Advances in Applied Mathematics 应用数学进展, 2021, 10(11), 4114-4122
Published Online November 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.1011437

Hans汉斯

四川省艾滋病发病人数模型及预测

李　姣，戴家佳

贵州大学数学与统计学院，贵州　贵阳

收稿日期：2021年10月29日；录用日期：2021年11月19日；发布日期：2021年11月30日

摘　要

为探究四川省艾滋病发病人数的趋势，本文利用残差修正GM(1,1)模型和BP (Back Propagation)神经网络模型对发病人数进行预测并对预测效果进行比较。根据四川省2005年第1季度至2016年第4季度艾滋病发病人数建立的残差修正GM(1,1)模型和BP神经网络模型，对2017年第1季度至第4季度发病人数进行预测。残差修正GM(1,1)模型预测出2017年四川省艾滋病发病人数的平均绝对误差(MAE)、平均绝对百分比误差(MAPE)分别为1019和0.4023；BP神经网络模型预测出2017年四川省艾滋病发病人数的MAE、MAPE分别为236和0.0697。BP神经网络模型相较于残差修正GM(1,1)模型能更好地拟合四川省艾滋病的发病趋势，因此更适用于四川省艾滋病发病人数的短期预测。

关键词

艾滋病，残差修正GM(1,1)模型，BP神经网络模型

Model and Prediction of AIDS Incidence in Sichuan Province

Jiao Li, Jiajia Dai

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Oct. 29th, 2021; accepted: Nov. 19th, 2021; published: Nov. 30th, 2021

Abstract

In order to explore the trend of the number of AIDS cases in Sichuan province, this paper uses the residual correction GM(1,1) model and the BP (Back Propagation) neural network model to predict the number of cases and compare the prediction results. Based on the data on the number of AIDS cases in Sichuan Province from the first quarter of 2005 to the fourth quarter of 2016, a residual correction GM(1,1) model and a BP neural network model were established to predict the

文章引用: 李姣, 戴家佳. 四川省艾滋病发病人数模型及预测[J]. 应用数学进展, 2021, 10(11): 4114-4122.
DOI: 10.12677/aam.2021.1011437

number of cases from the first quarter to the fourth quarter of 2017. The residual correction GM(1,1) model predicted that the average absolute error (MAE) and average absolute percentage error (MAPE) of the number of AIDS cases in Sichuan Province in 2017 were 1019 and 0.4023, respectively; the BP neural network model predicted that MAE and MAPE of the number of AIDS cases in Sichuan Province in 2017 were 236 and 0.0697, respectively. Compared with the residual correction GM(1,1) model, the BP neural network model can better fit the incidence trend of AIDS in Sichuan Province, therefore, it is more suitable for short-term prediction of AIDS incidence in Sichuan Province.

Keywords

AIDS, GM(1,1) Model, BP Neural Network Model

1. 引言

艾滋病，全称获得性免疫缺陷综合症(acquired immune deficiency syndrome, AIDS)，是由艾滋病病毒(HIV)引起并导致人类免疫功能受损，对人类造成严重危害的一种恶性传染性疾病[1]。艾滋病具有传染性强、致死率高的特点，并且目前尚无有效的彻底治愈方法，已经成为一个全球性的公共卫生问题[2]。本病的传播途径主要是母婴传播、性传播和血液传播等[3]。

四川是一个人口大省，艾滋病的发病人数也在全国位居前列，所以对艾滋病的预测和防控就显得尤为重要。国内外研究传染病发病人数的方法有很多，例如时间序列模型、灰色预测模型、动力学模型[4]、机器学习方法等，并且都取得了不错的效果，其中应用最广的是时间序列模型。但采用灰色 GM(1,1)模型和机器学习方法对四川省艾滋病发病人数进行分析的文献相对较少，基于此，为了比较这两种方法在预测效果上的优劣以及得到较为可靠的未来艾滋病发病人数，本文采用残差修正 GM(1,1)灰色预测模型以及 BP 神经网络模型来对四川省艾滋病发病人数进行预测并比较，从而探究四川省艾滋病的流行趋势，为疾病的预防控制提供参考意见。

2. 研究方法与模型建立

2.1. 残差修正 GM(1,1)模型

灰色系统理论是以部分信息已知、部分信息未知的小样本、贫信息的不确定性系统为研究对象。GM(1,1)模型作为灰色预测模型的代表，由于其所用原始数据较少，预测精度较高等优点近年来被广泛地应用于医学卫生领域[5]。

2.1.1. 建立 GM(1,1)模型

GM(1,1)模型建立过程如下：

1) 构建原始数据序列：设原始数据序列为 $y^{(0)}=\left(y^{(0)}(1),y^{(0)}(2),\cdots,y^{(0)}(n)\right)$，对原始数据做一次累加生成处理，生成的序列为：$y^{(1)}=\left(y^{(1)}(1),y^{(1)}(2),\cdots,y^{(1)}(n)\right)$，其中，$y^{(1)}(k)=\sum_{i=1}^{k}y^{(0)}(i)$，$k=1,2,\cdots,n$；

2) 均值生成：对一次累加生成值 $y^{(1)}(k)$ 进行均值处理，得到数据序列：

$z^{(1)}=\left(z^{(1)}(2),z^{(1)}(3),\cdots,z^{(1)}(n)\right)$，其中 $z^{(1)}(k)=\alpha y^{(1)}(k)+(1-\alpha)y^{(1)}(k-1)$，$k=2,3,\cdots,n$，$0\le\alpha\le1$，通常可取 $\alpha=0.5$；

3) 建立白化微分方程：建立灰色 GM(1,1)白化形式的微分方程：

$$\frac{\mathrm{d}y^{(1)}}{\mathrm{d}t}+ay^{(1)}=b$$

其中 a 表示发展系数，b 表示灰色作用量。利用 $y^{(0)}$、$y^{(1)}$ 和 $z^{(1)}$ 分别建立数据矩阵 B 和数据向量 M_n：

$$B=\begin{bmatrix}-\frac{1}{2}z^{(1)}(2) & 1\\ -\frac{1}{2}z^{(1)}(3) & 1\\ \vdots & \vdots\\ -\frac{1}{2}z^{(1)}(n) & 1\end{bmatrix},\quad M_n=\begin{bmatrix}y^{(0)}(2)\\ y^{(0)}(3)\\ \vdots\\ y^{(0)}(n)\end{bmatrix}$$

求解 a 和 b：

$$\begin{bmatrix}a\\ b\end{bmatrix}=\left(B^TB\right)^{-1}B^TM_n$$

4) 得出预测序列：$\hat{y}^{(1)}=\left(\hat{y}^{(1)}(1),\hat{y}^{(1)}(2),\cdots,\hat{y}^{(1)}(n)\right)$，其中：

$$\hat{y}^{(1)}(k)=\left[y^{(0)}(1)-\frac{b}{a}\right]\mathrm{e}^{-a(k-1)}+\frac{b}{a},\quad k=2,3,\cdots,n;$$

将预测模型得到的序列作递减还原，得到估计的原始序列，其中：

$$\hat{y}^{(0)}(k)=\hat{y}^{(1)}(k)-\hat{y}^{(1)}(k-1)=\left[y^{(0)}(1)-\frac{b}{a}\right]\mathrm{e}^{-a(k-1)}\left(1-\mathrm{e}^{a}\right),\quad k=2,3,4,\cdots,n.$$

以及 $\hat{y}^{(0)}(1)=y^{(0)}(1)$。

2.1.2. 检验精度

选择后验差检验法来确定预测序列 $\hat{y}^{(0)}$ 与原始序列 $y^{(0)}$ 的拟合精度。检验方式的标准如表 1 所示[6]，其中方差比 $C=\frac{S_e}{S_x}$（S_e 为残差数列标准差，S_x 为原始数列标准差）：

Table 1. Model accuracy comparison table
表 1. 模型精度对照表

模型精度等级	方差比 C	小概率误差 P
1 级(好)	$C\le0.35$	$P>0.95$
2 级(合格)	$0.35<C\le0.5$	$0.80<P\le0.95$
3 级(勉强)	$0.5<C\le0.65$	$0.70<P\le0.80$
4 级(不及格)	$C>0.65$	$P\le0.70$

2.1.3. GM(1,1)模型的修正

当数据的模型精度等级为 2~4 级时，认为 GM(1,1)模型的预测效果不好，这时选择对模型进行修正。在此可以对原始数据序列与拟合数据序列计算得到的残差序列也使用灰色 GM(1,1)模型进行拟合预测，最后再将得到的残差预测数据 $\hat{\varepsilon}^{(0)}(k)$ 加到已经算出的预测值 $\hat{y}^{(0)}(k)$，$k=1,2,\cdots,n$ 上，用这样的方法来修正原灰色 GM(1,1)模型。同样残差序列使用 2.1.1 中的方法来进行拟合，得到残差拟合值：

$$\hat{\varepsilon}^{(0)}(k)=\hat{\varepsilon}^{(1)}(k)-\hat{\varepsilon}^{(1)}(k-1)=\left[\varepsilon^{(0)}(1)-\frac{b}{a}\right]\mathrm{e}^{-a(k-1)}\left(1-\mathrm{e}^{a}\right),k=2,3,4,\cdots,n$$

此时修正后的预测数据序列可表示为 $\hat{y}^{(0)}=\left(\hat{y}^{(0)}(1),\hat{y}^{(0)}(1),\cdots,\hat{y}^{(0)}(n)\right)$，其中 $\hat{y}^{(0)}(k)=\hat{y}^{(0)}(k)+\hat{\varepsilon}^{(0)}(k)$，$k=1,2,\cdots,n$。

对残差修正后的预测数据进行精度检验，当模型的精度达到合格时，GM(1,1)模型残差修正停止。

2.2. BP 神经网络算法

BP 神经网络(back-propagation neural network)是一种多层前馈的神经网络(见图 1)，该网络的主要特点是信号向前传递，误差反向传递；在前向传递的过程中，需要输入的数据作为输入信号从输入层经过隐藏层逐层处理，到输出层输出；如果与真实输出信号存在误差，网络就转入误差反向传播过程，并根据误差大小来调整各层神经元之间的连接权值和偏倚。当误差达到预期时，网络学习的过程就会结束，之后得到调整后的输出值[7] [8]。

对于隐藏层层数的确定问题，有理论证明：隐藏层节点数越多，网络结构越复杂，网络的训练误差越小；但如果隐藏层节点数过多，虽然网络的续联误差会变小，但是容易出现过度拟合的情况。在研究中一般取隐藏层层数为 1~2，隐藏层节点数为 $\sqrt{m+n}+a$，其中 m 代表输入层节点数，n 代表输出层节点数，a 是 0~10 之间的常数[9]。

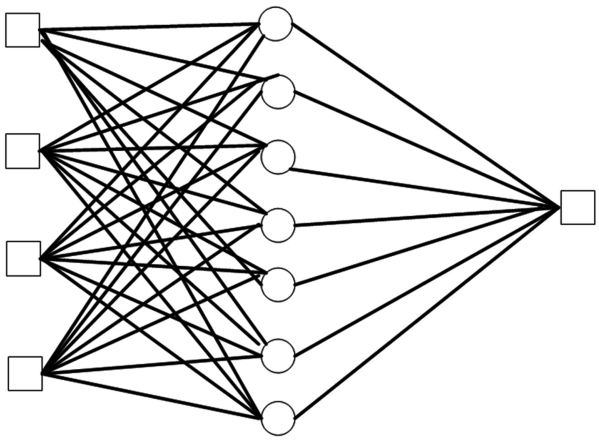

Figure 1. BP neural network structure diagram
图 1. BP 神经网络结构图

2.3. 预测评价指标

本文用平均绝对误差(MAE)、平均绝对百分比误差(MAPE)用于评价模型的预测效果。计算公式如下：

$$\mathrm{MAE}=\frac{1}{n}\sum_{k=1}^{n}\left|\hat{y}^{(0)}(k)-y^{(0)}(k)\right|$$

$$\mathrm{MAPE}=\frac{100\%}{n}\sum_{k=1}^{n}\left|\frac{\hat{y}^{(0)}(k)-y^{(0)}(k)}{y^{(0)}(k)}\right|$$

3. 实证研究分析

3.1. 数据来源

本文收集了 2005~2017 年的四川省艾滋病发病人数月度数据，所有数据来自中国公共卫生科学数据中心(https://www.phsciencedata.cn/Share/)。文中的发病人数由艾滋病感染对象的发病日期按季度累计得到。以 2005 年第 1 季度到 2016 年第 4 季度作为训练样本，2017 年第 1 季度到第 4 季度作为验证样本。

3.2. 残差修正 GM(1,1)模型建立

3.2.1. GM(1,1)模型

将 2005~2016 年的数据按季度分为四个部分，对每个部分进行 GM(1,1)模型的建立，得到每个季度的方差比 C 和小概率误差 P 如表 2 所示：

Table 2. Model-level evaluation of the number of patients in each quarter
表 2. 各季度发病人数模型级评价

	理论模型	C	P	精确等级
第 1 季度	$\hat{y}^{(1)}(t)=648.5982\mathrm{e}^{0.2366(t-1)}-618.5982$	0.276	100.0%	1 级
第 2 季度	$\hat{y}^{(1)}(t)=989.7316\mathrm{e}^{0.2396(t-1)}-959.7316$	0.327	91.67%	1 级
第 3 季度	$\hat{y}^{(1)}(t)=1076.1286\mathrm{e}^{0.252403(t-1)}-1046.1286$	0.460	91.67%	2 级
第 4 季度	$\hat{y}^{(1)}(t)=1027.5545\mathrm{e}^{0.2934(t-1)}-997.5545$	0.761	66.67%	4 级

由上表可以看出，第 1 季度和第 2 季度模型合格；第 3 季度和第 4 季度的模型精确等级不高，模型不合格，此时对这两个季度的模型进行修正。

3.2.2. 修正 GM(1,1)模型

采用 GM(1,1)模型残差修正建模方法，利用模型残差修正对第 3 季度和第 4 季度建模，经过一次修正，模型精度等级变为 1 级。此时得到第 3 季度、第 4 季度的残差拟合模型分别为：

$$\hat{\varepsilon}^{(1)}(t)=-271.2337\mathrm{e}^{0.214669(t-1)}+301.2337$$

$$\hat{\varepsilon}^{(1)}(t)=-206.7147\mathrm{e}^{0.2713(t-1)}+236.7147$$

由上述公式计算出 2005~2016 年第 3、4 季度的残差拟合值并加到 3.2.1 算出的原拟合值上，由此得到的值作为最终的第 3、4 季度四川省艾滋病发病人数的预测值，并与实际数据进行对比，得到四川省 2005~2017 年各季度艾滋病发病人数实际值和预测值的对比图，如图 2 所示：

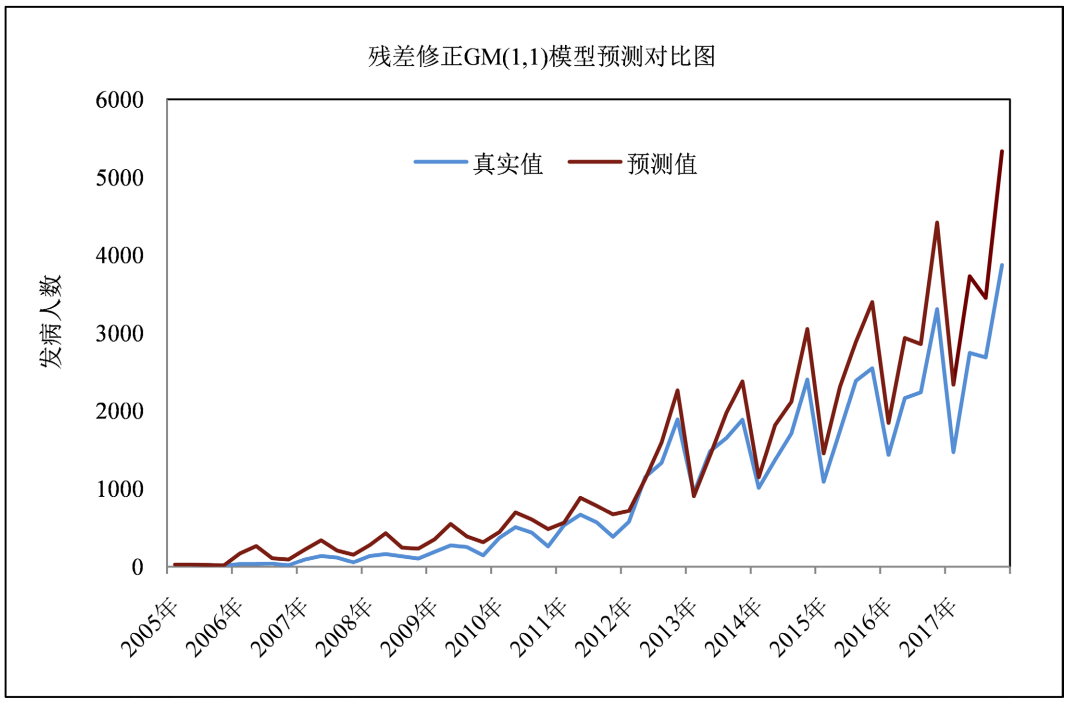

Figure 2. Comparison of GM(1,1) model prediction
图 2. GM(1,1)模型预测对比图

3.3. BP 神经网络模型

由于 BP 神经网络模型采用的是 sigmoid 转换函数，所以输入和输出的数据处理在 0~1 之间(称为数据的归一化处理)。本研究将数据按 $x' = (x - x_{\min}) / (x_{\max} - x_{\min})$ 进行标准化后分析，处理后的数据可得到表 3。

Table 3. Standardization of raw data in each quarter
表 3. 各季度原数据标准化

年份	标准化数据			
	第 1 季度	第 2 季度	第 3 季度	第 4 季度
2005	0.004255	0.003647	0.002736	0.000000
2006	0.005775	0.006383	0.007295	0.001824
2007	0.024012	0.03769	0.031003	0.012766
2008	0.037082	0.045289	0.035866	0.027356
2009	0.053799	0.078116	0.07234	0.039818
2010	0.108815	0.150152	0.128267	0.07538
2011	0.156839	0.197872	0.168997	0.112158
2012	0.171429	0.344985	0.400608	0.569605
2013	0.282979	0.446505	0.497568	0.568693
2014	0.303343	0.411854	0.515805	0.72614
2015	0.326748	0.52462	0.721581	0.769605
2016	0.442857	0.654103	0.676596	1.000000

利用标准化后的数据建立 BP 神经网络模型，将它作为输入数据，即从 2005 年第 1 季度开始到 2016 年第 4 季度结束，以每四季度标准化的数据作为输入，四个季度后的下一季度标准化数据作为输出，如此滚动运行。例以 2005 年第 1、2、3、4 季度标准化数据作为输入，2006 年第 1 季度的标准化数据作为输出等，对神经网络进行训练。

本研究正向传递过程的输入信号长度为 $m=4$，输出信号长度为 1，隐藏层取 1，根据经验公式 $\sqrt{m+n}+1$ 隐藏层节点数根据经验公式取 4~13，通过调节隐藏层节点数，获得了 10 个不同的训练结果。选择其中 MAE、MAPE 最小的训练结果作为最终的 BP 神经网络模型的参数。各节点数对应的 MAE 和 MAPE 如表 4 所示：

Table 4. Evaluation of the corresponding number of nodes in the hidden layer of the BP neural network model
表 4. BP 神经网络模型隐藏层各节点数对应的评价

节点数	4	5	6	7	8	9	10	11	12	13
MAE	2569	2277	1273	1203	2486	1921	1246	578	899	237
MAPE (%)	92.07	77.57	43.92	46.48	86.99	72.29	39.56	18.94	29.47	6.97

由表 4 可以看出，当隐藏层节点数为 13 时，MAE 和 MAPE 最小，分别为 237、6.97%，因此确定隐藏层节点数为 13。建立好 BP 神经网络模型后对数据进行拟合和预测，并与实际发病人数数据进行对比，得到四川省 2005~2017 年各季度艾滋病发病人数实际值和预测值的对比图，如图 3 所示：

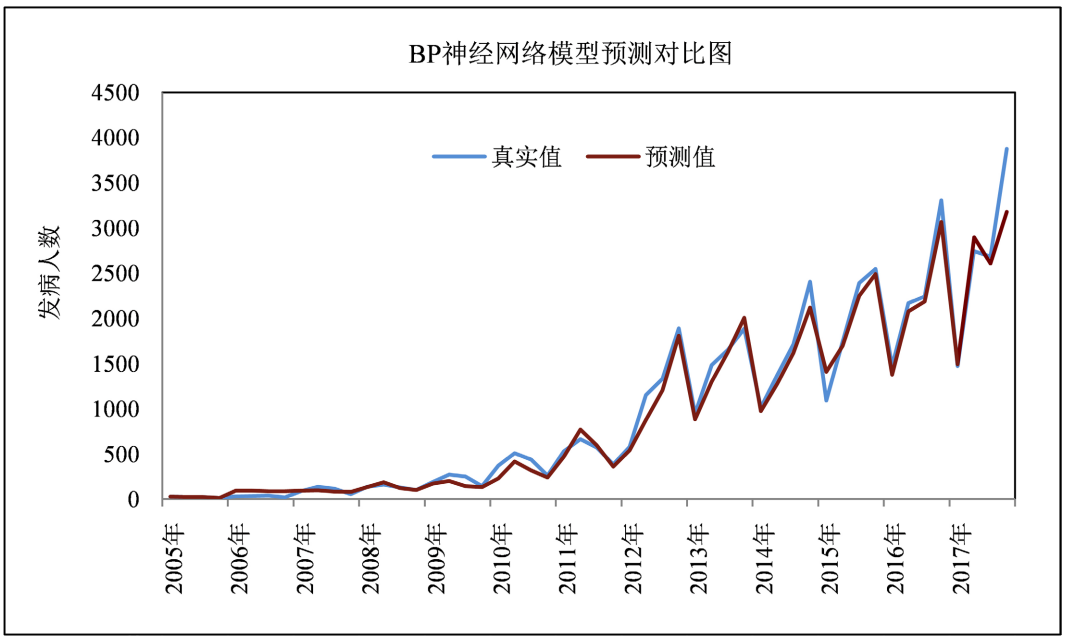

Figure 3. Comparison of BP neural network model prediction
图 3. BP 神经网络模型预测对比图

4. 模型预测结果比较

本文使用 2005 第 1 季度~2016 年第 4 季度作为测试数据，2017 年第 1 季度到第 4 季度作验证数据，预测结果如表 5 所示：

Table 5. Comparison of prediction results between GM(1,1) and BP neural network (2017)
表 5. GM(1,1)与 BP 神经网络的预测结果比较(2017 年)

2007 年	真实值/例	GM(1,1)模型		BP 神经网络模型	
		预测值/例	绝对误差/例	预测值/例	绝对误差/例
第 1 季度	1473	2339	−865	1494	−21
第 2 季度	2744	3732	−988	2897	−153
第 3 季度	2686	3451	−764	2608	78
第 4 季度	3874	5333	−1459	3179	695
MAPE(%)		0.4023		0.0697	
MAE		1019		236	

由表 5 可见，利用残差修正 GM(1,1)模型最终得到 2017 年四川省艾滋病发病人数预测的 MAE、MAPE 分别为 1019、40.23%，利用 BP 神经网络模型最终得到 2017 年四川省艾滋病发病人数预测的 MAE、MAPE 分别为 236、6.97%。通过比较 MAE、MAPE 以及绝对误差这三个评价指标，发现 BP 神经网络模型的预测精度高于残差修正 GM(1,1)模型，在预测四川省艾滋病发病人数方面具有更好效果，并且两种模型的预测结果表明四川省艾滋病发病人数仍然会呈现出上升的趋势。因此有关部门应提前做好应对措施、制定防控方案。

5. 讨论

艾滋病在全球范围内受到极大重视，正确地预测艾滋病的发展趋势有助于艾滋病的控制和防护。

时间序列分析是利用数据序列的时间效应和记忆效应通过建立模型来预测疾病的未来发展趋势，GM(1,1)灰色预测模型属于一种短期的时间预测模型，优点是所需样本少，利用较少的数据预测数据的发展趋势，再对模型得到的残差进行修正，可达到较大的精确度；机器学习算法近年来产生了巨大的发展，其中 BP 神经网络在各界运用较为广泛，技术相对成熟，在数据方面也发挥了极大的作用。

所以，本研究决定使用 GM(1,1)灰色预测模型和 BP 神经网络模型对四川省艾滋病发病人数进行预测并比较，结果证明 BP 神经网络机器学习模型的预测要比残差修正 GM(1,1)模型的更加精确，可以用于艾滋病发病人数的有效预测。但是该模型与其他常用的数学模型一样，主要是从数据来反映疾病的发展趋势，得到的结果是建立在历年数据的基础上，如果数据参数发生了变化，预测的值也相应地会发生改变。另外，在实际生活中，艾滋病的发病因素还有很多未被考虑到本文的模型中，这些因素也会影响模型的预测效果。因此，在制定艾滋病的预防措施时，还应该考虑其他因素对预测结果的影响。

致　谢

在本次论文的撰写中，我得到了我的导师戴教授的精心指导，不管是从开始定方向还是在查资料，查数据准备的过程中，一直都耐心地给予我指导和意见，使我在撰写论文方面都有了较大提高。在此，我对戴老师表示诚挚的感谢以及真心的祝福。

还要感谢的是各位师兄师姐，没有师兄师姐的解惑以及对我的学业的帮助，是没有这篇论文产生的，在论文写作期间，他们不仅给予了我学习上的帮助，创造了学习的氛围，更会在生活上给我提出很有效的建议。再次感谢戴老师和各位师兄师姐的帮助！

参考文献

[1] He, N., Zhang, J., Yao, J., Tian, X., Zhao, G., Jiang, Q. and Detels, R. (2009) Knowledge, Attitudes, and Practices of Voluntary HIV Counseling and Testing among Rural Migrants in Shanghai, China. *AIDS Education & Prevention*, **21**, 560-5812. https://doi.org/10.1521/aeap.2009.21.6.570

[2] 戴色莺, 沈张伟, 范引光, 程晓莉, 叶冬青. 我国艾滋病预防控制中流行病学研究进展[J]. 中华疾病控制杂志, 2015, 19(12): 1282-1285.

[3] 王楚雯, 胡颖, 侯颖. 广西壮族自治区艾滋病模型及预测分析[J]. 检验检疫学刊, 2020, 30(2): 6-9.

[4] 彭志行, 陈峰. HIV/AIDS 传播动力学模型研究进展[J]. 中国卫生统计, 2011, 28(6): 730-734.

[5] 代玉巧, 严运楼, 刘政. 基于 GM(1,1)模型的北京市卫生总费用及构成变化趋势预测分析[J]. 现代预防医学, 2021, 48(11): 1996-2000.

[6] 程燕, 刘如春, 谢红卫. ARIMA 模型与灰色系统 GM(1,1)模型在长沙市艾滋病发病率预测中的效果比较[J]. 职业与健康, 2017, 33(22): 3091-3094.

[7] 杨召, 叶中辉, 赵磊, 薛庆元, 梁淑英, 王重建. ARIMA-BPNN 组合预测模型在流感发病率预测中的应用[J]. 中国卫生统计, 2014, 31(1): 16-18.

[8] 陈涛. 基于 BP 神经网络的艾滋病预测模型[J]. 科学技术与工程, 2007, 7(16): 4176-4178.

[9] 张晓玲. 基于 BP 神经网络的大理州艾滋病流行现状分析和疫情预测模型的研究[D]: [硕士学位论文]. 昆明: 云南大学, 2013.