李　环,孙素芬,罗长寿. 基于 NARX 神经网络的粮食产量预测模型[J]. 江苏农业科学,2020,48(22):228 – 232.
doi:10.15889/j.issn.1002 – 1302.2020.22.043

基于 NARX 神经网络的粮食产量预测模型

李　环[1], 孙素芬[2], 罗长寿[2]
(1.北京农学院,北京 102206; 2.北京市农林科学院农业信息与经济研究所,北京 100097)

摘要:中国是世界上重要的粮食生产大国,保证粮食产量和粮食安全关系到我国国民经济的健康发展。而做好粮食产量的预估工作,对于指导经济的健康发展十分重要。在借鉴相关研究成果和中国统计年鉴的基础上,选择 7 组与粮食产量相关的统计指标,并根据数据与粮食产量作出典型相关性分析,证明 7 组统计量与粮食产量之间的相关性,进而构建一种基于时间序列非线性自回归神经网络的粮食产量预测模型。经过检验发现,该模型的准确率和性能都取得较好的效果,在测试数据集上的平均误差为 1.5%。

关键词:典型相关分析;粮食产量预测;非线性自回归;时间序列
中图分类号: F326.11　　**文献标志码**: A　　**文章编号**:1002 – 1302(2020)22 – 0228 – 05

粮食安全是支撑我国国民经济健康发展的重要力量,也是国家维护社会稳定的重要基石,还是保证国家长治久安的头等大事。其中,人均粮食占有量和粮食自给率是判断粮食安全程度的重要指标[1-2],所以分析我国粮食生产过程中的变化规律以及影响粮食生产的各种因素、掌握合理预测粮食产量的方法具有重要意义。当前国内外学者对粮食产量的预测进行了大量研究,并提出多种粮食产量的预测方法,其中较传统的预测方法主要有基于时间序列的预测方法[3]、滑动平均的预测方法[4]、多元线性回归的预测方法[5]。传统的预测方法虽然具有方法简单、便于操作的优点,但也存在预测值和实际粮食产量相差较大以及模型稳健性差的缺点,可见传统的粮食产量预测模型具有一定的局限性。随着人工智能和机器学习技术的发展,许多学者开始引进机器学习的方法进行粮食产量的预测,其中主要方法有基于支持向量机(SVM)的方法和基于 BP 神经网络的方法。宗宸生等建立一种改进粒子群优化神经网络的粮食产量预测模型,该模型是将普通 BP 神经网络与改进的粒子群算法相结合,并建立 IPSO – BP 模型[6]。该模型使用粒子群算法进行 BP 神经网络初始权重的优化,解决了传统 BP 神经网络模型权重参数优化容易陷入局部最优化的问题,但是该模型存在对粮食产量长期变化预测能力不足的问题。程伟等提出一种基于支持向量机的粮食产量预测方法,选用径向基函数作为核函数[7],并取得比传统多元线性回归、指数平滑模型和灰色模型更好的预测效果。但支持向量机更多的是应用在分类问题上,选择不同核函数对模型精度的影响较大,而如何选取正确的核函数须要进一步研究。因此,本研究提出一种基于时间序列

收稿日期:2020 – 02 – 09
作者简介:李　环(1994—),男,山西大同人,硕士研究生,主要从事农业工程与信息技术研究。E – mail: lh339966@163.com。

结果的置信度还会进一步提高。可以尝试扩增训练集的植物种类,将这一技术进行推广和应用,协助植物资源调查人员进行植物辨识,为中药资源的普查和研究,植物资源研究等各方面工作提供强大的技术支持。

参考文献:

[1]梅星宇,李新华,鲍文霞,等. 基于复频域纹理特征的植物叶片识别算法[J]. 江苏农业学报,2019,35(6):1334 – 1339.
[2]Xiang Z,Zhao W Q,Luo H Z,et al. Plant recognition via leaf shape and margin features[J]. Multimedia Tools and Applications,2019,78(19):27463 – 27489.
[3]于慧伶,麻峻玮,张怡卓. 基于双路卷积神经网络的植物叶片识别模型[J]. 北京林业大学学报,2018,40(12):132 – 137.
[4]张善文,张晴晴,齐国红. 基于 Fourier 描述子和 LBP 相结合的植物叶片识别方法[J]. 江苏农业科学,2019,47(14):273 – 276.
[5]马　娜,李艳文,徐　苗. 基于改进 SVM 算法的植物叶片分类研究[J]. 山西农业大学学报(自然科学版),2018,38(11):33 – 38.
[6]丁常宏,高　鹏. 基于动态时间规整算法的药用植物叶片识别方法[J]. 中医药导报,2019,25(11):63 – 66.

非线性自回归神经网络(nolinear autoregressive models,NARX)的粮食产量预测模型,最终预测结果显示,非线性自回归神经网络结合时间序列模型后的预测效果较好,能够在我国粮食生产预测中提供有效的方式。

1　指标体系的建立

1.1　主要指标数据的准备

构建粮食产量的预测模型首先须要设置模型的指标,本研究数据来源于《中国统计年鉴》(1978—2018 年的数据,数据不包括港澳台地区),选取有效灌溉面积(x_{1n},千 hm^2)、化肥施用量(x_{2n},万 t)、农村用电量(x_{3n},亿 kWh)、农业机械总动力(x_{4n},万 kW)、农业劳动力人数(x_{5n},万人)、粮食作物播种面积(x_{6n},千 hm^2)、受灾面积(x_{7n},千 hm^2)等 7 项数据与粮食总产量(y_n,万 t)进行相关性分析。

在构建模型之前对数据进行预处理,将数据值映射到(-1,1)的区间中,即对数据进行标准化处理,进而可将有量纲数据转变为无量纲数据,用来消除量纲对最后回归结果的影响。本研究使用 min-max 标准化方法使运算后的结果映射到(-1,1)区间中,公式为

$$x_{in}^{*}=2\times\frac{x_{in}-\min(x_{in})}{\max(x_{in})-\min(x_{in})}-1\ 。$$

式中:x_{in}^{*}为各指标 x_{in}的无量纲数值,这样处理后的数据就被映射到(-1,1)区间中,便于之后模型的建立。i 表示指标种类数,$i\in[1,7]$;n 表示 1978—2018 年所对应的年数,$n\in[1,40]$。

1.2　主要指标因素分析

典型相关分析(canonical correlation analysis)是一种使用综合变量的分析方法,通过综合变量之间的相关关系反映 2 组指标之间的整体相关性的多元统计方法,在原来 2 组待研究的变量中提取 2 个具有代表性的变量(w 和 v 的含义为 2 个待考察变量组的线性组合),利用这 2 个变量反映 2 组指标之间的相关性。

简单推导过程如下:设向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$,其中($\boldsymbol{X}\subset\boldsymbol{R}^k$,$\boldsymbol{Y}\subset\boldsymbol{R}^l$),$w=\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}$,$v=\boldsymbol{b}^{\mathrm{T}}\boldsymbol{Y}$。即:

$$w=a_1x_1+a_2x_2+\cdots a_kx_k;$$

$$v=b_1y_1+b_2y_2+\cdots b_ly_l。$$

其中:$a_1\sim a_k$、$b_1\sim b_l$ 是标准化的典型系数。

再使用 相关系数度量 w 和 v 之间的关系

$$\rho_{wv}=\frac{\mathrm{Cov}(w,v)}{\sigma_w\sigma_v}。$$

式中:σ_w、σ_v 分别代表 w 和 v 的方差。

此时,要寻求 1 组 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 的最优解,使得 ρ_{wv}最大化,这样处理后得到的 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 就是使得 w 和 v 有最大相关性的典型系数。这时就可以用 ρ_{wv}代替 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的相关性,从而达到降维的目的。$\boldsymbol{a}$ 和 $\boldsymbol{b}$ 可以用 SAS 软件进行求解。

将 $x_{1n}\sim x_{7n}$作为影响粮食产量的自变量,将粮食产量 y_n 作为另一因变量,然后使用 SAS 软件进行典型相关分析(表 1)。

表 1　原变量的标准化典型系数

变量	典型相关系数
有效灌溉面积(x_{1n})	-0.342 2
化肥施用量(x_{2n})	0.812 5
农村用电量(x_{3n})	0.687 2
农业机械总动力(x_{4n})	0.068 2
农业劳动力人员(x_{5n})	0.354 0
粮食作物播种面积(x_{6n})	0.266 4
受灾面积(x_{7n})	-0.144 4

由表 1 可知,从原有变量提取出来的综合变量为

$$w=-0.342\ 2x_{1n}+0.812\ 5x_{2n}+0.687\ 2x_{3n}+0.068\ 2x_{4n}+0.354\ 0x_{5n}+0.266\ 4x_{6n}-0.144\ 4x_{7n}。$$

由表 2 可以得出原始变量 $x_{1n}\sim x_{7n}$分别与 y_n 的综合变量 v 的相关系数。

表 2　原变量和 v 之间的相关系数

变量	典型系数
有效灌溉面积(x_{1n})	0.944 2
化肥施用量(x_{2n})	0.927 8
农村用电量(x_{3n})	0.927 0
农业机械总动力(x_{4n})	0.454 3
农业劳动力人员(x_{5n})	-0.575 8
粮食作物播种面积(x_{6n})	0.811 4
受灾面积(x_{7n})	-0.590 4

由表 3 可知,原变量组 $x_{1n}\sim x_{7n}$的综合变量 w 和 y_n 的综合变量 v 之间的典型相关系数为 0.983 223 3,具有很强的相关性,所以选出的统计变量可以用来预测粮食产量。

表 3　典型相关分析的结果

统计量指标	数值
典型相关系数	0.983 233
调整典型相关系数	0.981 105
估计标准误	0.005 258
典型相关系数平方	0.966 747

2　模型分析

时间序列是一组按照时间顺序排列的数据，分析时间序列根据时间序列的数据进行曲线拟合和参数估计，是一种定量预测的方法。其基本原理包括：第一，承认事物发展过程中是延续不断的，即应用过去的数据就可以掌握事物的发展规律；第二，考虑事物发展的随机性，任何事物的发展都受到偶然因素的影响，所以用统计学原理对数据进行进一步加工处理。本研究粮食产量的预测模型中所使用的数据是时间序列，所以可以使用时间序列的统计方法构建时间序列模型（图 1）。

非线性自回归模型是一种典型的非线性动态神经网络，NARX 主要由输出层、隐藏层和输出层构成。

NARX 神经网络模型为可以表示为

$$y(t)=f[y(t-1),y(t-2),\cdots,y(t-h),x(t-1),x(t-2),\cdots,x(t-h)]。$$

式中：$y(t)$ 的取值取决于上一个时刻的 $y(t-1)$ 和 $x(t-1)$ 的取值；$f(\cdot)$ 函数是一个非线性函数，且 $f(\cdot)$ 的确立是根据已有数据训练得到的；h 表示模型当中的延时量（图 2）。

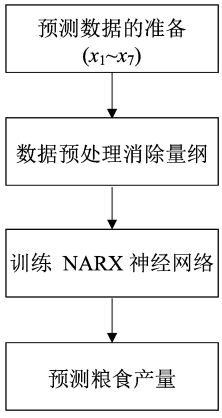


图1　NARX 网络的预测流程

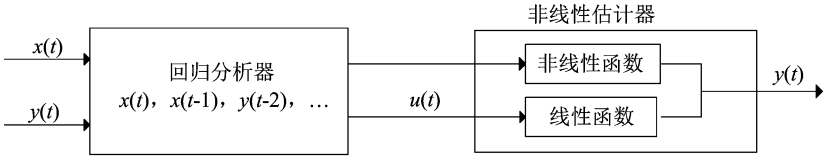
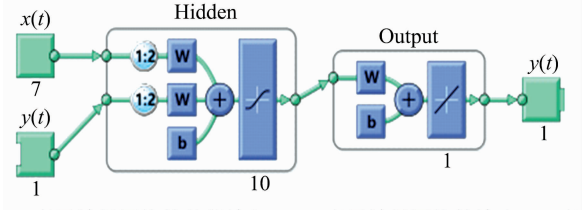


图2　NARX 模型的详细结构

NARX 模型是一个黑箱模型，即 NARX 内部具体的工作流程和运算过程无法清楚地解释，其中各节点的权重也没有明确的含义，但可以通过预测结果评价该模型的性能。

3　预测粮食产量模型的建立

粮食产量预测模型的数据包含 2 个部分，即模型的输入时间序列数据（x_{1n} ~ x_{7n}）、输出时间序列数据（历年的粮食总产量 y），输入延时量为 2。本研究使用 Matlab 进行编程，建立 NARX 预测模型（图 3）。



$x(t)$表示神经网络的外部输入；$y(t)$ 表示神经网络的输出；w 表示联接权值；b 表示阈值，且模型中隐藏层的神经元有 10 个

图3　NARX 粮食产量预测模型结构

由图 3 可知，训练数据占数据集的 70%，验证数据占数据集的 15%，测试数据占数据集的 15%。训练模型时，训练数据、验证数据、测试数据是随机划分的，所以每次训练的结果都有所差异。训练结果见图 4，NARX 神经网络在训练 4 个周期后，在验证集上的误差上升，所以模型的训练可以结束，整个验证集的均方误差（*MSE*）为 0.005 522 6。

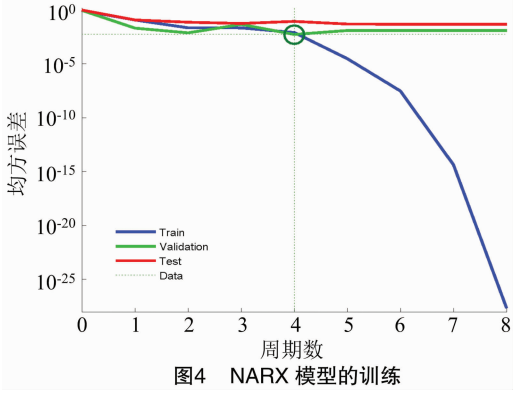
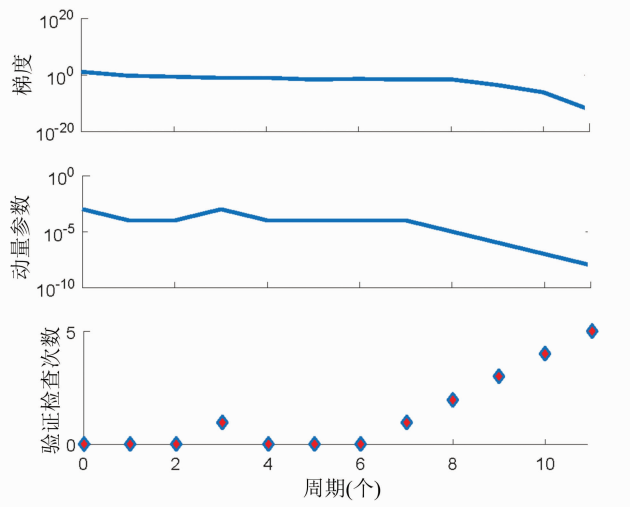


图4　NARX 模型的训练

该模型在训练过程中梯度等参数变化见图 5。粮食产量预测模型的效果可以通过分析误差图、误差自相关图、输入与误差相关图。在观测误差图中，黄色线表示误差线，即表示实际值与预测值之间的误差，误差越小表示模型的预测效果越好（图 6）。误差自相关图（图 8）中，误差在输入延迟（lag）为0时取得最大值，其他lag值处在置信区间为最



11 个周期时梯度为 6.095×10^{-13}、动量参数为 1.0×10^{-8}、验证检查 5 次

图5　NARX 训练过程中的参数变化

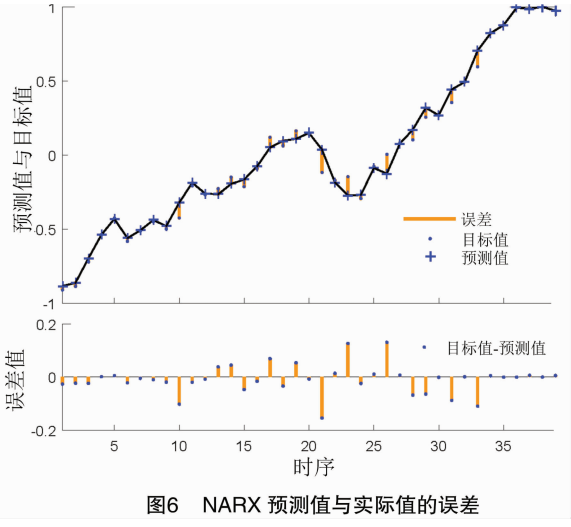


图6　NARX 预测值与实际值的误差

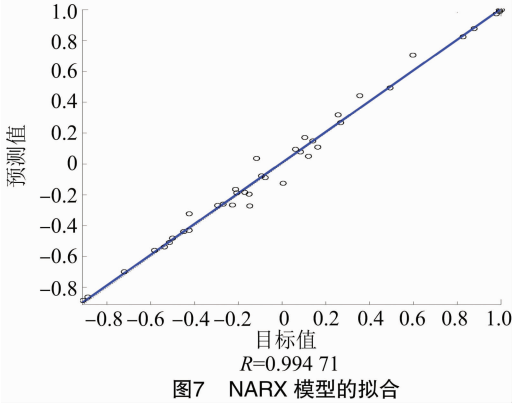


图7　NARX 模型的拟合

佳。在输入与误差相关图(图 9)中,输入与误差的相关系数越接近 0 越好,从图像上分析可知,NARX 模型预测粮食产量的效果较好,预测准确度达到 99.3%,符合预期效果。

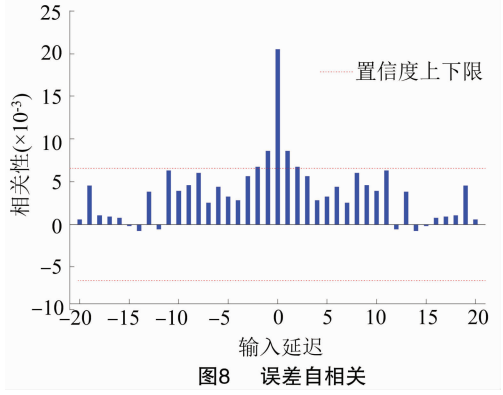


图8　误差自相关

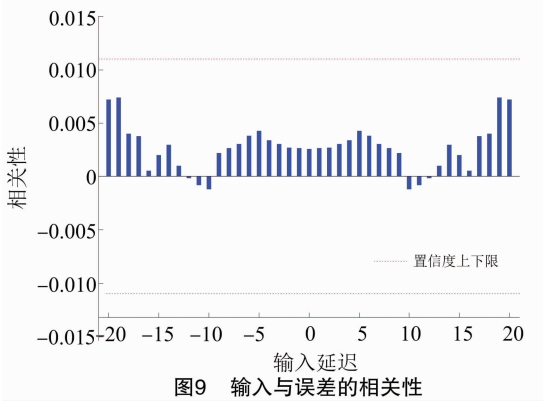


图9　输入与误差的相关性

4　模型的检验

为了进一步验证模型的准确性和精度,用相同数据构建多元线性回归方程建立多元线性回归模型,根据数据拟合得到回归方程

$y_n = 52\,066 - 0.307\,2x_{1n} + 4.291\,7x_{2n} + 2.067\,2x_{3n} + 0.003\,4x_{4n} + 1.212\,5x_{5n} + 0.553\,1x_{6n} - 0.14x_{7n}$。并对 2 个模型进行验证。

由图 10 可知,NARX 粮食产量的预测模型的预测效果较好,与实际值相差不大,且没有出现过拟合和欠拟合的情况,而多元线性回归模型的预测数据与实际值相差较大,预测数据的波动也较大,所以 NARX 模型在粮食产量预测方面较多元线性回归模型的精度更高。由表 5 可知,当使用没有经过训练的数据进行预测时,NARX 模型平均误差为 1.5%,多元线性回归平均误差为 10.88%,相比之下 NARX 模型的精度更高,预测得出的结果也更加接近真实值。

5　小结

本研究从影响粮食产量的因素入手,选取 7 组与粮食产量相关的统计指标,并进行相关性分析,用统计学方法证明指标选取的合理性,接着引入时

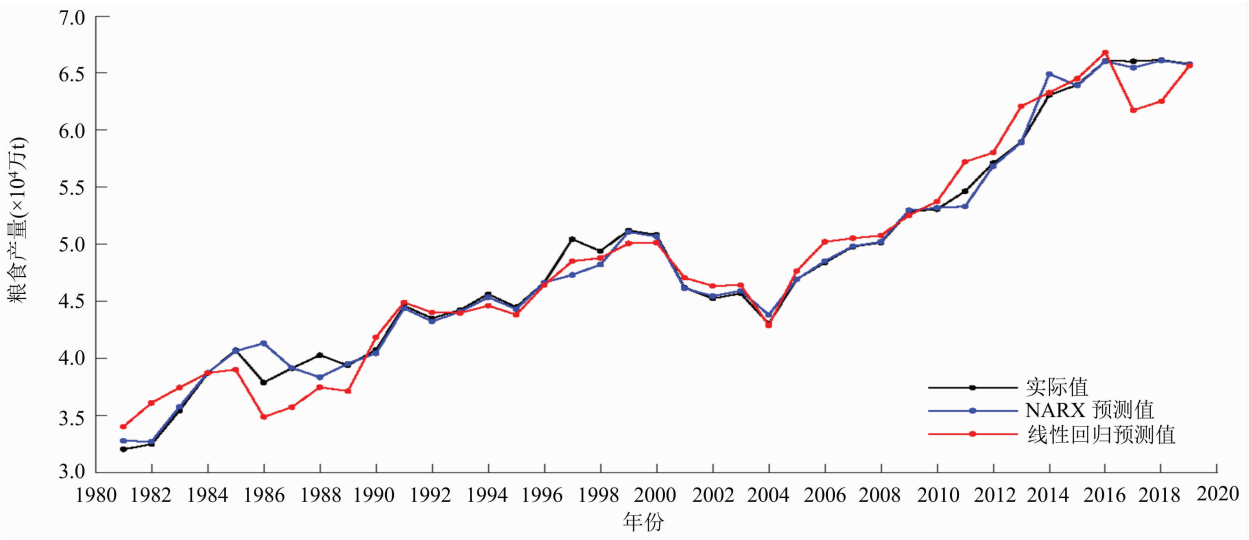


图10　粮食产量预测值与实际值

表 5　粮食产量模型的验证结果

年份	实际产量（万 t）	NARX 神经网络		多元线性回归模型	
		预测产量（万 t）	误差（%）	预测产量（万 t）	误差（%）
1981	32 502	32 738	0.73	37 631	15.78
1987	40 298	38 369	4.70	41 757	3.62
1994	44 510	44 335	0.39	48 622	9.24
2003	43 069	43 832	1.70	50 117	17.13
2011	57 121	56 860	0.46	62 066	8.66

间序列非线性自回归神经网络构建粮食产量的预测模型。在模型检验部分通过与多元线性回归模型进行比较，证明该模型的精度优于多元线性回归模型，且该模型的平均误差仅为 1.5%。总体来说，本研究构建的粮食产量预测模型可以应用在我国粮食生产的预测领域，为国家制定相关政策方针提供帮助。

NARX 模型的优势在于决定当前预测值时要考虑 2 个因素，一是输出时间序列的过去值，二是输入时间序列的当前值，这样可以使 NARX 的预测效果更佳。但是由于 NARX 神经网络中的延时阶数以及隐藏的神经元数，无法用科学的方法得出，只能依靠经验获取，即 NARX 模型依旧是黑盒模型，所以这些问题成为 NARX 模型发展的限制因素，这些因素在今后还有待进一步研究。

本研究为我国粮食产量的预测提供了一种行之有效的方法，该方法不仅可以应用在粮食生产预测方面，在其他领域也有进一步研究的空间。

参考文献：

[1]赵和楠，侯石安. 新中国 70 年粮食安全财政保障政策变迁与取向观察[J]. 改革，2019(11)：15－24.

[2]罗海平，邹　楠，潘柳欣，等. 生态足迹视域下中国粮食主产区粮食生产安全态势的时空属性研究：2007—2025[J]. 江苏农业学报，2019，35(6)：1468－1475.

[3]林彩云. 云南省粮食主产区产量差异时间序列分析[J]. 中国农业资源与区划，2017，38(7)：17－21.

[4]侯云先. 产量预报中滑动平均法的改进[J]. 河南农业大学学报，1994(3)：417－422.

[5]张　煜. 关于农业投入产出的线性回归模型[J]. 农村经济与科技，2019，30(18)：227－228.

[6]宗宸生，郑焕霞，王林山. 改进粒子群优化 BP 神经网络粮食产量预测模型[J]. 计算机系统应用，2018，27(12)：204－209.

[7]程　伟，张燕平，赵　姝. 支持向量机在粮食产量预测中的应用[J]. 安徽农业科学，2009，37(8)：3347－3348.