序列。

3、对 2015 年的预测校检光凭一张图一段描述不够，需要有具体的预测值与实际值的对照表，及各点偏差。

在本文初稿中用 ARIMA（0,1,1）×（0,1,1）12 模型对 2015 年 1-12 月 HIV/AIDS 月发病率进行了预测，如下表。

但是考虑到艾滋病实际月发病率比较敏感，单位学术部门依旧不建议将下表列在文章中。采纳专家意见，为了使实际值和预测值对照更加明显，将图 3 更改为“2015 年 1-12 月 HIV/AIDS 实际发病率、预测发病率及预测值 95%可信区间”折线图。

2015 年 1-12 月 HIV/AIDS 实际月发病率与预测发病率比较（1/10 万）

月份	实际发病率	预测发病率	95%CI 下限	95%CI 上限
1 月	4.27	4.61	1.60	10.52
2 月	1.87	2.85	0.97	6.59
3 月	5.52	5.89	1.96	13.82
4 月	5.61	4.89	1.59	11.63
5 月	5.61	5.13	1.63	12.36
6 月	6.68	6.39	1.99	15.59
7 月	7.93	6.72	2.05	16.61
8 月	4.36	5.80	1.73	14.50
9 月	4.63	5.21	1.52	13.17
10 月	4.63	4.39	1.25	11.22
11 月	4.99	5.05	1.41	13.06
12 月	4.54	7.60	2.09	19.89

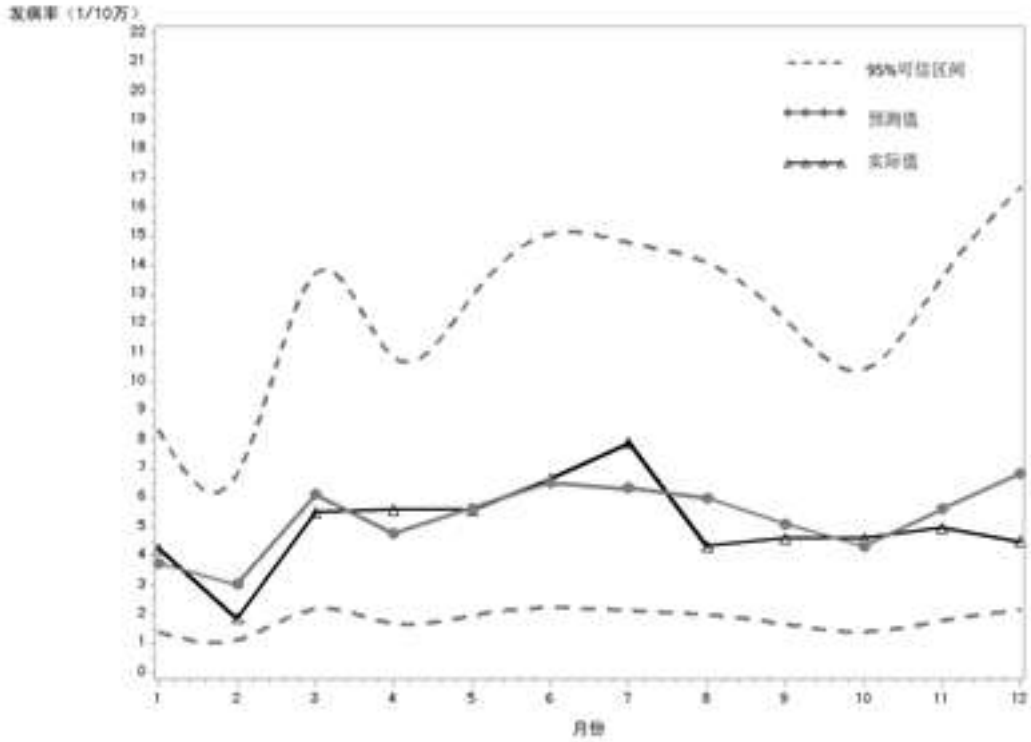

图 3 北京市东城区 2015 年 1-12 月 HIV/AIDS 实际发病率、预测发病率及预测值 95%可信区间

————————————————————定稿会意见与答复————————————————————

定稿会意见：

建议进一步修改： 1. 引言中第一段和第二段衔接感觉不紧密，建议第一段“近些年，我国艾滋病主要传播途径为性传播为主，其中男性同性性传播比例上升明显。由于流动人口断向不大中型城市聚集，大中城市的艾滋病疫情呈上升趋势[2]。”前移至“可导致死亡的综合征”后，删除“2005 年以来,全国范围内 HIV/AIDS 新发病例数持续上升,”一句。2. 讨论中第二段是北京市艾滋病流行特征，与模型预测区域及研究结果无直接关系，作者写此段想表达的意思不明确，建议删除此段。

1. 引言中第一段和第二段衔接感觉不紧密，建议第一段“近些年，我国艾滋病主要传播途径为性传播为主，其中男性同性性传播比例上升明显。由于流动人口断向不大中型城市聚集，大中城市的艾滋病疫情呈上升趋势[2]。”前移至“可导致死亡的综合征”后，删除“2005 年以来,全国范围内 HIV/AIDS 新发病例数持续上升,”一句。
2. 讨论中第二段是北京市艾滋病流行特征，与模型预测区域及研究结果无直接关系，作者写此段想表达的意思不明确，建议删除此段。
处理：已按照专家意见进行了相应“增、删、改”处理

本文经这次修改后，基本达到要求，可以发表，谢谢！