神经网络和卡尔曼算法在石油需求预测中的应用

孙梅，邓星，董同明
（江苏大学 能源发展与环境保护战略研究中心，江苏 镇江 212013）

摘 要：分析中国1978—2009年影响石油需求的8个相关指标数据。将指标分成3组，通过每组指标的数据分别用广义回归神经网络和误差反向传播神经网络(GRNN和BPNN)方法对2013年的中国石油需求量进行预测，并对其预测结果进行比较。进一步采用神经网络平均影响值(Mean Impact Value，MIV)方法，从影响石油需求时间序列的相关指标数据中筛选出对石油需求影响最大的5个变量。用选出的5个变量，根据AIC准则确定了时间序列的阶数，并建立了石油需求的AR时间序列模型。采用卡尔曼滤波算法和Rauch-Tung-Striebel(RTS)算法对AR模型进行了后验估计。卡尔曼滤波算法使得模型参数得以更新，且相关仿真结果表明，对于AR模型的输出起到较好的修正作用，从而提高了模型的预测精度。

关键词：石油需求预测；神经网络；自回归模型；卡尔曼滤波；RTS算法
中图分类号：TK01+8 **文献标识码**：A **文章编号**：1009-3370(2013)06-0010-06

改革开放以来，中国的石油消费在不断增加，准确地估计石油消费，对国家石油需求和相关市场的操作运行有着至关重要的意义。近年来，不少学者提出了不同的能源预测方法。Yu shiwei(2012)通过数经系数分析研究了中国能源需求的影响因素[1]，并提出了混合算法，粒子群优化和遗传算法的最优能源需求估计(PSO算法，遗传算法EDE)[2]。Tang xu(2010)利用URR模型分析了中国大庆油田的地位和最终可采储量[3]。Toshihide Ito(2010)根据区域能源需求趋势预测了未来能源需求趋势，给出了中国能源需求上限，研究表明实际的能源消费可能超过需求上限但不会高很多[4]。神经网络是近年来新兴的一种数学建模分析方法，已广泛应用于信号处理、数据压缩、故障诊断等众多领域。由于其具有良好的函数逼近能力，神经网络可以用来进行有关数据的预测。Yetis Sazi Murat(2006)使用社会经济和交通运输的相关指标，采用监督神经网络方法预测了交通运输部门的能源需求[5]。孙伟(2009)选用合适的BP神经网络建立了安徽省GDP的神经网络预测模型[6]。吕彬(2009)用人工神经网络对电力负荷进行了预测[7]。刘映(2006)采用时序—神经网络模型建立福建省能源消费预测模型[8]。邢小军(2008)运用主成份分析法辅助协整和误差校正模型建立了中国能源需求的预测模型[9]。Geem Z W(2009)利用人工神经网络分析了韩国的石油需求量问题[10]。王珏采用定性与定量相结合的方式，分析了影响我国能源需求的主要因素，建立了基于小波神经网络的我国能源需求非线性预测模型[11]。张跃军(2013)对2013年国际石油价格进行了分析与预测[12]。

但是，不少学者在采用人工神经网络方法对石油需求时间序列进行预测时很少考虑预测指标的个数会影响预测精度。为了研究不同输入对神经网络结果的影响，本文将影响石油需求的8个指标——人口数量(x_1)、GDP(x_2)、石油进口量(x_3)、石油出口量(x_4)、汽车数量(x_5)、原油价格(x_6)、探明储量(x_7)和炼油能力(x_8)分成3组：第1组(4输入)：x_1~x_4；第2组(5输入)：x_1~x_5；第3组(8输入)：x_1~x_8，对3组数据分别用不同的神经网络方法进行预测，并将其预测结果进行比较。

为了进一步验证研究结果的合理性，通过AIC准则确定阶数后，建立了石油需求自回归AR模型。另外采用卡尔曼(Kalman)算法和Rauch-Tung-Striebel(RTS)算法对建立的模型进行了后验估计，使得在不同置信水平下能够准确地对数据进行估计。

一、基于广义回归和误差反向传播神经网络的石油需求预测

(一)神经网络的石油需求预测

神经网络适用于解决非线性问题，即使在获悉

收稿日期：2013-06-13
基金项目：国家自然科学基金资助项目(71073071；71273119)
作者简介：孙梅(1964—)，女，教授，博士生导师，E-mail：sunm@ujs.edu.cn

数据比较小的情况下，预测结果仍然非常准确。通常情况，在数据进行训练前，数据必须进行归一化，使得数据值的范围在[0,1]之间。如果人工神经网络中使用的数据没有调整到一个适当的范围内，那么网络的训练将不会收敛，同时也不会产生有意义的结果。因此，使用以下的规范化公式

$$\bar{x}_i=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}},\ x_{\max}\neq x_{\min}$$

其中，$x_i(i=1,\cdots,n)$表示输入或者输出数据；$x_{\max}$ 和 $x_{\min}$ 分别代表数据的最大值和最小值。

对于 BP 神经网络，一个至关重要的因素则是隐层神经元的数量。本文利用如下公式计算隐层神经元的数量[13]80-86

$$l=(0.43mn+0.12n^2+2.54m+0.77n+0.35+0.5\sqrt{m+n}\ )^{1/2},\quad l\in N^+$$

其中，m、n 和 l 分别代表了输入层神经元的数量、输出层神经元的数量和隐层神经元的数量。设输入层神经元的数量分别是 4、5 和 8，分别代表 4 输入，5 输入和 8 输入，每个输入向量涵盖了 1978—2009 年的 32 个相关数据。现把数据分为训练组和测试组，利用 1978—2005 年的 28 组数据作为训练组进行训练，误差设为 0.000 1，2006—2009 年的 4 组数据作为测试组用来验证本文预测所得结果的准确性，部分隐层神经元的数量如表 1 所示。

表 1　隐层神经元的数量

类型	GRNN	BPNN
预测（4 输入）	5	3
预测（5 输入）	4	5
预测（8 输入）	5	4

表 2 给出了采用广义回归神经网络(GRNN)和误差反向传播神经网络(BPNN)用 1978—2009 年的关于人口数量、GDP、石油进口量、石油出口量、汽车数量、原油价格、探明储量和炼油能力的相关数据对 2010—2013 年的数据进行预测的结果。

表 2　石油需求量预测

类型	2010 年	2011 年	2012 年	2013 年
GR(4)	367.81	376.81	382.81	398.81
BP(4)	393.07	390.70	397.58	391.31
GR(5)	364.54	376.62	384.66	397.63
BP(5)	430.89	458.56	488.06	517.15
GR(8)	367.63	387.78	412.81	437.78
BP(8)	337.13	352.06	328.49	380.03

注：GR(4)代表 GRNN 四输入，BP(4)代表 BP 四输入，以此类推。

这里 4 输入代表 4 个输入变量，包括人口数量、GDP、石油进口量和石油出口量；5 输入代表 4 输入中增加汽车数量在内的 5 个变量；8 输入代表了以上提到的所有变量。

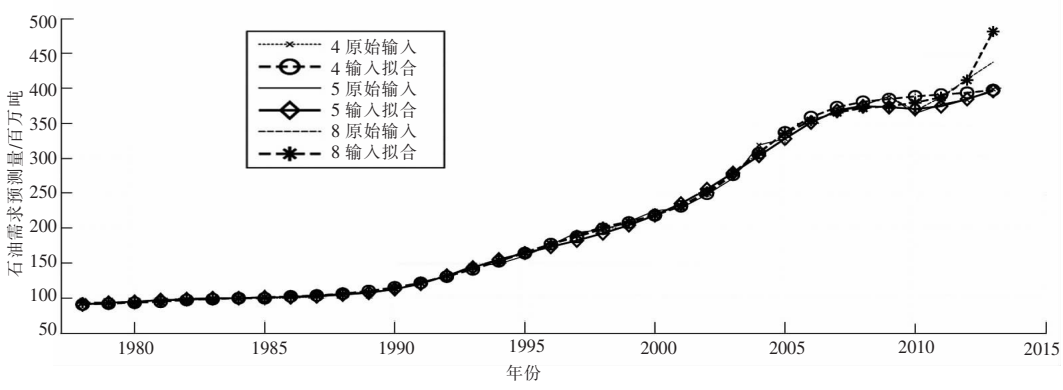

图 1　石油需求预测(GRNN)

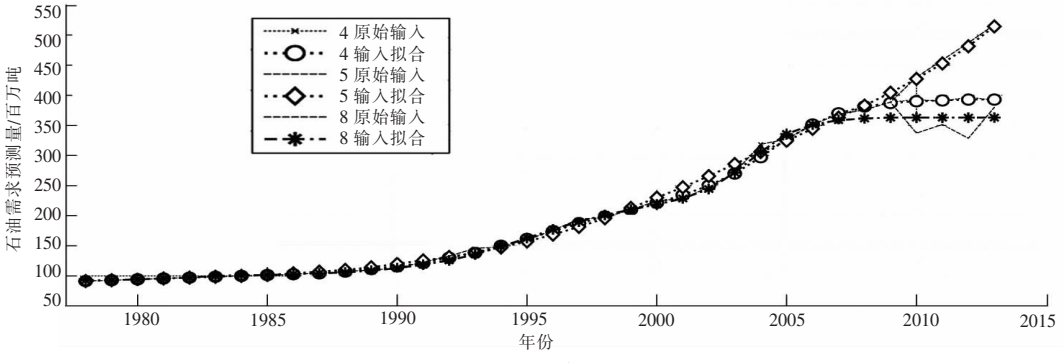

图 2　石油需求预测(BP)

图 1 和图 2 对相关数据进行了拟合。其中图 1 表示的是利用 GRNN 分别对三组数据(4 输入、5 输入和 8 输入) 进行的石油需求预测结果的拟合曲线;图 2 表示的是利用 BPNN 分别对三组数据(4 输入、5 输入和 8 输入)进行的石油需求预测结果的拟合曲线。

(二)相对误差和最佳估计

假设相对误差由如下公式进行描述

$$E_{r,e}=\frac{|V_p-V_t|}{V_t}$$

其中,$E_{r,e}$ 是相对误差;V_p 为由 GRNN 和 BPNN 预测得到的石油需求量;V_t 为真实的石油需求量。其中 2010 年、2011 年的真实数据来至《中国统计年鉴 2012》,本文所用 2012 年的真实数据来自于石油消费的表观数据[14]。将得到的相对误差分别列在表 3 中, 表 3 中的三个误差分别为 2010 年、2011 年, 2012 年的误差。

从表 3 中可以看出 BP 神经网络对 5 输入的预测效果相对较好。

表 3 石油需求预测的相对误差的百分比

类型	2010 年	2011 年	2012 年
GR(4)	0.149 5	0.168 9	0.218 8
BP(4)	0.091 1	0.138 3	0.188 6
GR(5)	0.157 0	0.169 3	0.215 0
BP(5)	0.003 6	0.011 4	0.004 0
GR(8)	0.149 9	0.144 7	0.157 5
BP(8)	0.220 4	0.223 5	0.329 6

注:GR(4)代表 GRNN 4 输入,BP(4)代表 BPNN 4 输入,其余的以此类推。

表 4 回归数值的百分比

类型	4 输入	5 输入	8 输入
BP	94.325	99.168	80.515
GRNN	98.712	76.334	98.646

在表 4 中进一步给出了回归数值的百分比。

综合表 3 和表 4,最好的拟合曲线是 BP 关于 5 输入的预测,说明并不是越多的输入指标带来的预测结果越准确。结合以上的讨论,2013 年最好的预测结果为 517.15 百万吨,最佳数据在图 3 中得以拟合输出。

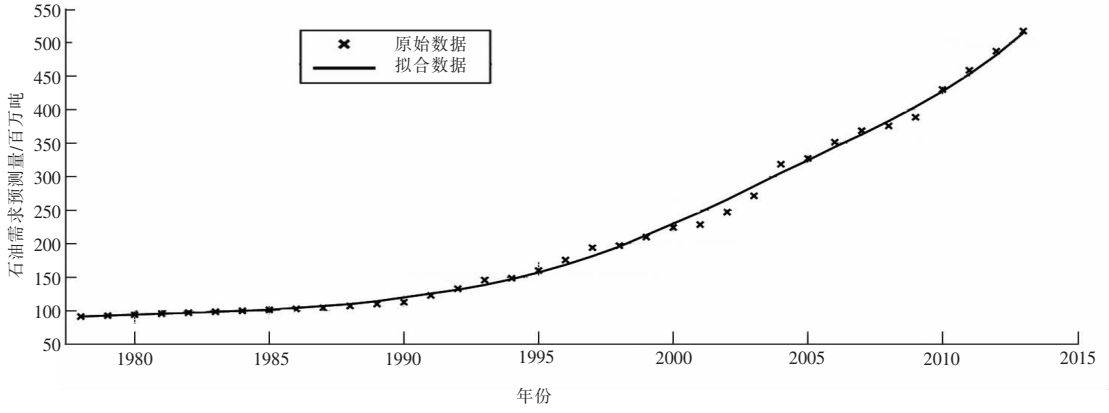

图 3 最佳石油预测估计

二、AR 模型建立与参数估计

(一)数据处理

Dombi 提出用 BP 神经网络的平均影响值(Mean Impact Value,MIV)来反映神经网络中权重矩阵的变化情况,并作为评价各个自变量对于因变量影响的重要性大小指标。前一部分讨论的神经网络中所包含的网络输入变量是研究者根据专业知识和经验预先选择好的。然而许多实际应用中,由于没有清晰的理论依据,神经网络所包含的自变量及网络输入特征难以预先确定。如果把一些不重要的自变量也引入神经网络,会降低模型的精度。因此选择有意义的自变量特征作为网络输入数据常常是应用神经网络分析预测问题中很关键的一步。这里将对以上给出的影响石油需求的相关数据进行重新整理,结合 BP 神经网络平均影响值(MIV)方法进行自变量筛选,并进行时间序列建模和分析[15]。

首先选取了之前提到的影响石油需求的 8 个变量,用神经网络 MIV 方法进行自变量筛选。为了计算出来的 MIV 值有更高的精度,对每组变量的值都进行了归一化处理。考虑到归一化过程中训练的准确性,对 8 组数据分别用式(1)和式(2)进行归一化,处理范围分别是[0,1]和[0.1,0.9]。然后分别求其 MIV 值,经过多次测试后选出最具代表性的两组数据进行对比 (表 5)。训练过程中发现归一化到[0.1,0.9]的训练效果更好,收敛速度也更快。

$$\bar{x}_i=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}},x_{\max}\neq x_{\min},i=1,\cdots,n \qquad (1)$$

$$\bar{x}_i=0.8\left(\frac{x_i-x_{\min}}{x_{\max}-x_{\min}}\right)+0.1, x_{\max}\neq x_{\min}, i=1,\cdots,n \quad (2)$$

其中，$x_{\max}$ 和 $x_{\min}$ 分别代表数据的最大值和最小值。

从表 5 中可以看出影响国内石油需求时间序列的抑制因素为石油出口量(x_4)、原油价格(x_6)、探明储量 (x_7) 和炼油能力 (x_8)。人口数量(x_1)、GDP (x_2)、石油进口量(x_3)和汽车数量(x_5)为影响石油需求的积极因素。对每一组数据的 MIV 值求其均值如表 5 的第 4 行，选出对石油需求时间序列影响最大的 5 个变量：人口数量(x_1)、GDP(x_2)、石油进口量(x_3)、汽车数量(x_5)、炼油能力(x_8)。其中，炼油能力(x_8)为负相关，其他均为正相关。

表 5　BP 神经网络平均影响值(MIV)

归一范围	[0, 1]	[0.1,0.9]	平均值
MIV_x_1	0.009 4	0.011 8	0.106 0
MIV_x_2	0.008 8	0.017 3	0.013 1
MIV_x_3	0.023 4	0.030 7	0.027 1
MIV_x_4	−0.004 8	−0.004 6	−0.004 7
MIV_x_5	0.010 6	0.019 0	0.014 8
MIV_x_6	−0.001 7	−7.886 6e−04	−0.001 3
MIV_x_7	−0.003 9	−0.003 6	−0.003 8
MIV_x_8	−0.018 1	−0.028 7	−0.023 4

(二)模型识别与参数估计

在统计学和时间序列分析中，自回归 AR 模型是一类典型的随机过程，常用作描述和预测各类自然现象。自回归模型是一类典型的线性预测模式，通常是基于之前的时间序列输出作为变量来预测输出准确的时间序列。现代谱估计从方法上大致可分为参数模型谱估计和非参数模型谱估计两种，前者有 AR 模型、MA 模型、ARMA 模型、PRONY 指数模型等；后者有最小方差方法、多分量的 MUSIC 方法等。在此对于时间序列的数学建模具体步骤如图 4 所示。

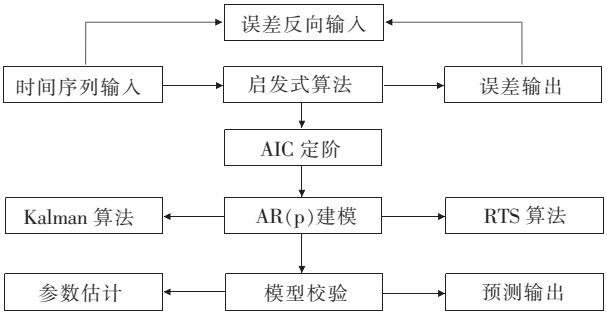

图 4　数学建模流程图

根据神经网络的拟合输出与优化输出效果、训练误差、回归率以及 AIC 准测判定，确定误差最小长度为 p=4。计算得到如下的石油需求时间序列模型

$$\begin{cases} x_t=0.985\ 9x_{t-1}-0.0\ 308x_{t-2}-0.0\ 022x_{t-3}-0.0\ 297x_{t-4}+\varepsilon_t \\ t\in N_4^+=\{t|t\in N, t>4\}, \varepsilon_t\in WN(0,\sigma^2) \end{cases} \quad (3)$$

其中，ε_t 满足零均值白噪声。

考察式(3)的特征多项式

$$A(z)=1-\sum_{j=1}^{p}a_j z^j=1-\sum_{j=1}^{4}a_j z^j, z\in Z$$

经过计算，得到式(3)的根为

$z_{1,2}=-1.8\ 704\pm 2.9\ 040i$

$z_3=2.5\ 676$

$z_4=1.0\ 991$

由于最小相位条件$|z_i|>1, i=1, \cdots, 4$ 在单位圆外，说明该模型下的石油需求时间序列是一个平稳自回归序列。

三、Kalman 滤波与 RTS 算法估计

随着随机过程统计学等理论的发展，时间序列分析方法得到了广泛的应用和发展，其理论也逐渐成熟起来。但是传统的模型建立方法几乎都是着重于单纯从数据自身的相关特性来研究时间序列的各种性质，而忽略了引发数据变化的原因，所以在实际中未得到广泛应用。

近年来许多以状态空间模型为框架的时间序列预测方法应运而生，这种方法首先将时间序列转化为状态空间，然后采用卡尔曼滤波和 Rauch-Tung-Striebel(RTS)平滑算法对非平稳时间序列进行外推预测内插以及平滑，同时还可利用这两种算法对模型的未知参数进行极大似然估计[13]80-86。

基于上面的讨论，利用卡尔曼滤波算法和 Rauch-Tung-Striebel(RTS)平滑算法对建立的石油需求 AR 模型进行估测和递推计算。以自适应的方式追踪石油需求时间序列，并对其进行对比估计。根据式(3)，假定有如下的状态方程和观测方程

$$\begin{cases} \begin{bmatrix} X_t \\ X_{t-1} \\ X_{t-2} \\ X_{t-3} \end{bmatrix}=\begin{bmatrix} 0.985\ 9 & -0.030\ 8 & -0.005\ 5 & -0.029\ 7 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} X_{t-1} \\ X_{t-2} \\ X_{t-3} \\ X_{t-4} \end{bmatrix}+\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}\varepsilon_t \\ \begin{bmatrix} Y_t \\ Y_{t-1} \end{bmatrix}=\begin{bmatrix} 0.895\ 0 & -0.029\ 8 & -0.002\ 2 & -0.029\ 7 \\ 0.998\ 0 & -0.030\ 0 & -0.002\ 0 & -0.029\ 0 \end{bmatrix}\begin{bmatrix} X_{t-1} \\ X_{t-2} \\ X_{t-3} \\ X_{t-4} \end{bmatrix}+\begin{bmatrix} 1 \\ 1 \end{bmatrix}\varepsilon_t \end{cases} \quad (4)$$

对式(4)的真实数值，观测数值以及进行卡尔曼滤波后数值在图 5 中予以显示，其中假设 X 的维数是 4×20，Y 的维数是 2×20，式(4)中状态方程的初始状态分别是 388.2、376.0、369.3、351.2。

从图 5 可以看出，经过滤波后的数值与实际数值比较接近，这显示了卡尔曼滤波算法的有效性。为了显示区间对于数据之间关联的问题，采用不确定的椭圆将其包围。如果样点越少，则椭圆越大，并且样点的快速椭圆拟合能达到 Ricatti 稳定状态的

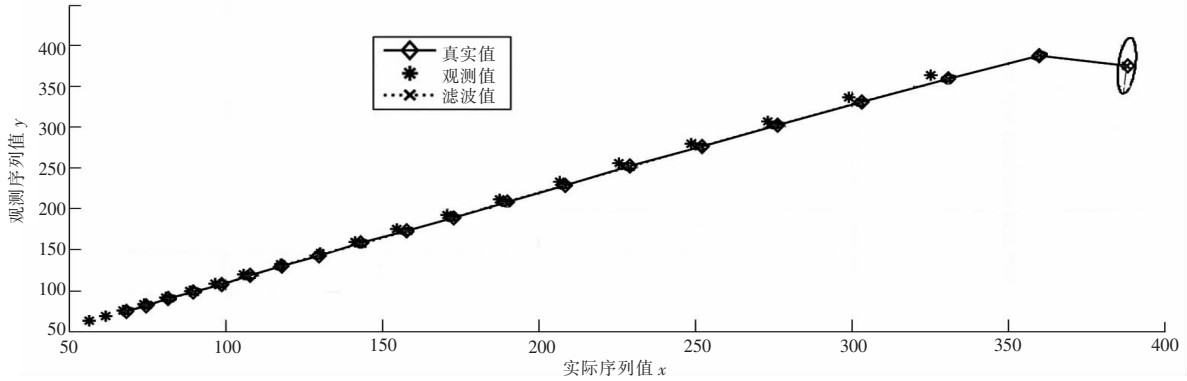

图 5 置信水平为 0.9544 下的真实数值,观测值与 Kalman 滤波图形

值。对于图形中的椭圆形是绘制了一个以序列协方差(x 和 y 分别表示实际序列值和观测序列值)作为椭圆形的长轴和短轴的二元高斯分布。相应的置信水平越高,所对应的置信区间就会越大。

同样地,采用 Rauch-Tung-Striebel(RTS)平滑算法来进行比较,RTS 平滑算法是一种有效的双向区间平滑算法,其通常假设如下的条件[16]:

1.前向:常规的卡尔曼滤波算法。

2.后向:$\hat{X}_{k|n}=\hat{F}_k\hat{X}_{k+1|n}+\hat{K}_k\hat{X}_{k+1|k}$ 这里$\tilde{F}_k=F_k^{-1}(I-Q_kP_{K+1|k}^{-1})$,$\tilde{K}_k=F_k^{-1}Q_kP_{K+1|k}^{-1}$ 对式(3)的真实数值、观测数值以及行 RTS 算法后的数值,如图 6 所示。

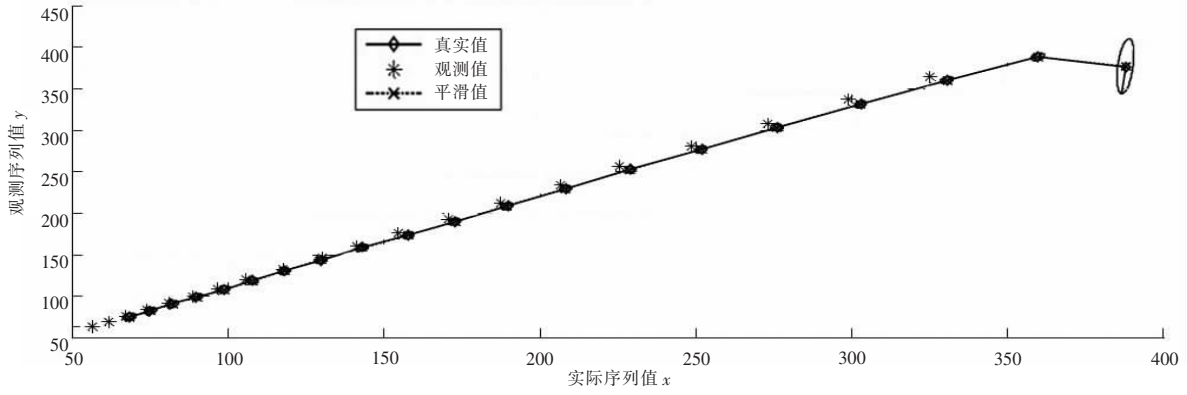

图 6 置信水平为 0.954 4 下的真实数值,观测值与 RTS 平滑算法图形

其中 x 轴和 y 轴表示实际序列和观测序列的值。通过程序计算得到:均方误差拟合估计为 2.134 5,优化估计为1.007 5,显然RTS 优化估计值比均方估计值要小得多。图 5 和图 6 分别显示了利用卡尔曼和 RTS 算法对模型(4)的真实数值,观测数值以及进行 RTS 优化后的数值。在石油需求时间序列中卡尔曼滤波算法和 RTS 算法的效果差别不大,并且不同序列和参数下两种算法又各有优劣。这在一定程度上比较和验证了卡尔曼滤波的正确性。

四、结论

本文采用《中国统计年鉴 2012》[17]和《BP 世界能源统计年鉴 2012》[18]得到的数据,结合经典的人工神经网络智能算法和时间序列预测模型对中国石油需求进行了预测和递推估计。通过改变输入变量的数目,对中国石油需求量进行了预测,结果表明,最好的拟合曲线是 BP 关于 5 输入的预测。利用训练好的 5 输入模型, 预测出 2013 年的石油需求量为 517.15 百万吨。并通过利用卡尔曼滤波和 RST 平滑算法对所建立的时间序列预测模型进行后验估计,同时利用观测方程来追踪设定的状态方程。卡尔曼滤波算法使得模型参数得以更新,从而大大的提高了模型的预测精度。结果表明,卡尔曼滤波算法为石油需求量的预测提供了另一条新的可行的方法,在石油需求预测中应用前景良好。

参考文献:

[1] Yu Shiwei, Zhu Kejun. A hybrid procedure for energy demand forecasting in China [J]. Energy, 2012, 1(37): 396-404.

[2] Yu Shiwei, Wei Yiming, Wang Ke. A PSO-GA optimal model to estimate primary energy demand of China [J]. Energy Policy, 2012(42): 329-340.

[3] Tang Xu, Zhang Baosheng, Mikael Höök, Feng Lianyong. Forecast of oil reserves and production in Daqing oilfield of China [J]. Energy, 2010, 35(7): 3097-3102.

[4] Toshihide Ito, Youqing Chen, Shoichi Ito, Kaoru Yamaguchi. Prospect of the upper limit of the energy demand in China from regional aspects [J]. Energy, 2010, 35(12): 5320-5327.

[5] Yetis Sazi, Murat, Halim Ceylan. Use of artificial neural networks for transport energy demand modeling [J]. Energy policy, 2006, 34(17): 3165-3172.

[6] 孙伟, 李柏年. BP 神经网络在安徽省 GDP 预测中的应用 [J]. 科技和产业, 2009, 9(9): 90-92.

[7] 吕彬, 曾洁. 基于 BP 神经网络的电力负荷预测研究 [J]. 现代商贸工业, 2009(21): 254-255.

[8] 刘映, 张岐山. 能源消费的——时序神经网络预测模型 [J]. 能源与环境, 2006(5): 26-28.

[9] 邢小军, 周德群. 中国能源需求预测函数:主成份辅助的协整分析 [J] . 数理统计与管理, 2008, 27(6): 945-951.

[10] Geem Z W, Roper W E. Energy demand estimation of South Korea using artificial neural network [J]. Energy Policy, 2009, 37(10): 4049-4054.

[11] 王珏, 鲍勤. 基于小波神经网络的中国能源需求预测模型 [J]. 系统科学与数学, 2009, 29(11): 1542-1551.

[12] 张跃军, 王婧, 王姿懿, 张璐, 王小越. 2013 年国际原油价格分析与趋势预测[J]. 北京理工大学学报: 社会科学版, 2013, 15(1): 1-4.

[13] 高大启. 有教师的线性基本函数前向三层神经网络结构研究 [J].计算机学报, 1998, 21(1): 80-86.

[14] 路透社. 中国 2013 年石油表观消费量预计增长 4.8%——中石油 [EB/OL].(2013-01-31) [2013-06-13].http://cn.reuters.com/article/cnInvNews/idCNCNE90TOB820130130.

[15] 史峰, 王小川, 郁磊, 李洋, 张延亮. MATLAB 神经网络 30 个案例分析 [M]. 北京: 北京航空航天大学出版社, 2010.

[16] 邓自立. 卡尔曼滤波与维纳滤波:现代时间序列分析方法[M]. 哈尔滨: 哈尔滨工业出版社, 2001.

[17] Rauch H E, Tung F, Striebel C T. Maximum likelihood estimates of linear dynamic systems [J]. AIAA Journal, 1965, 8(3): 1445-1450.

[18] 中国统计出版社. 中国统计年鉴 2012 [EB/OL]. (2012-09-25) [2013-06-13].http://www.stats.gov.cn/tjsj/ndsj/2012/indexch.htm.

[19] 中国统计出版社. BP 世界能源统计年鉴 2012 [EB/OL]. (2012-06-26) [2013-06-13].http://www.doc88.com/p-975854588099.html.

Application of Neural Networks and Kalman Algorithm in Oil Demand Prediction

SUN Mei, *DENG Xing*, *DONG Tongming*

(Jiangsu University Center for Energy Development and Environmental Protection, Jiangsu University, Zhenjiang Jiangsu 212013, China)

Abstract: This paper analyzes the eight indicators which affected oil demand from the year 1978 to 2009. Setting the indicators into three groups and applying different group data, China's oil demand is estimated in 2013 by the method of generalized regression neural network (GRNN) and back propagation neural network (BPNN). Five indicators which have great influence on oil demand are picked via Mean Impact Value method. Based on the five variables, the order of autoregressive (AR) model is established by employing Akaike information criterion. A posteriori estimation of the AR model is carried out by the Kalman algorithm and Rauch-Tung-Striebel (RTS) algorithm. The results indicate that Kalman filter algorithm could amend the AR model well by updating the parameters and improve the accuracy of the prediction.

Key words: oil demand; neural network; autoregressive model; Kalman; RTS algorithm

[责任编辑：孟青]