基于遗传规划的地下水盐分动态模拟

张传奇[1,2]，温小虎[1]，王 勇[1]，王 德[1]，王 琼[1,2]
（1. 中国科学院烟台海岸带研究所，山东 烟台 264003；2. 中国科学院大学，北京 100049）

摘 要：地下水盐分动态研究对认识地下水盐化规律以及合理规划、利用和管理地下水资源具有重要的意义。本文以地下水盐分动态时间序列数据为基础，利用自回归综合移动平均（ARIMA）模型确定输入向量，建立了地下水盐分动态遗传规划模型。以黄河三角洲为例，使用该遗传规划模型对地下水盐分动态进行模拟。为检验 GP 模型的有效性，与 ARIMA 模型进行对比。结果表明：地下水盐分动态遗传规划模型适于地下水盐分动态模拟研究，而且模拟精度比单纯使用 ARIMA 模型有显著提高。

关键词：遗传规划；地下水盐分；电导率；时间序列
中图分类号：P641.8 **文献标识码**：A **文章编号**：1672-643X(2013)03-0018-05

Dynamic simulation of groundwater salinity based on genetic plan

ZHANG Chuanqi[1,2], **WEN Xiaohu**[1], **WANG Yong**[1], **WANG De**[1], **WANG Qiong**[1,2]
(1. *Yantai Institute of Coastal Zone Research*, *Chinese Academy of Sciences*, *Yantai* 264003, *China*;
2. *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

Abstract: The research on groundwater salinity dynamic plays an important role in understanding the rules of groundwater salinization and the reasonable plan, utilization and management of groundwater resource. A genetic programming (GP) model by using groundwater salinity time series was proposed to model groundwater salinity dynamic. The autoregressive integrated moving average (ARIMA) model was used to determine the input vectors of GP model. The GP model was applied to model groundwater salinity dynamic in the Yellow River Delta. Besides, the GP model was compared with ARIMA model to test the validity. Results indicated that the GP model of groundwater salinity dynamic was suitable for model groundwater salinity dynamic; and the modelling accuracy of GP model is higher than that of ARIMA model.

Key words: genetic programming; groundwater salinity; conductivity; time series

地下水污染研究是当今学科前沿普遍关注的热点[1]。地下水盐化是地下水污染最为普遍的形式，已经成为全球需要迫切解决的环境问题[2]。地下水盐分动态模拟研究有利于认识地下水盐化规律，有利于对地下水资源进行合理的规划、利用和管理。

很多学者采用数值模拟[3-4]、时间序列[5]、灰色模型[6]以及人工神经网络理论[7-8]等方法模拟或预测地下水盐分动态，但这些方法在准确描述地下水盐分动态时仍存在一些问题。数值模拟法虽然能准确反映地下水盐分的运移特征，但建立模型需要详细的水文地质资料以及大量的历史水文数据，时间和经济的消费大；时间序列法应用简单且容易计算，却未充分考虑系统中的非线性关系；灰色模型的微分方程形式是固定的，对复杂动态数据模型的拟合和预测精度较低；人工神经网络理论处理非线性能力强，但无法直观定量输入与输出之间的依赖关系。

遗传规划（Genetic Programming，简称 GP）是在遗传算法的基础上加以延伸而形成的一种新的演化算法，由于与计算机程序直接紧密结合，可以用于实现问题求解程序的优化设计和程序代码自动生成，在建模、预测、控制、信息压缩、人工智能、机器学习、符号处理等众多领域已得到成功应用[9]。相比数值模拟等方法，GP 模型不仅建立简单、处理非线性能力强而且可以显性表达输入输出间非线性关系。本文使用地下水盐分动态时间序列数据，结合自回归综合移动平均模型（ARIMA）确定输入向量，建立地下水盐分动态 GP 模型，并将该模型应用于黄河三角洲地下水盐分动态模拟；此外，为检验 GP 模型的有效性，与 ARIMA 模型进行对比。

收稿日期：2013-01-26； **修回日期**：2013-02-28
基金项目：国家自然科学基金项目（41001013、41001378、41001360）
作者简介：张传奇（1988-），男，山东人，硕士研究生，从事地理信息系统、水环境数值模拟研究。
通讯作者：温小虎（1978-），男，甘肃人，副研究员，硕士生导师，从事环境科学、水资源与水环境研究。

1　遗传规划

遗传规划是 Koza 在遗传算法基础上提出的一种新的启发式随机搜索方法[10]。该方法改变了个体的表达方式,使用树状结构替代了遗传算法的染色体形式。树结构以自变量或常数作为叶节点,以运算符(如 +、-、×、÷、sin、cos、ln 等)作为非叶节点。树状结构如图 1 所示,该树结构代表的表达式是 $y = a + bx$。遗传规划法的寻优结果是与观测值拟合效果最好的自变量与因变量之间的表达式。

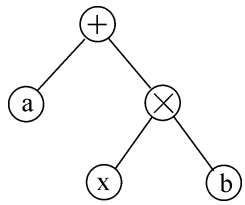

图 1　树状结构示意图

GP 的求解过程:首先利用终止符集和运算符集随机产生一个初始种群,然后根据个体的适应度,按照达尔文的适者生存原则,用遗传算子(复制、交叉和变异)处理高适应度的个体,产生下一代种群,如此循环下去,直到满足终止条件(达到最大进化代数或达到了求解精度),然后输出最优的表达式。

2　ARIMA 模型

ARIMA 模型是统计学家 Box 和 Jenkins 提出的一整套关于时间序列分析预测与控制方法中最重要的基本模型之一[11]。ARIMA 模型把时间序列视为随机过程并用一个数学模型来描述或模拟;当确定该数学模型时,就可以用时间序列的过去值和现在值来预测未来值。

若序列 $\{y_t\}$ 能通过 d 次差分后变成平稳序列 $\{Z_t\}$,则:

$$Z_t = \Delta^d y_t = (1 - B)^d y_t \tag{1}$$

于是可以建立 ARMA(p,q) 模型:

$$\Phi(B) Z_t = \theta(B) \varepsilon_t \tag{2}$$

经 d 阶差分后的 ARMA(p,q) 模型称为 ARIMA(p,d,q) 模型,其一般表示形式为:

$$\Phi(B)(1 - B)^d y_t = \theta(B) \varepsilon_t \tag{3}$$

式中:B 为后移算子,$\Phi(B)$ 为自回归算子,$\theta(B)$ 为移动平均算子,p 为自回归模型的阶数,q 为移动平均模型的阶数,ε_t 为一个白噪声序列。ARIMA 模型的建模流程:首先检验序列的平稳性,若不平稳则通过差分将其转换为平稳序列;然后通过分析样本序列特征初步确定模型的阶数并根据一定的准则确定模型参数。

3　地下水盐分动态 GP 模型

地下水盐分动态数据构成了一个时间序列 $\{x_i\} = \{x_1, x_2, x_3, \cdots, x_n\}$。依据时间序列理论,时间序列的历史数据能揭示现象随时间变化的规律,并且这种规律可以延伸到未来,以进行预测[12]。对 x_{t+1} 进行预测,就是要寻找 x_{t+1} 与前 m 个时刻地下水盐分含量值 $x_1, x_2, \cdots, x_{t-m+1}$ 之间的关系,即 $x_{t+1} = f(x_t, x_{t-1}, \cdots, x_{t-m+1})$,其中 $f(x_t, x_{t-1}, \cdots, x_{t-m+1})$ 是一个非线性函数,表示地下水盐分含量未来值与历史值之间的非线性关系。研究中使用 ARIMA 模型来寻找这些历史值作为地下水盐分动态 GP 模型的输入向量。因此,地下水盐分动态 GP 模型的构建可以分成两部分:第一部分是建立 ARIMA 模型确立输入向量;第二部分是利用确定的输入向量,并设置 GP 的基本控制参数,运行 GP,建立地下水盐分动态模拟的 GP 模型。

3.1　ARIMA 模型部分

地下水盐分动态受多种因素影响,通常属于非平稳时间序列。因此需要对数据进行平稳化处理。在确定处理后的序列平稳之后,通过该平稳序列的自相关函数(ACF)和偏自相关函数(PACF)来确定 ARIMA(p,d,q) 模型的阶数 p 和 q。确定了适当的 p 和 q 后,可以根据 AIC 准则综合确定模型的参数。一般而言,较小的 AIC 值表示滞后阶数是合适的。最后根据 ARIMA 模型确定时滞变量以作为后续 GP 模型的输入向量。

3.2　GP 模型部分

利用 ARIMA 模型确定的输入向量作为终止符集,并根据需要设置合适的运算符号集。考虑到地下水盐分动态的非线性特征,运算符集常包含 +, -, ×, /, sin, cos, ln 等。运行 GP 需要设置的基本控制参数还包括种群大小、进化代数、树的最大树深(或最大节点数)、交叉和变异概率等。

根据确定的终止符集、运算符集以及 GP 的控制参数,运行 GP,建立地下水盐分动态 GP 模型。

4　黄河三角洲地下水盐分动态模拟

4.1　研究区概况

黄河三角洲是我国三大河口三角洲之一,位于山东省北部,渤海西南岸,属于暖温带半湿润季风气候区[13]。多年平均降水量 600 mm,其中 70% 的降

水集中在7、8月份,蒸降比为3∶1[14]。区内地下潜水埋深较浅(平均埋深为1.14 m),盐化现象严重[15]。相关研究显示,该地区只有23.6%的地方地下水矿化度小于2 g/L,主要分布在南部的山前倾斜平原和西部远离大海的区域;其余部分皆为咸水和半咸水;从内陆向滨海潜水矿化度逐渐增高,东南部地区潜水矿化度大于50 g/L,有的地方甚至高达100 g/L[16]。由于该区地下水受海水入侵、黄河侧渗、蒸发浓缩以及人类活动等多重因素影响,地下水盐分变化特征复杂[17]。传统的基于物理过程的数值模拟等方法难以快速准确地描述地下水盐分动态变化。鉴于水中盐分的测定较复杂,而电导率测定简单快速而且能反映盐分的状况,因此,研究中使用地下水电导率表征地下水盐分含量[18]。

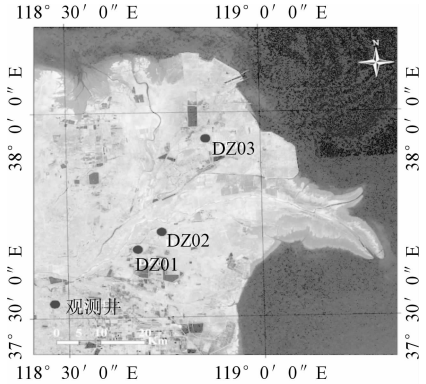

图2 采样点位置分布图

4.2 数据获取与模型构建

4.2.1 *数据获取* 本研究数据来自地球系统科学数据共享网的黄河三角洲2004-2005年野外定点观测数据集[19]。地下水样采集于区内黄河两岸的三个观测井(DZ01、DZ02和DZ03),采样点位置如图2所示。样品采集期为2004年5月5日-2005年12月28日,每间隔5 d采集一次水样。每口观测井共采集了119个水样。其中2004年5月5日-2005年6月16日数据作为训练样本;剩余时间的数据作为测试样本。地下水电导率利用H-BD5W水质测试仪在室内测量。表1给出了地下水电导率数据的基本统计值资料。

表1 地下水电导率数据的基本统计值 μs/cm

观测井	最大值	最小值	平均值	标准差	偏度	变异系数
DZ01	994.7	861.9	921.3	36.0	0.28	0.039
DZ02	946.6	794.5	823.5	21.5	2.28	0.026
DZ03	1022.3	862.8	936.6	39.6	0.52	0.042

4.2.2 *模型构建*

(1)ARIMA模型部分。观测井DZ01、DZ02和DZ03的训练样本时间序列见图3。从图3可以直观看出训练样本序列随时间变化呈现下降的趋势,不是平稳序列。对训练样本序列分别进行一阶差分,差分结果见图4。从图4可以看出差分后序列在零均值上下波动,无明显变化趋势,可以视为平稳时间序列。

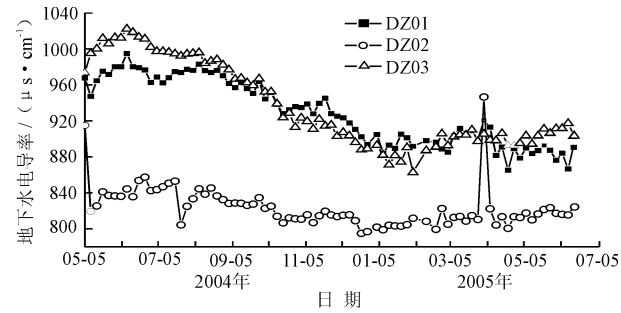

图3 训练样本时序图

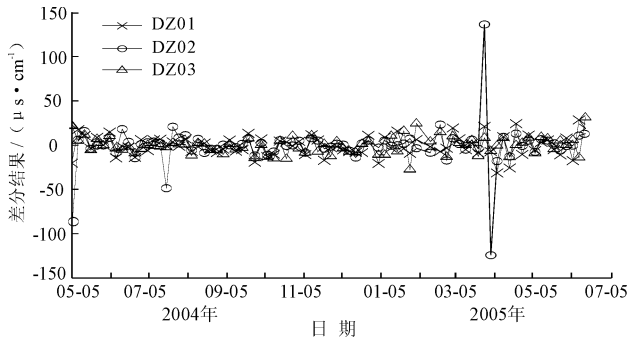

图4 训练样本的差分序列图

根据差分后序列,绘制其自相关函数(ACF)和偏自相关函数(PACF)图(图5)。通过分析图5发现,序列均表现出拖尾特性,符合ARMA(p,q)特征,因此可以建立ARIMA(p,d,q)模型。

通过ACF和PACF图分析可知,对观测井DZ01,显著不为零的PACF个数为7或12,所以p取7或12;显著不为零的ACF个数为1或6,所以q取1或6;因此初建的模型有ARIMA(7,1,1)、ARIMA(7,1,6)、ARIMA(12,1,1)和ARIMA(12,1,6),根据AIC准则,最终取$p=7$,$q=6$。同理,对观测井DZ02,最终取$p=2$,$q=1$;对观测井DZ03,最终取$p=5$,$q=14$。以上分析确定观测井DZ01、DZ02和DZ03的最优模型分别为ARIMA(7,1,6)、ARIMA(2,1,1)和ARIMA(5,1,14)。

(2)GP模型部分。根据确定的ARIMA模型,可得观测井DZ01、DZ02和DZ03的终止符集分别为:$\{x_t,x_{t-1},\cdots,x_{t-7}\}$、$\{x_t,x_{t-1},x_{t-2}\}$和$\{x_t,x_{t-1},\cdots,x_{t-5}\}$。考虑到地下水盐分未来值与时滞变量之间可能存在的非线性关系,运算符集被设置为{+, -,

×,/,ln,sin,cos}。遗传规划法的一些基本的参数设置见表 2。GP 模型的建立和模拟计算是在 Matlab 2010b 平台支持下,利用 GPLAB 2.0 版工具箱[20]实现。经过多次运行,最终确定地下水盐分动态 GP 模型。

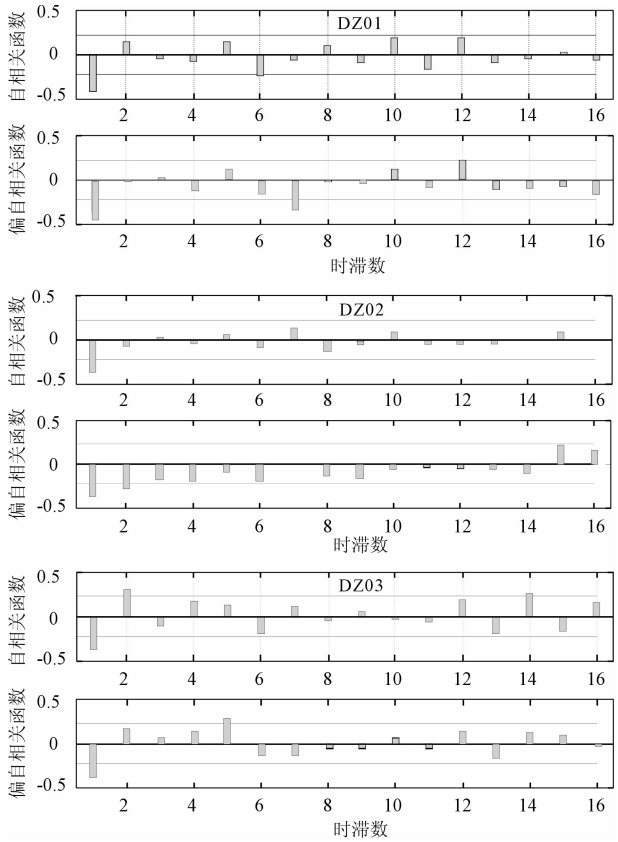

图 5　差分序列的自相关函数(ACF)和偏自相关函数(PACF)图

表 2　遗传规划基本参数设置

种群大小	进化代数	交叉概率	变异概率	树的最大节点数	子代生成选择方法[20]
200	100	0.9	0.1	30	Lexictour

4.3　结果与讨论

地下水盐分动态 GP 模型建立后,需对其精度进行检验;同时选用 ARIMA 模型作为对比。这是因为 ARIMA 模型本身就是一个适于非平稳时间序列预测且预测精度较高的模型[21]。为定量评价模拟效果,模拟结果使用相关系数 r、均方根误差 RMSE 以及平均绝对百分比误差 MAPE 作为评价指标。r 越接近 1,模拟与实测值的变化趋势越一致;RMSE 和 MAPE 越接近 0,模拟与实测的数值大小越接近,精度也越高。它们的计算公式如下:

$$r = \frac{\sum_{i=1}^{n}(x_i^o - \overline{x^o})(x_i^p - \overline{x^p})}{\left(\sqrt{\sum_{i=1}^{n}(x_i^o - x^o)^2}\right)\sqrt{\sum_{i=1}^{n}(x_i^p - \overline{x^p})^2}} \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i^o - x_i^p)^2}{n}} \tag{5}$$

$$MAPE = \frac{1}{n}\sum_{i}^{n}\left|\frac{x_i^p - x_i^o}{x_i^o}\right| \times 100 \tag{6}$$

式中:x^o 和 x^p 分别指地下水电导率的实测值和模拟值;$\overline{x^o}$ 和 $\overline{x^p}$ 分别表示实测和模拟的电导率平均值;n 表示数据个数。

GP 模型和 ARIMA 模型的模拟结果如表 3 所示,结果显示观测井 DZ01、DZ02 和 DZ03 的 GP 模型的 r 分别到达了 0.882、0.896 和 0.821,同时 RMSE(MAPE)分别为 13.3(1.0%),8.3(0.8%)和 11.41(0.9%)μs/cm;ARIMA 模型的 r 分别到达了 0.811、0.783 和 0.635,同时 RMSE(MAPE)分别为 16.9(1.4%),11.8(1.2%)和 11.4(1.9%)μs/cm。以上数据表明 GP 模型和 ARIMA 模型的模拟值的变化趋势都接近实际,而且模拟值与实测值的大小也很接近,反映出两模型都具有较高的模拟精度,适于进行地下水盐分动态模拟。

表 3　ARIMA 和 GP 模型的模拟效果统计参数

μs/cm, %

观测井	模型	r	RMSE	MAPE
DZ01	ARIMA	0.811	16.9	1.4
	GP	0.882	13.3	1.0
DZ02	ARIMA	0.783	11.8	1.2
	GP	0.896	8.3	0.8
DZ03	ARIMA	0.635	21.3	1.9
	GP	0.821	11.4	0.9

对比分析 GP 和 ARIMA 模型,经计算,对观测井 DZ01,GP 模型的 r 提高了 7.1%,RMSE 和 MAPE 分别降低了 21.3% 和 28.6%;对观测井 DZ02,GP 模型的 r 提高了 14.4%,RMSE 和 MAPE 分别降低了 29.7% 和 33.3%;对观测井 DZ03,GP 模型的 r 提高了 29.3%,RMSE 和 MAPE 分别降低了 46.5% 和 52.6%。由计算结果可知 GP 模型比 ARIMA 模型具有更高的模拟精度。此外,图 6 显示了 GP 模型和 ARIMA 模型的模拟值与实测值的时间序列对比,图 6 显示的结果也验证了 GP 模型模拟效果优于 ARIMA 模型的结论。

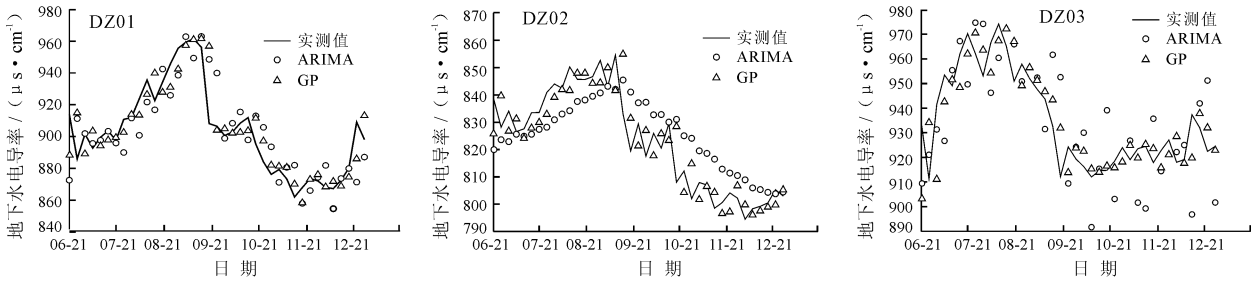

图 6 模拟值与实测值时间序列对比(2005-06-21 – 28)

5 结 语

地下水盐分动态模拟研究有利于认识地下水盐化规律,有利于对地下水资源进行合理的规划、利用和管理。本文根据遗传规划理论,利用地下水盐分时间序列数据,构建了地下水盐分动态 GP 模型。同时,利用自回归综合移动平均模型(ARIMA)确定了该 GP 模型的输入向量。此外,将该 GP 模型应用于黄河三角洲地下水盐分动态模拟,同时,为检验 GP 模型的有效性,与单纯的 ARIMA 模型进行了对比。结果显示地下水盐分动态 GP 模型适用于地下水盐分动态模拟,而且模拟精度高,具有比 ARIMA 模型更好的模拟性能。地下水盐分动态 GP 模型为研究地下水盐分变化规律提供了一种有效的分析工具。但是本文建立的 GP 模型在模拟地下水盐分动态时仍然存在较多不足之处。首先影响地下水盐分动态的环境因素较多,地下水盐分的时间序列数据不能充分反映多重因素对地下水盐分动态的影响;其次,使用 ARIMA 模型确定时滞变量,忽略了系统中与目标值非线性相关较强的时滞变量。因此今后需要更多改进模型的尝试,以便更真实地反映地下水盐分动态变化规律。

参考文献:

[1] 张新钰,辛宝东,王晓红. 我国地下水污染研究进展[J]. 地球与环境,2011,39(3):415 – 422.

[2] Trabelsi R, Zairi M, Dhia H B. Groundwater salinization of the Sfax superficial aquifer, Tunisia[J]. Hydrogeology Journal, 2007,15:1341 – 1355.

[3] Padilla F, Mendez A, Fernandez R, et al. Numerical modeling of surface water & groundwater flows for freshwater & saltwater hydrology: the case of the alluvial coastal aquifer of the low Guadalhorce river, Malaga, Spain[J]. Environmental Geology, 2008,55:215 – 226.

[4] 吴吉春,薛禹群,黄 海,等. 山西柳林泉局部区域溶质运移二维数值模拟[J]. 水利学报,2001,32(8):38 – 42.

[5] Garcia – Diaz J C. Monitoring and forecasting nitrate concentration in the groundwater using statistical process control and time series analysis: a case study[J]. Stochastic Environmental Research and Risk Assessment, 2011, 25(3):331 – 339.

[6] 冯 娟. 开采条件下德州地区地下水水质演化研究[D]. 青岛:中国海洋大学,2011.

[7] 杨劲松, 姚荣江. BP 神经网络在浅层地下水矿化度预测中的应用研究[J]. 中国农村水利水电,2008(3):5 – 8.

[8] Banerjeea P, Singha VS, Chatttopadhyay K, et al. Artificial neural network model as a potential alternative for groundwater salinity forecasting[J]. Journal of Hydrology, 2011,398:212 – 220.

[9] 刘大有,卢奕南,王 飞,等. 遗传程序设计方法综述[J]. 计算机研究与发展,2001,38(2):213 – 222.

[10] Koza J R. Genetic programming: on the programming of computers by means of natural selection[M]. Cambridge: the MIT Press, 1992.

[11] 谭秀娟,郑钦玉. 我国水资源生态足迹分析与预测[J]. 生态学报,2009,29(7):3560 – 3568.

[12] 谷 赫. 时间序列的数据挖掘在证券预测分析中的应用研究[D]. 长春:吉林大学,2005.

[13] 关元秀,刘高焕,刘庆生,等. 黄河三角洲盐碱地遥感调查研究[J]. 遥感学报,2001,5(1): 46 – 52.

[14] 刘高焕,叶庆华,刘庆生,等. 黄河三角洲生态环境动态监测与数字模拟[M]. 北京:科学出版社,2003.

[15] 范晓梅,刘高焕,唐志鹏,等. 黄河三角洲土壤盐渍化影响因素分析[J]. 水土保持学报,2010,24(1):139 – 144.

[16] 关元秀,刘高焕,王劲峰. 基于 GIS 的黄河三角洲盐碱地改良分区[J]. 地理学报,2001,56(2):198 – 205.

[17] 安乐生,赵全升,叶思源,等. 黄河三角洲浅层地下水化学特征及形成作用[J]. 环境科学,2012,33(2):370 – 377.

[18] 蔡阿兴,陈章英,蒋正琦,等. 我国不同盐渍地区盐分含量与电导率的关系[J]. 土壤,1997(1):54 – 57.

[19] 地球系统科学数据共享网项目:黄河三角洲 2004 – 2005 年野外定点观测数据. [DB/OL][2006]. http://www.geodata.cn.

[20] Silva S, Jonas A. GPLAB: A genetic programming toolbox for Matlab[EB/OL]. [2012-01-31]. http://www.doc88.com.

[21] 张 勃,刘秀丽. 基于 ARIMA 模型的生态足迹动态模拟和预测——以甘肃省为例[J]. 生态学报,2011,31(20):6251 – 6260.