基于基因表达式编程的人口预测模型*

刘萌伟，黎 夏，刘 涛
（中山大学地理科学与规划学院，广东 广州 510275）

摘 要：提出一种基于基因表达式编程算法（GEP）的人口预测新方法，并将该方法应用于东莞市人口预测实例问题研究。实验结果表明：由于基因表达式编程算法采用基因型与表现型相统一的编码方式、高效的遗传算子以及全局搜索的寻优方式，基于GEP算法的人口预测模型能够在样本少的情况下给出相对准确的预测结果。其验证数据的预测绝对值平均误差为0.96%，与灰色系统GM（1，1）预测模型及径向基人工神经网络预测模型相比，预测精度分别提高了18.34%、30.54%。GEP人口预测模型能够更好地挖掘人口发展的复杂非线性模式，有效防止过度拟合现象的发生，提供更为准确、合理的拟合及预测结果。

关键词：基因表达式编程；人口预测；时间序列；灰色模型；人工神经网络
中图分类号：TP391 **文献标志码**：A **文章编号**：0529-6579（2010）06-0115-06

A Gene Expression Programming Algorithm for Population Prediction Problems

LIU Mengwei, *LI Xia*, *LIU Tao*
(School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: Predicting the size or development tendency of population is a complicated geographical problem. This kind of problem often involves multiple geographical components that interact in a complex way In this article, a new technique based on a gene expression programming (GEP) algorithm is presented, which can be used to address population prediction problems. In the context of GEP algorithm, population prediction problems are formulated by designing encoding strategies, evolutionary operations and fitness function. The population prediction model based on GEP approach is finally constructed and applied to predict population of Dongguan city. Compared with grey model and artificial neural network model, the predicting precision is improved by 18.34% and 30.54%, respectively. GEP model has better accurateness of predicting the size and development tendency of population. It can accurately fit nonlinear population development tendency and avoid overfitting to a certain extent. Gene expression programming algorithm can be used to effectively solve population prediction problems.

Key words: gene expression programming; population prediction; temporal series; grey model; artificial neural network

人口预测是指根据人口的现状情况及影响人口发展的各种制约因素，来对未来人口规模、水平和趋势进行测算。人口预测研究对于一个国家和地区制定国民经济发展政策、城市发展规划以及资源与环境可持续发展纲要具有重要参考价值。合理准确的人口预测，对于城市建设和发展具有重要的

* 收稿日期：2010-03-31
基金项目：国家自然科学基金重点资助项目（40830532）
作者简介：刘萌伟（1982年生），男，博士研究生；通讯作者：黎夏；E-mail：lixia@mail.sysu.edu.cn

意义。

人口的增长受到自然、经济、政策等多种因素的影响，其系统是一个十分复杂的非线性系统。一般的统计模型难于准确的预测人口的发展水平与趋势。在线性回归人口预测模型之后，灰色模型、马尔萨斯模型、宋健模型以及人工神经网络模型开始应用于人口预测研究[1-6]。但是，虽然马尔萨斯模型及宋健模型具有较好的准确性，但所涉及的预测变量与模型参数较多，增加了决策者应用模型的难度。近来研究较多的神经网络预测模型虽然具有较好的自学习、自适应能力，但是要取得合理可靠的预测结果，因素的选取、隐含层的设计、原始数据的选择等方面都将对预测结果产生极大的影响[5]，同时由于神经网络模型采用局部学习的方式，所以模型容易出现局部最优及过度拟合的现象。

基因表达式编程算法（Gene Expression Programming，GEP）作为新型的进化算法，已经在许多领域开展了相关研究，在人口预测研究领域还未发现相关报道。本文拟以1988－2008年东莞市人口数量作为时间序列，采用GEP算法建立人口预测模型，预测2009－2015年东莞市人口，并与灰色模型以及人工神经网络模型进行比较。

1 基因表达式编程算法

基因表达式编程算法是Candida Ferreira借鉴生物遗传的基因表达规律提出的一种数据挖掘新技术[7]。GEP算法融合并发展了现有的遗传算法（GA）及遗传编程算法（GP），使用线性、定长、树状的基因编码形式，具有极强的函数发现能力和很高的搜索效率。GEP与GA、GP同属于进化算法，但是GEP特有的编码方式以及遗传算子，克服了“GA简单编码只能处理简单问题”及“GP复杂编码导致代码膨胀”的不足，具备解决人口预测问题的能力。

1.1 GEP的基因和染色体结构

GEP算法的根本优势源于其独特的遗传编码方式。其编码方式实现了树状编码（表现型编码）表示实际复杂问题，线性编码（基因型编码）参与简单遗传操作完美统一。其中，表现型编码和基因型编码可以相互转化。在GEP算法中，基因是构成染色体的基本单位，染色体由一个或多个基因通过构成。染色体表示待解决问题的可行解。

GEP的编码环境可以描述为一个包含函数集合与终结符集合的二元组 $GEP = <E, F>$[7]。其中，GEP的基因型编码是一个线性串，由头部（Head）和尾部（Tail）组成。头部可以包含终结符和函数符号，而尾部只能包含终结符。函数符是指用来连接终结符的函数操作符；终结符是指基因表达式编程中程序的输入、常量以及没有参数的函数。头部的长度 h 根据问题的需要而定，尾部的长度 t 是 h 和 n 的函数，其中 n 是函数符的最大操作数，函数关系式如下：$t = h \times (n-1) + 1$。表现型编码是一棵表达式树，其可由基因型编码按照从左到右的顺序逐个读取基因中的字符，并按照语法规则和层次顺序构成。每个染色体可由单个或多个基因构成，每个基因构成一棵子表达式树，而多个子树可由函数符相互连接。举例说明如下：假设一个 $b \times (a/b)$ 的实际问题，考虑函数集 $\{+, -, \times, /\}$ 和终结符集 $\{a, b\}$。如果选定头部长度 $h = 6$，则由头部和尾部的函数关系可得，$t = 7$；所以基因的长度为 $gLength = 13$。假设基因 g 的基因型编码如下：$\times b/abbabbaaab$，则表现型编码方式如图1所示。

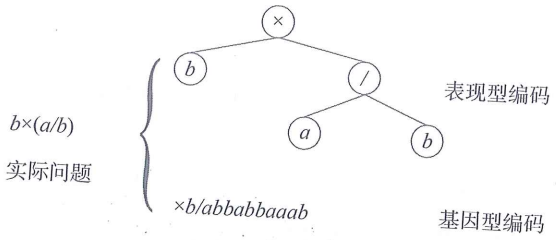

图1 GEP算法基因型编码与表现型编码

Fig. 1 Encodings of genotype and phenotype

从上例可以发现，表达式树的基因长度为5，而实际基因长度为13，这正是GEP的编码特点。基因型编码后部这一未利用的编码区域称为非编码区域，这些非编码区域的存在为程序进化提供了很大的空间。这一区域使得进化过程中可能产生中性变异。很多学者认为，中性变异是保持系统种群多样性的重要因素[8]，所以GEP这种采用了遗传编码形式和个体表现型不同的思路实现了遗传编码的简洁性与表现型的复杂性的统一，克服了GA功能复杂性的不足和GP难以再产生新的变化的缺陷，使得种群在进化的时候不易陷入早熟性收敛。

1.2 GEP算法的遗传算子

在迭代进化过程中，种群中的每个染色体要经过遗传算子的作用，以推进整个种群不断地进化。

1.2.1 选择算子 基本的GEP算法在选择操作上沿用进化算法常用的选择算子，并无特殊性。本文模型选用轮转赌盘算子，根据染色体的适应度值大

小计算当前染色体进入下代种群的概率，并采用精英策略，使得当前种群中最优染色体直接进入下一代种群。

1.2.2　变异算子　与传统的 GA 算法类似，GEP 算法的变异算子作用于单个染色体，可以发生在染色体的各个位置上。其中，发生在基因头部的变异可以选取函数符或终结符作为变异符号；发生在基因尾部的变异只可选取终结符作为变异符号。

1.2.3　重组算子　GEP 算法中的重组算子丰富了传统 GA 算法中的交叉算子，重组算子分为单点重组、两点重组以及基因重组。其中单点重组即从染色体中随机选择一个位置，选取交叉位置到染色体末端的子串作为交叉串，交换两个父代染色体对应的交叉串，构成两个新的子代染色体。双点重组充染色体中随机选择两个交叉位，两个父代染色体交换两个交叉位置之间的子串。基因重组从染色体中随机选择一个基因，两个父代染色体交换这个位置上对应的整个基因。更为丰富的交叉操作保证了进化迭代中的进化效率，可以使得搜索过程快速收敛。

1.2.4　插串算子　插串算子是 GEP 所特有的遗传算子。它随机的在基因中选择一段子串，然后将其插入到其基因的头部位置上，按照是否插入到头部起始位置，可以将插串操作分为插串以及根插串两种方法。其中插串算子是将随机选择的子串插入到除头部起始位置之外的任何位置，插入位置之后的编码向后顺延，超过头部长度的编码将被截去。根插串则是在头部随机选择一个位置，然后在随机位置之后寻找第一个函数符，并以该函数符的位置为起始点选择一段子串，并将子串插入到基因头部的起始位置，头部的编码向后顺延，同样超出头部长度的编码将被截去。如果在随机位置之后没有找到函数符，则不进行任何操作。插串操作类似于生物科学中的隐性基因码激活操作，即将染色体中不活跃的编码段激活并再次组合，这一算子可以在一定程度上避免种群多样性的缺失，可避免进化过程陷入早熟。

2　基于 GEP 的人口预测模型

人口增长呈现明显的非线性趋势，利用 GEP 算法进行人口预测，其基本思想就是利用 GEP 算法强大的空间搜索能力，获取符合人口数据非线性的发展趋势的最佳拟合函数，并据此函数给出未来时期的人口预测值。

2.1　人口预测模型遗传编码

人口预测问题遗传编码环境是一个的二元组 $<F, T>$。其中，终结符结合 T 为原始数据经过相空间重构而得到的时间序列数据的变量集合，$T=\{V1, V2, \cdots\cdots, Vn\}$；函数集合 F 为连接这些终结符的函数符号集，为了能够充分的表示人口增长的非线性趋势，GEP 预测模型选用{ +，-，*，/，Exp，Sqrt，X^2，X^3，Sin，Cos，Tan，Cot，Arcsin，Arctan，Arccsc，tanh，csch，sech}构建函数符集合。头部长度根据问题的复杂性而确定，较短的头部不能完全的表示待求解问题，而较长的头部会引起进化搜索过程的缓慢。针对长期的人口预测非线性的特点，本文的人口预测模型基因头部长度确定为 8。每个染色体由 6 个基因组成，多个基因通过‘+’连接。

2.2　适应度函数构建

适应度函数是问题寻优过程的“导向器”，用来对种群中的染色体进行评估，以确定进入下一代的染色体。适应度函数设计的好坏直接影响到所建立模型的准确度。本文采用基于相对平方根误差的适应度函数，适应度函数如下：

$$f_i = M \times \frac{1}{1 + E_i} \tag{1}$$

式（1）中，f_i 表示染色体 i 的适应度值，M 表示适应度的取值范围，E_i 取值如式（2）所示：

$$E_i = \sqrt{\frac{\sum_{j=1}^{n} (p_{ij} - T_j)^2}{\sum_{j=1}^{n} (T_j - \overline{T})^2}} \tag{2}$$

其中，p_{ij}表示染色体 i 对于样本 j 的预测值，T_j 表示样本 j 的真实值，$\overline{T}$ 取值如下：

$$\overline{T} = \frac{1}{n}\sum_{j=1}^{n} T_j \tag{3}$$

根据公式（1）-（3）可知，当搜寻到较优人口预测函数时，E_i 取值趋近于 0，f_i 取值趋近于 M。相应的适应度的取值范围是 0 - M，本文模型 M 取值为 1000。

2.3　人口预测模型流程

本文基于 GEP 算法的人口预测模型输入为某市历年总人口数据集，模型输出为未来年份总人口预测数据集。模型描述如下：

①根据嵌入维数以及时滞，将原始数据进行相空间重构，生成时间序列数据；②根据二元组 $<F, T>$ 初始化染色体，并形成种群 Population；③利用适应度函数计算种群中的每个染色体的适应度值；④判断种群中的最优染色体是否满足要求，如

果满足则跳至⑦，否则继续向下执行；⑤对当前种群进行轮转赌盘选择操作，并保留种群中最优染色体，使其直接进入下一代种群；⑥随机选择种群中的染色体执行变异、单点重组、双点重组、基因重组、插串、根插串操作，各个遗传算子按照一定的概率进行；执行完毕跳至③；⑦根据获取的最优染色体得到的人口预测函数关系预测未来年份的人口数值。

3 模型应用

3.1 数据处理及模型参数

采用东莞市1988－2008年人口数据进行实证分析和检验，其中1988－2005年数据作为训练数据，2006－2008年数据作为验证数据（见表1）。人口的变化波动受到自然、经济、政策等多种因素的影响，在非线性分析之前，首先需要对数据进行相空间重构[9]，由相空间重构技术确定嵌入维数为8，时间延迟系数为1，从而得到状态空间重构后的时间序列数据，预测模型的终结符集合为 $T = < V1, V2, V3, V4, V5, V6, V7, V8 >$。

根据上述模型定义，并通过Matlab 7.8编程实现GEP人口预测模型，模型详细参数设置如下：进化代数500，种群大小40，基因头部长度8，染色体基因数目6，嵌入维数8，时间延迟1，变异率0.05，单点重组率0.3，两点重组率0.3，基因重组率0.1，插串率0.1，根插串率0.1。

3.2 实验结果与分析

运行构建的GEP人口预测模型，可得到最优染色体的适应度值为936.96，模型的拟合系数为0.996。每棵表达式子树为一个基因，各个子树通过‘+’连接构成染色体。最优染色体各子树如图2所示。

表1 东莞市总人口统计数据[1)]

Table 1 Dongguan demographic data

年份	1988	1989	1990	1991	1992	1993	1994	1994	1996	1997	1998
人口/万人	163.7	177.0	197.4	214.2	250.5	260.6	280.5	285.8	288.6	291.8	347.9
年份	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	
人口/万人	395.6	407.3	611.7	589.8	599.4	648.9	750.6	755.1	729.1	727.4	

1）数据来源：东莞市统计年鉴

将最优染色体转化为函数表达式形式，即可获取第 t 年份东莞市总人口预测函数关系式 X_t：

X_t = subTree1 + subTree2 + subTree3 + subTree4 + subTree5 + subTree6

其中，

$\text{subTree1} = [\text{Sech}(\cos X_{t-1}) + \text{Tan}((X_{t-6} - X_{t-5}) - (X_{t-2} * X_{t-3}))]^2$

$\text{subTree2} = \text{Exp}(\text{Exp}(\text{Sin}(X_{t-8}))) - \text{Tan}((X_{t-5} * X_{t-4}) * (X_{t-2} * X_{t-7}))$

$\text{subTree3} = \text{Tan}((X_{t-1} + X_{t-8}) + (X_{t-5}/X_{t-8}) - (X_{t-2} * X_{t-1}) + \text{Sqrt}(X_{t-2}))$

$\text{subTree4} = \text{Tan}(\text{Tan}(X_{t-4}) + X_{t-1})$

$\text{subTree5} = \text{Tan}((X_{t-3} * X_{t-4})^3 - ((X_{t-5} * X_{t-8}) - \text{Cot}(X_{t-4})))$

$\text{subTree6} = X_{t-1} + \text{Arcsin}(\text{Tanh}(X_{t-3} * X_{t-6}) + (\text{Sech}(X_{t-8}) * X_{t-1}))$

上述函数关系式中，X_{t-1}，X_{t-2}，…，X_{t-8} 分别代表 $V1$，$V2$，…，$V8$ 表示当前年份 t 前 n 年的总人口数据。

近年来，灰色系统模型及人工神经网络模型被广泛应用于人口预测研究。其中，人工神经网络模型凭借其强大的学习能力以及能够从未知模式的复杂系统中发现规律的特点在人口预测领域中取得了较好的研究成果。本文同时利用灰色系统GM（1，1）预测模型及径向基人工神经网络（RBFNN）模型结合东莞市历年人口数据进行了预测实验分析，以期在“灰色理论模型”及“启发式智能模型”两个层次上与GEP人口预测模型进行详细的对比分析，结果如图3所示。

由图3可以看出，本文的GEP人口预测模型的可以较好的拟合东莞市历年人口发展趋势。其中，GEP预测模型2006－2008年的预测人口分别为762万、719万、723万，与实际人口的误差率分别为－0.91%、1.38%，－0.57%，预测验证数据的绝对值平均误差为0.96%，预测人口基本符合东莞市实际人口情况。灰色系统GM(1,1)模型只获取到大致东莞人口整体逐步增多的趋势，对于东莞市人口的波动情况预测能力不足，其拟合程度

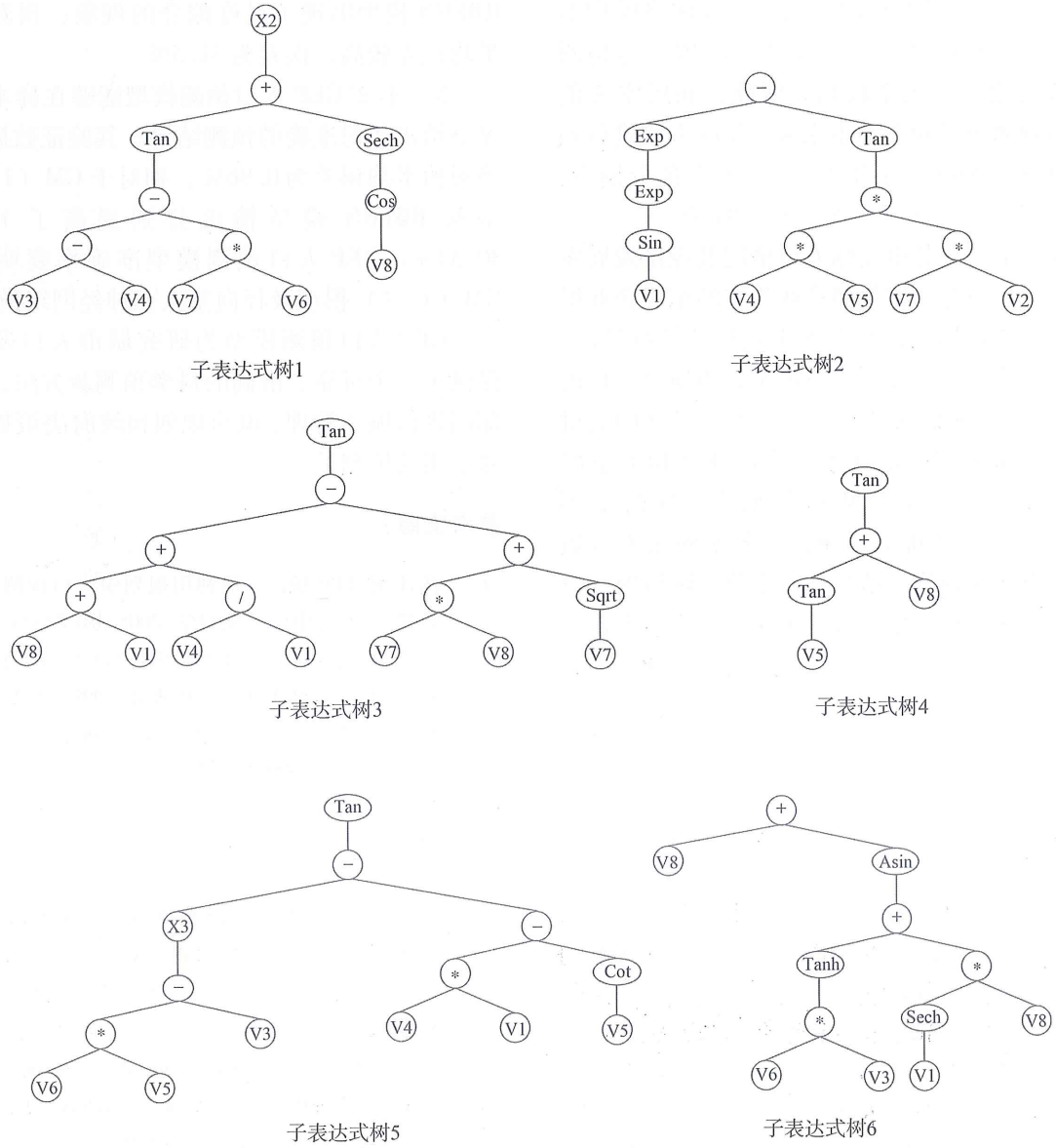

图 2　最优染色体结构

Fig. 2　Structure of optimal solution

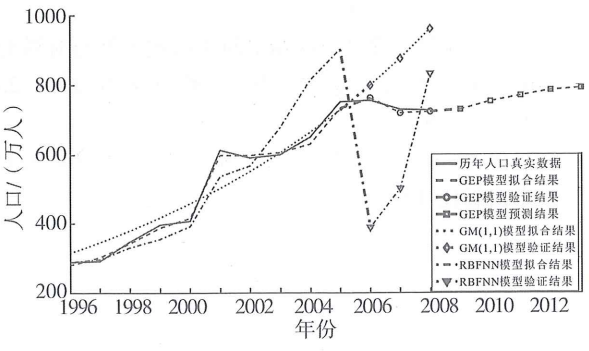

图 3　GEP 人口预测模型实验结果比较图

Fig. 3　Experimental results by using different prediction models

较差。其中，GM(1,1)模型 2006－2008 年预测人口分别为 798 万、876 万、961 万，与实际人口的误差率分别为－5.7%、－20.1%、－32.2%，预测验证数据的绝对值平均误差为 19.3%。RBF 人工神经网络预测模型可以在训练数据上相对较好的拟合人口发展的非线性趋势，但在预测验证数据时出现预测精度低的问题，即出现了过度拟合的现象。造成这种现象的原因在于东莞历年人口训练数据中存在噪音或者训练数据太少，而 RBF 神经网络模型在训练数据少、数据存在噪音的情况下给出较好的预测结果的能力显然不足。RBFNN 模型 2006－2008 年预测人口分别为：388 万、501 万、833 万，与实际人口的误差率分别为 48.6%、31.3%、－14.5%，预测验证数据的绝对值平均误差为 31.5%。对比 GEP 人口预测模型、灰色系统

GM（1，1）模型以及径向基人工神经网络模型不难发现，GEP模型在非线性系统上的拟合与预测能力比较突出，准确率较高。据此，利用本文的GEP人口预测模型对东莞市未来5年的人口进行预测，2009年－2013年东莞市总人口预测值分别为：729万、752万、769万、784万、791万。

此外，结合东莞市实际人口情况及经济发展发展情况可知：以劳动密集型产业为主的东莞企业提供了大量的就业岗位，吸引外来人口的持续迁入。据东莞市统计年鉴，外来人口占东莞当地总人口的2/3以上，外来人口数量是东莞市人口发展的关键因素。近年来，东莞正处于一个产业结构调整时期，知识密集型产业比重逐年增多，同时受到2007年全球经济危机的影响，东莞市外来人口数量增长呈现减缓趋势。结合这一东莞实际情况，分析本文模型的预测结果：未来5年东莞市总人口增速减慢，整体增长趋于平稳的预测结果是比较合理的。

4 结 论

本文首次将基因表达式编程算法应用到人口预测研究中，结合人口预测研究的特点，建立了GEP人口预测模型。GEP算法采用表达式树结构的遗传编码方式、高效的遗传算子以及全局搜索的寻优方式，使之具备较强的非线性空间全局搜索能力，能够从人口时间序列数据中挖掘出较好的拟合函数，从而进行人口预测分析研究。通过东莞市人口预测研究实验可得以下结论：

1）灰色系统GM（1，1）模型能够在本文东莞人口数据样本少、相关信息量不足的情况下给出较为准确的预测结果，其与东莞市2006－2008年实际人口验证数据的平均误差为19.3%，预测精度高于RBFNN模型。但是，其仅仅能够大体反映东莞人口逐步递增的发展趋势，对于人口的非线性发展趋势拟合能力不足。

2）径向基神经网络模型能够较好的拟合东莞人口的非线性发展趋势，但是由于采用局部学习方式，在本文训练样本少、相关信息量不足的情况下RBFNN模型出现了过度拟合的现象，预测绝对值平均误差较高，误差为31.5%。

3）本文GEP人口预测模型能够在样本少的情况下给出相对准确的预测结果，其验证数据的预测绝对值平均误差为0.96%，相对于GM（1，1）模型及RBFNN模型精度分别提高了18.34%、30.54%。GEP人口预测模型准确率要明显优于GM（1，1）模型及径向基人工神经网络模型。

GEP人口预测模型为研究城市人口发展趋势提供了一个可靠、准确的科学预测新方法，为有关部门进行城市管理、城市规划和政府决策提供了有效实用的依据。

参考文献：

[1] 汤江龙，赵小敏. 土地利用规划中人口预测模型的比较研究[J]. 中国土地科学，2005，19(2)：14－20.

[2] 刘兆德，刘西雷. 人口规模预测的GM(1，1)模型应用初探[J]. 资源开放与市场研究，1999，15(1)：25－26.

[3] 杨青生. 基于灰色系统理论的广州市人口预测[J]. 统计与决策，2009，11：49－51.

[4] 王争艳，潘元庆，皇甫光宇，等. 城市规划中的人口预测方法综述[J]. 资源开发与市场，2009，25(3)：237－240.

[5] 赖红松. 基于灰色预测和神经网络的人口预测[J]. 经济地理，2004，24(2)：197－201.

[6] 尹春华，陈雷. 基于BP神经网络人口预测模型的研究与应用[J]. 人口学刊，2005，(2)：44－48.

[7] FERREIRA C. Gene Expression Programming：A New Adaptive Algorithm for Solving Problems [J]. Complex Systems，2001，13(2)：87－129.

[8] SALVATORE A E，KRAMER F R. Multiplex detection of single-nucleotide variations using molecular beacons [J]. Genetic Analysis：Biomolecular Engineering，1999，14：151－156.

[9] 王仲君，高健. 股票市场相空间重构嵌入维与时滞的确定[J]. 重庆工学院学报：自然科学版，2009，23(6)：31－35.

[10] 邓聚龙. 灰色系统基本方法[M]. 武汉：华中理工大学出版社，1987.