人口统计数据空间化的一种方法

廖一兰[1,2], 王劲峰[1], 孟 斌[3], 李新虎[4]
(1. 中国科学院地理科学与资源研究所, 北京 100101; 2. 中国科学院研究生院, 北京 100039;
3. 北京联合大学应用文理学院, 北京 100083; 4. 中国科学院城市环境研究所, 厦门 361003)

摘要: 人口空间分布信息在环境健康风险诊断、自然灾害损失评估和现场抽样调查比较等地理学和相关学科研究中占有重要的地位。目前随着对地观测技术和地理信息科学的飞速发展, 如何精确地进行人口数据空间化成为了研究的难点和热点。针对采用传统方法解决人口空间化问题所遇到的困难和不足, 设计了遗传规划 (genetic programming, GP)、遗传算法 (genetic algorithms, GA) 和 GIS 相结合的方法, 以 GIS 确定量化影响因子权重, 以 GP 建立模型结构, 以 GA 优化模型参数, 成功建立研究区—山西省和顺县的人口数据格网分布表面。实验证明与传统建模方法 (如逐步回归分析模型和重力模型)相比, 所提方法建模过程更为智能化与自动化, 模型结构更为灵活多样, 而且数据拟合精度更高。

关键字: 空间插值; 曲面建模; 遗传规划; 遗传算法; 地理信息系统; 山西省

1 引言

人口增长已经给全球资源、环境承载能力造成了巨大压力: 耕地和林地面积骤减, 生态多样性破坏严重, 人类生存条件日益恶化等。及时获取不同尺度上精确的人口空间分布及其变化信息对于解决这些社会、经济和环境问题, 提高人口、资源和环境的综合管理能力有着重要意义。而人口数据通常是按照行政单元来逐级统计和汇总的。这种统计方法往往造成研究中人口和其他数据所依附的空间单元尺度不同, 使得数据间融合成为难题。另外由于人口的增长和迁移, 还需要大量的精力和财力来维持人口信息的实时性。因此非常必要将人口普查数据进行空间化, 通过建模来模拟人口真实的空间分布状况和动态变迁的过程[1]。近年来人口数据空间化研究发展较快, 成果较多[2-4]。但是这些方法大多需要大量的先验知识来建立人口和影响因子数据之间的数学关系, 并且在建模过程中过多地注重影响因子的选择和量化, 很少顾及被选因子之间的相关性。这样不仅容易造成信息的冗余而且增加了问题的复杂程度。另外人口和影响因子之间的关系因地而异, 对这些方法而言很难用一个统一的模型结构来准确估计不同区域的人口分布。

GP 是一种计算机自动编程技术, 能够发掘数据间潜在关系并以数学方式表达出来[5]。在GP 中, 模型的构建被看作是在由所有可能的计算机程序组成的搜索空间里寻找某一特殊程序, 通过这个程序可以根据给定的输入获取想要的输出[6]。而 GA 则是一种搜索寻优技术, 最早是由 John Holland 和他的同事于 1975 年提出来的[7]。GA 一般是从某一初始群体出发, 遵照一定的操作规则, 不断迭代计算逼近最优解。它的特点是无需提供复杂的数学表达式就能解决十分复杂的优化运算问题[8]。与传统优化方法不同的是, GA 使用定长的线性字符串 (染色体) 为遗传物质, 采用简单编码的方式来在很大的非线性搜索空间

收稿日期: 2007-01-22; 修订日期: 2007-07-02
基金项目: 国家 973 项目 (2001CB5103); 国家 863 项目 (2006AA12Z15); 国家自然科学基金项目 (70571076; 40471111); 中科院知识创新工程 (KZCX2-YW-3-8) [**Foundation:** National "973" Program, No.2001CB5103; National "863" Program, No.2006AA12Z15; National Natural Science Foundation of China, No.70571076; No.40471111; Knowledge Innovation Program of the CAS, No.KZCX2-YW-3-8]
作者简介: 廖一兰 (1980-), 女, 博士, 主要研究方向为地理信息技术与空间分析方法。E-mail: liaoyl@lreis.ac.cn
通讯作者: 王劲峰, E-mail: Wangjf@lreis.ac.cn

中找到十分理想的答案[9]。GA 和 GP 都是遗传进化计算技术之一，它们仿效生物的进化与遗传，根据“生存竞争”和“优胜劣汰”的原则，借助复制、交换、突变等操作，以适应度为目标函数，一代一代地变化，逐步得到最优的结果。它们之间最大的区别是对问题的不同表达方式[10]。GA 的字符串表达方式虽然简单但是不能反映问题的所有性质，尤其是不能提供解决问题的层次结构[11]。而 GP 的表达方式是长度和大小可变的计算机程序树，这种动态树状表示方式灵活、自然，更接近人类解决问题的自然方式，在很大程度上克服了遗传算法的局限性。

因此本文利用进化算法自适应、自组织、自学习特性，设计了将 GP 和 GA 相结合的混合进化建模算法，以 GP 建立模型结构，以 GA 优化模型参数，成功实现了人口数据空间化建模过程的自动化。GIS 为整个建模过程提供了输入数据预处理功能。试验结果证明与常用的逐步回归分析模型和重力模型相比，GP/GA 模型不但结构多样，而且便于解释，拟合精度也更高。

2　人口数据空间化建模方法

自从 1857 年第一张人口密度等值线图产生之后，人口数据空间化研究迅速发展起来。在系统分析了这些方法之后，Deichmann[12]曾将它们分为两大类：面插值和曲面建模。面插值是将人口数据在不同的面域单元内进行转换，目的是将源区的人口普查数据转换到目标区上。而人口分布曲面建模则是利用适当的公式将普查数据分配到一个规则的格网系统中去，系统中的每个格网都包含了一个其特定位置上的人口估算值[13]。这类方法能获取比面域数据更为详尽的人口分布空间信息，而且在分析区域选择时不再受地理分层的限制。人口空间化方法能提供大量人口空间分布和变化信息，但是它们都不可避免地存在一个难题—寻找影响因子和人口数据之间的数学关系。在国内外很多此类研究中是通过建立因子和人口之间线性或非线性回归模型的方法来实现人口空间化的。逐步回归是求取回归模型最为常用的方法之一[14]。逐步回归是一种“有进有出”的计算方法，它按变量的重要性逐一选出重要变量，而且还考虑到已入选回归方程的某些变量，有可能随着其后另一些变量选入而失去原有的重要性，这样的变量就要及时地从回归方程中剔除出去，于是最终的回归方程只保留重要的变量。这里所谓变量的重要性是按它在回归平方和中的贡献多少来衡量的。虽然逐步回归操作简单，结果便于解释，但是要求预先定义模型结构和模型参数，这个往往是很难确定的。

此外，还有其他学科一些成熟的模型被引进用来估算人口，例如物理学上的重力模型就在联合国环境署及国际热带农业中心 (CIAT) 等支持的非洲、亚洲和拉丁美洲的洲际人口数据库建库项目中得以调整应用。人口空间化重力模型建立前提是假设人们都趋于生活在城市及其周边地区，要不就是其生活区域与城市中心有很好的交通连接；即使在农村，人口密集区多依交通干线分布，而且越接近市区，人口密度越要比腹地高。类似于物理重力模型思路，这些关于人口空间分布程式化模式通过衡量各个位置上的平均可接近度来实现，而每个既定位置上的平均接近度是指其到达商场、服务中心此类目的地的方便程度[15]。给定一个距离域值，潜在的人口分布可以表达为[16]：

$$P_{ij} = \sum_{k=1}^{M} \frac{S_k}{(d_{ijk})^a} \tag{1}$$

式中：P_{ij} 为位置 (i, j) 内的人口数，S_k 为目的地 k 的规模，d_{ijk} 为位置 (i, j) 和目的地 k 之间的距离，M 为在给定距离内所涉及到的目的地总数，a 为需要模拟的指数。除了这些目的地以外，事实上人口空间分布还受到其它自然和社会经济因素的影响。因而对于此方法，选择哪些人口分布影响因子变量输入模型同样也是个难题。如果集中分析模型中某

一因子变量，很容易引入偏倚。

3 遗传规划、遗传算法和 GIS 结合解决人口空间化问题

本文的最终目标是结合进化算法 GP、GA 和地理信息来建立一个格网化的人口分布表面。众所周知，人口曲面建模分为三个基本步骤：① 建立一个针对研究区域的规则格网体系，在此基础上生成权重因子分布表面；② 利用辅助数据资料来调整第一步中得到的基本权重；③ 依照前面步骤建立起来的权重比例把研究区域总人口分配到相应的格网中[16]。按照这个思路，本文人口空间化过程分为三个部分：GIS 预处理数据、进化算法建立人口分布模型和依照模型分配人口普查数据。人口分布模型的建立是成功进行人口空间化最为重要的一步，需要首先找到最符合实测数据的模型函数形式，然后寻找满足需求的所有常量和参数。本文所用 GP/GA 方法最大特点是能在 GIS 多维数据中自动便利地找到模型结构和优化答案，无需使用复杂计算。其关键是根据具体问题定义相应的个体和适应度函数，具体步骤如下。

3.1 从 GIS 获取空间数据

GIS 提供人口分布模型所需的基本数据。本文挑选的人口分布影响因子主要涉及到自然和社会经济等方面。相应的数据层被集中输入到 GIS 数据库中，然后 GIS 软件计算出各个因子的原始属性值，对这些值进行归一化处理之后将其作为变量样本值输入到 GP 中去。研究中用于数据处理的 GIS 软件有 ArcGIS 9.0i 和 Geoda 095i。

3.2 基于 GP 的模型结构构建

从数学角度上，人口分布和输入因子变量之间的关系常被表达成：

$$popu(x) \rightarrow f(x_1, x_2, x_3, x_4, x_5, \cdots, x_r) \quad (2)$$

式中：$popu(x)$ 是每个格网内的人口估算值；$x_1, x_2, x_3, x_4, x_5, \cdots, x_r$ 分别是所选的 r 个输入因子变量。利用何种算法准确有效地确定这种关系是各种人口数据空间化方法的难点之一。GP 是一种不依赖于具体问题领域特定知识的机器自动学习的软方法，其基本思想是随机产生一个适合于给定问题环境的初始群体，即问题搜索空间，依据自然选择原则，用遗传算子对初始群体进行相关处理，得到适应度最高的个体组成下一代群体，多次迭代后使问题逐渐逼近最优解[17]。与其它建模方法相比，GP 进化模型是根据环境自动确定的，不需事先确定或限制最终答案的结构或大小。而且在计算过程中输入、中间结果和输出都是问题的自然描述，无需或少需对输入数据的预处理和对输出结果的后处理。最后产生结果也具有层次性，便于理解。鉴于这些特点，本文确定用 GP 来构建人口分布和影响因子间的数学关系式结构。

3.2.1 初始种群的生成　在 GP 中，每个可能的人口分布模型结构 (个体) 都以二叉树结构来显示，树上的终端叶子结点表示一个终止符 (随机常量和输入的影响因子变量)，其余的结点则表示一个函数 (操作符 +, -, , /, exp, ln)。初始群体采取混合法的方式产生，就是初始群体中每个个体深度在 2 至给定的最大深度 (叶子深度是指叶子距树根的层数) 之间均匀选择，每一种深度下的初始个体数所占百分比 p 为：

$$p = 100/(\max\_depth - 2 + 1) \quad (3)$$

max_*depth* 是指给定的最大深度。在同一深度中，50%的初始个体用完全法产生，另50%的初始个体用生长法产生[10]。当待定结点的深度小于给定的最大深度时，完全法中该结点仅在函数集内产生而生长法则可以在函数集与终止集的并集内选择；当待定结点的深度等于给定的最大深度时，在这两种方法里该结点都仅在终止符集内产生。

3.2.2 适应度评价　适应度评价是影响 GP 运行效率的重要因素，并且随问题的不同而变化。由于研究中未能获取到每个格网内真实的人口数，所以本文只能通过在普查单元

里汇总用个体间不同符号表达式所求得的格网人口估算值，求取其相关系数 (R^2) 来评价适应度优劣的。相应的适应度函数定义为：

$$F = \frac{\sum_{j=1}^{N} (P(j) - \overline{P})\,(P'(j) - \overline{P'})}{\sqrt{\sum_{j=1}^{N} (P(j) - \overline{P})^2 \sum_{j=1}^{N} (P'(j) - \overline{P})^2}} \tag{4}$$

式中：N 是普查单元个数，$P'(j)$ 和 $P(j)$ 分别是普查单元 j 的估算和实际人口值，而 $\overline{P'}$ 和 $\overline{P}$ 则分别是研究区域所有普查单元的人口估算和实际人口平均值。$P'(j)$ 通过下式可以获得：

$$P'(j) = \sum_{i=1}^{n} popu(i, j) \tag{5}$$

式中：$popu(i, j)$ 是指普查单元 j 内格网 i 的人口估算值，n 代表普查单元所包含的格网数。显然适应度越大，个体性能越好。

3.2.3 遗传操作　GP 能够不断地在以适应度为基础选择出来的个体上产生新个体，主要是通过执行复制、交叉和变异这些遗传算子来实现的。复制操作是从当代群体中选择优良个体使之自我复制繁衍的过程，体现了“适者生存”的自然选择原则。本文 GP 采取的是竞争选择策略，即随机从群体中选取一组个体，比较该组每个成员的适应度，选出实际最好的个体进行复制以取代最差的个体。值得注意的是，当代个体是有放回的选取，所以同一个体可能会被多次选中或复制。交换和突变是 GP 里主要的进化手段。其中交叉是仿照生物杂交的原理，用互相交换两个被挑选个体的任意选定部分的办法来产生新种群里子代个体的。本文 GP 交叉就是随机选取两个个体的交叉点，然后相互交换这两个交叉点以下的子树来生成两个新个体。突变则是通过任意变异所选个体的选定部分来实现产生新个体的目的。本文 GP 采取收缩变异来实现个体突变，也就是随机选定父代个体的变异点及其下属分支子树后，删除突变点，再用其下属分支子树来代替它。

3.2.4 终止条件　GP 是一个重复迭代的搜索方法，通过多次进化只能逼近最优解而不是正好达到最优解。为了有效地利用计算资源，本文根据以下两个条件来终止运算过程：

(1) IF　总的遗传代数 g = 规定的遗传代数　THEN　运算停止

(2) IF　R2 ≥ a　THEN　运算停止

式中：a 是[0, 1]之间的一个实数。两条件满足其一，运算便停止。此时得到的最优个体便是待求的研究区域最佳格网人口分布模型结构。

3.3 基于实数 GA 的参数优化

由于 GP 搜索空间过大，不能对计算机程序中某单个结点进行优化，所以模型结构确定后，模型参数优化成为提高人口分布模型精度的关键。应用传统的优化搜索方法，如最小二乘、EM 算法等，进行人口分布模型参数优化计算，很容易陷入局部最优解。而 GA 作为一种仿生算法，通过全面模拟自然选择和遗传规律，形成一种“生成 + 检验”特征的搜索寻优机制，具有全局最优解、智能式搜索、渐进式优化、简单通用性强和优化精度高的特点，恰恰是解决此问题的有效途径[18]。通过对人口数据空间化问题的具体分析，结合遗传算法的基本原理，确定了遗传算法对模型的优化进程：① 通过分析模型最后需要达到的各项要求，建立适应度评价函数，以便于进行结果的评价选择；② 采用实数编码方式，选择合适的群体大小，随机生成初始群体；③ 计算群体中每个个体所对应的评价函数值，根据其值大小，通过优胜劣汰，淘汰适应度差的个体，对幸存的个体根据其适应度的好坏，按概率选择，进行复制、交叉和突变的操作，产生子代。④ 对子代群体重复步骤 (3) 的操作，进行新一轮遗传进化过程，直到找到了最优解。

3.3.1 适应度函数定义　与 GP 相似，本文 GA 也是通过计算普查单元里的人口估算值

和实际值的方差来评判个体优劣。为了避免适应度计算所产生的早熟收敛问题，GA 适应度函数定义为：

$$F\_GA_k = 1 / \left(\frac{S \times \sum_{j=1}^{N} (P'_k(j) - P(j))^2}{\sum_{k=1}^{S}\sum_{j=1}^{N} (P'_k(j) - P(j))^2} + \varphi\right) \quad (6)$$

式中：N 是普查单元个数，S 是种群规模，$P'_k(j)$ 是利用个体 k 估算出来的普查单元 j 人口数，$P(j)$ 则是普查单元 实际人口数，φ 是位于 (0, 1) 的常数。

3.3.2 染色体编码 传统 GA 常常用二进制字符串 (染色体) 来表示问题可能的解决方案。但是对于很多问题而言，二进制编码方式会产生过多的字符串，从而减慢进化过程。并且如果字符串长度不够，GA 只能靠近而不能达到全局最优的目标。现实世界里大多数优化问题所涉及到的参数都为实数，直接在原始实数空间内对其操作比在离散空间中要好[19]。因此在本文中，GA 个体都表现为一个由 GP 进化模型参数组成的 t 维向量 ($\beta_1\beta_2...\beta_t$)，并且所有遗传算子仅对个体基因进行操作。实践证明，这种编码方式是十分便于实现这些操作的。

4 应用及结果分析

4.1 研究区域及空间数据

本研究区域选在山西省和顺县。以模拟和顺县人口分布为例，检验所提出的方法在人口数据空间化建模方面的有效性，并将该方法与逐步回归分析法和重力模型法进行对比。和顺县位于山西省东部，太行山中段，地理坐标是东经113°05'-113°56'，北纬 37°03'-37°36'。现辖 326 个行政村，总面积 2250 km^2。2001 年全县总人口将近 13.4 万人，人口密度达到 60 人 /km^2。和顺地形中高周低，多中低山地，总体呈西南—东北走向。境内有 15 条较大河流，分属黄河、海河两大流域。交通较为便利，有榆(次)—邢(台)公路纵横全县东西，平(定)—黎(城)公路、阳(泉)—涉(县)铁路贯通南北。

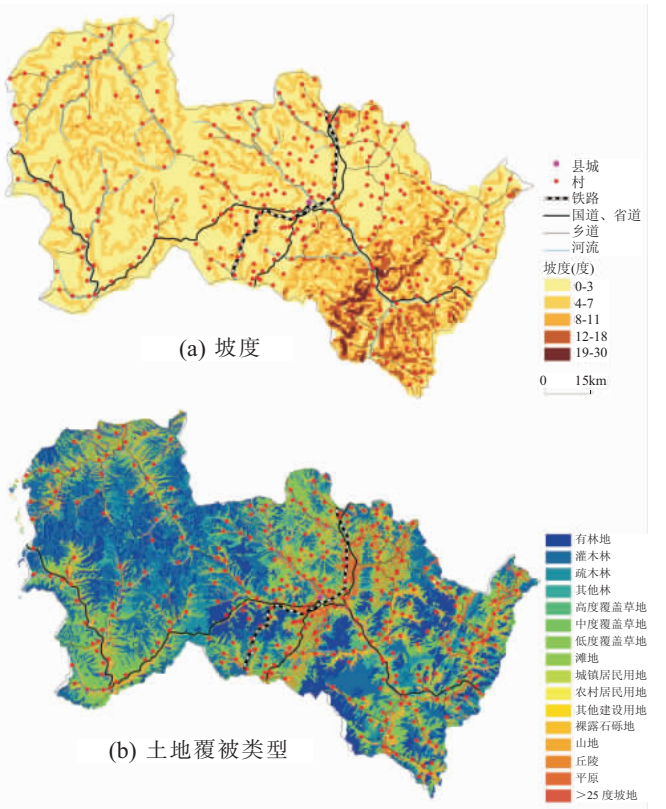

图 1 2001 年和顺县坡度 (a) 和土地覆被类型 (b) 分布

Fig. 1 Distribution of slope (a) and land-cover types (b) in Heshun(2001)

此次研究目标是在格网内人口分布相对可能性的基础上，利用一种精细内插方法，将和顺县 2001 年村普查数据分配到各个格网中去。根据和顺县人口分布实际情况，在空间化过程中建立了一个由 75×30 个 (2,250 个数据点) 1 km^2 大小格网组成的格网层。借鉴Dobson[20]、岳天祥[14]和 Nelson[21]等人研究经验，研究选取了以下几个影响因子 (图 1)：① 坡度，以坡度类型宜居程度为权重；② 河流，权重取值考虑格网到最

近河流的距离；③ 交通设施，权重是格网分别到最近铁路和主要道路距离；④ 土地覆被，直接将不同土地覆被类型上的人口密度作为权重；⑤ 邻近村镇，权重受邻近村庄、县城的人口及其和格网之间的距离影响。除了土地覆被数据由中国科学院地理科学与资源研究所资源环境数据中心提供以外，其他基本数据都来源于和顺县当地政府相关部门。

表 1　遗传规划计算参数

Tab. 1　Genetic programming parameters

项目	参数
群体规模	500
遗传代数	2000
最大生成深度	40
最大交叉深度	17
复制概率	0.60
交叉概率	0.98
突变概率	0.05
终止条件	最大代数：2000；或$R^2 \geqslant 0.9500$

4.2　GP/GA 建模

首先我们随机选择了 261 个村的格网数据(占总数的 80%) 作为训练样本来建立基于GP/GA 的人口分布模型。GP 软件采用英国 Salford 大学开发的 GPC++ 0.40 工具包，GA 程序是利用 Matlab 6.5 自行编码实现的。所有 GP 参数如表 1 所列。其中“最大生成深度”和“最大交叉深度”分别限定了初始个体和交叉后生成个体的规模大小，这样能避免 GP 生成结构复杂庞大的个体，便于最终得到的进化模型解释。

为了获取最能反映真实情况的模型结构，独立运行 GP 程序 100 次，运行结果如表 2 所示。因为 GP 有选择对模型有意义的输入变量能力，所以在 GP 中要测量某个变量重要性只要统计其在多次运行结果中被选择的次数便可。由此可见此次试验 GP 中，不同土地覆被类型上的人口密度和到最近主要道路的距离是最为重要的两个变量。

表 2　100 次 GP 运行中各变量被选择的次数

Tab. 2 Number of input variable selections in 100 GP runs

输入变量	坡度	河流	土地覆被	主要道路	铁路	周边村镇	总计
选中次数	18	6	71	58	15	25	193

最后这些模型中适应度最高的被选择作为和顺县 2001 年人口分布模型结构：

$$popu(i) = 22 - 3.24 \times \frac{\ln\left(\dfrac{road(i)}{205.5 \times lan\_\mathrm{cov}(i) \times slope(i)}\right)}{\exp(0.01 \times nei\_vil(i))} \tag{7}$$

式中：$slope(i)$ 为坡度的归一化值，$lan\_\mathrm{cov}(i)$ 为土地覆被人口密度归一化值，$road(i)$ 为格网到最近道路的距离归一化值，$nei\_vil(i)$ 为邻近村镇影响归一化值，在研究中任意格网所受到的邻近村镇影响等于所有邻近村 (包括县城) 到格网的距离与该村人口总数的比值之和。GA 被用来优化 GP 进化模型里的参数，所以研究中 GA 染色体长度为 4 字节。

在 GA 中个体适应度决定了其存活和繁殖下一代的几率，因而确定合适的适应度函数在整个进化过程显得尤为重要。研究中 GA 采用式 (6) 作为适应度函数，其中参数 ϖ 取 10^{-10}。适应度函数确定之后，GA 便可以根据适应度来选择优良个体进行复制和形成配对池。本研究采用比例选择模式来挑选复制个体的。而且为了避免计算中适应度比例取整时可能会造成新旧种群个体数目不一致问题，GA 还对复制前后所有个体数目差异进行排序，依次对损失较大的个体加 1 直到差异为 0。GA 的个体交叉是通过在每个待交叉个体上选取两个交叉点，互换两个待交叉个体的交叉点之间部分来实现的。与简单遗传算法设置固定交叉概率的做法不同，研究中 GA 的交叉概率是一个位于 (0.8, 1) 之间的随机值。由于所有个体都表现为一个 n 维向量，因此在保证突变后的个体仍在搜索范围内的前提下，GA 采取给所选个体加噪声的方法来实行个体突变。突变算子采用多级变异，突变概率也是一个间于 (0, 0.1) 的不确定值。在 GA 中，种群规模对于提高算法效率尤为关键。如果种群规模太大，运算速度便会放慢。研究中 GA 群体规模为 150，迭代 1200 代。经过 GA 优化，研究所用最终的和顺县 2001 年人口分布 GP/GA 模型为：

$$popu(i) = 28 - 2.86 \times \frac{\ln\left(\frac{road(i)}{172.5 \times lan\_cov(i) \times slope(i)}\right)}{\exp(0.02 \times nei\_vil(i))} \quad (8)$$

4.3 与逐步回归分析法和重力模型对比

逐步回归分析法和空间关系重力模型已经多次用来进行人口空间化建模。为了检验所提方法的有效性，我们使用相同的样本分别也建立了逐步回归分析模型和重力模型。

(1) 逐步回归分析模型：我们将上述 5 个因子权重的对数归一化值作为回归分析的输入变量，经过逐步回归最终得到的回归公式为：

$$popu(i) = 49.31 + 543.876 \times land\_cov(i) - 25.792 \times slope(i) + 243.764 \times nei\_vil(i) \quad (9)$$

复合相关系数为 0.835；回归标准差和 S 的比率 (F) 为 574.305。模型显著性检验发现此模型的显著性水平超过 95%，说明通过逐步回归建立的公式 (9) 是可靠的。

(2) 重力模型：事实上除了周边城市和交通设施等社会经济因素以外，人口分布还受到了很多自然要素的影响。所以研究中所使用的重力模型由相同 5 个因子权重变量组成，模型结构为：

$$popu(i) = 59.89 \times road(i)^{0.001} \times rail(i)^{0.002} \times slope(i)^{0.15} \times lan\_cov(i)^{1.03} \times river(i)^{0.007} \times ner\_vil(i)^{1.2} \quad (10)$$

式中：$river(i)$和 $rail(i)$ 分别为格网到最近河流和铁路的距离归一化值。

4.4 试验结果

我们用 GP/GA 模型、逐步回归分析模型和重力模型各自估算了剩下来的 65 个村里 1 km×1 km 网格内人口数目，由此来比较在此情况下三种方法的优劣程度 (图 2、图 4)。通过各种模型所得到的格网内人口数都汇总到所在村上统计，并求取与村的人口普查数据 (图 3) 之间的差异作为误差。

从表 3 可以看出 65 个村里 34%的 GP/GA 模型估算值与政府普查数据一致，有 77%的村比率误差不大于 10%。这些村子大多分布在和顺县的中部和东部地区，包含了整个县域内多数人口。而在重力模型估算中，仅有 51 %的村误差小于 10%，其余的村误差间于 10% -30%。逐步回归分析的结果更糟。只有 60%的村误差小于 30%，有些村的误差竟然超过 55%。从这就可以初步看出在估算精度方面，GP/GA 方法是三种方法里最好的 (图 4)。

表 4 显示了在和顺县所有村范围内的人口真实值 (x_s) 和用三种方法得到的人口估算值 (y_s) 之间的简单线形回归分析结果。回归分析三个指标：回归系数 (b)，相关系数 (R^2) 和均方差 (MSE) 被用来进行方法比较。逐步回归分析模型

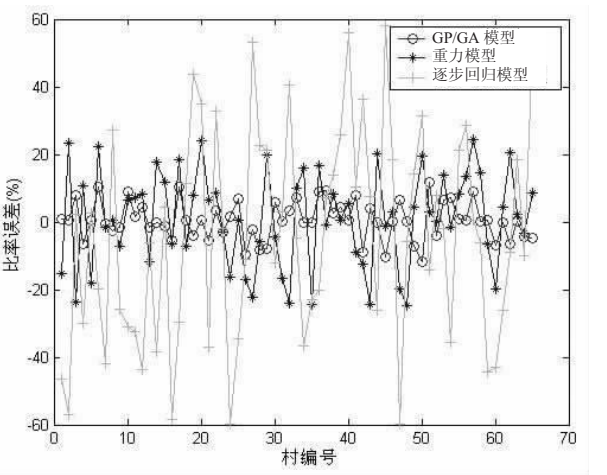

图 2 2001 年和顺县 65 个村人口模拟比率误差分布

Fig. 2 Relative errors in population estimation of 65 villages in Heshun(2001)

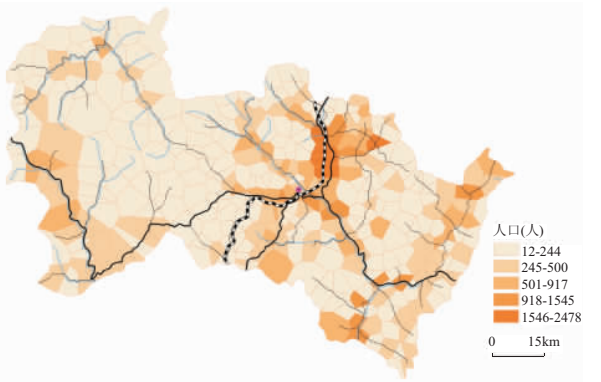

图 3 2001 年和顺县 65 个村实际人口分布

Fig. 2 Distribution of population of 65 villages in Heshun(2001)

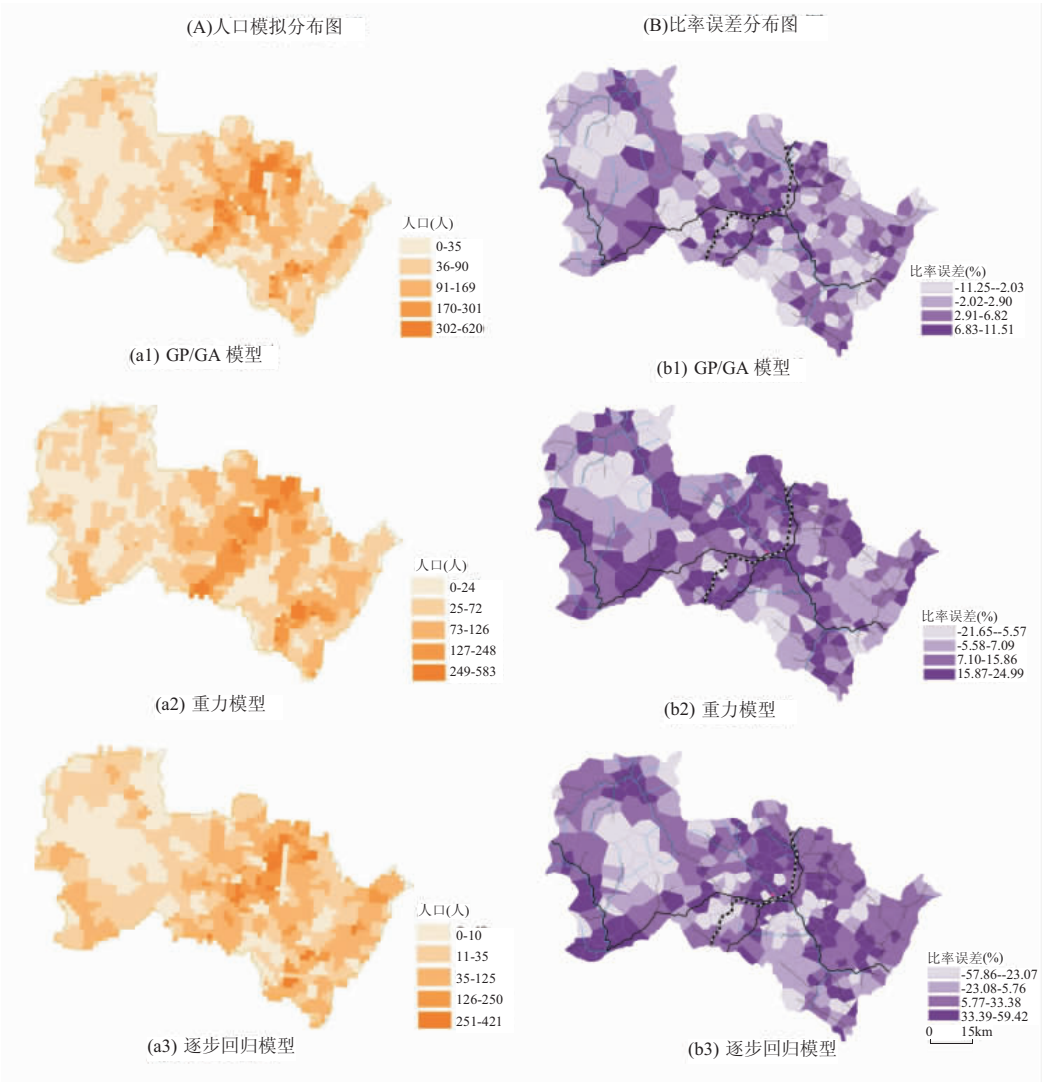

图 4　GP/GA 模型 (a) & (d)、重力模型 (b) & (e) 和逐步回归模型(c) & (f) 模拟结果与误差分布图

Fig. 4　Population estimations and relative errors from GP/GA model (a) & (d), the gravity model (b) & (e) and stepwise regression model (c) & (f)

的回归系数是最大的，0.997，其次是重力模型，0.983，GP/GA 模型的是 0.897，这可能显示了这三种内插方法的逐步低估趋势。逐步回归分析模型的 为 0.973，重力模型的为 0.936，GP/GA 模型的为 0.805。这反映了 GP/GA 模型的拟合度要高于重力模型，而重力模型要比逐步回归分析模型要好。均方差也表现出与 R^2 相同的次序。所以可以说相对于其他两种模型，GP/GA 模型是最适宜的研究区域人口数据内插模型。

表 3 65 个村比率误差统计

Tab. 3　Statistics of relative errors in 65 villages

比率误差	村 (个)		
	GP/GA 模型	重力模型	逐步回归分析模型
< -0.3	0	0	16
-0.3 ~ -0.2	0	6	7
-0.2 ~ -0.1	2	9	5
-0.1 ~ 0	20	13	6
0	22	4	2
0 ~ 0.1	18	16	7
0.1 ~ 0.2	3	11	6
0.2 ~ 0.3	0	6	6
> 0.3	0	0	10

通过比较这三种不同方法的人口分布模拟结果图，我们会发现在研究区的一些人口稀疏地方尤其是西部村庄，这三个模型都大

大高估了当地的人口数。从图中可以看到这些区域几乎都很靠近交通设施 (或) 相对发达乡镇，一小部分的城镇人口被分配到农村格网里去导致了这种高估结果。在那些人口总数非常少，从县级尺度几乎可以忽略的村庄，这种差异更为明显。人口高估最多的地方是在义兴，虽然在建模过程中它的各种变量值都非常好。这个结果很有可能与它作为县城拥有大量的与当地居住人口不成比例的市政公共和基础设施有关。同样在研究区域中部很多交通、经济中心，模型估算人口值都比普查数据要高。

表 4 三种模型得到的和顺县 326 个行政村人口估算值与真实值之间的回归分析结果

Tab. 4 Results from regression analysis between ´real´ population in 326 villages and population estimation from the three models

参数	回归系数	相关系数	均方差
GP/GA 模型	0.897	0.973	801.132
重力模型	0.983	0.936	3976.289
逐步回归模型	0.997	0.805	22794.268

5 结论

随着人口空间化研究的深入，越来越多的基于影响因子和普查数据的人口估算模型和方法得到应用。它们能够为理解众多社会政治过程与现象提供人口规模、行为和空间分布信息。但是在特定区域里到底是哪些因子对人口分布起作用呢？它们又是怎样影响的呢？没有哪个现有模型和方法能很好地解决这个问题。本文是第一次将智能算法运用到人口数据空间化研究中，开拓了人口数据内插方法研究新思路。

尽管本文所提出来的方法仅在山西省和顺县得到检验，但是它可以用来解决任何尺度下的人口空间化问题。因为 GP/GA 能在不同情况下通用，可以给不同区域看似不同的问题提供一个统一的解决方法。除此以外，它还可以用来同时进行好几个区域人口空间化。但是这种方法也存在着一些问题需要在未来的研究中继续解决。作为一种基于统计的算法，其精度的高低不仅受到算法本身优劣的影响，而且样本大小、样本类型及其特性都会起作用。因而这需要进行更为广泛的数据收集工作以确保 GIS 数据的实时性和精确性。还有不同区域、时段的人口分布所受到的各种影响因素不同，如何选择合适的影响因子来避免计算产生偏差仍旧是个难题。除此以外，GP/GA 算法本身也些待改进之处，例如局部搜索能力差，收敛速度慢等。但是随着 GIS、计算机技术的发展和人口空间分布知识的累积，这些问题都会被逐一解决。

参考文献 (References)

[1] Wang Xuemei, Li Xin, Ma Mingguo. Advance and case analysis in population spatial distribution based on remote sensing and GIS. Remote Sensing Technology and Application, 2004, 19(5): 320-327. [王雪梅, 李新, 马明国. 基于遥感和 GIS 的人口数据空间化研究进展及案例分析. 遥感技术与应用, 2004, 19(5): 320-327.]

[2] Tobler W R. Smooth pychophylactic interpolation for geographical regions. Journal of the American Statistical Assoc., 1979, 367(74): 519-530.

[3] Martin D. Mapping population data from zone centroid locations. Transactions of the Institute of British Geographers, 1989, 14(1): 90-97.

[4] Dobson J E, Bright E A, Coleman P R. Landscan: A global population database for estimating populations at risk. Photogrammetric Engineering and Remote Sensing, 2000, 66: 849-857.

[5] Kishore J K, Patnaik L M, Mani V et al. Genetic programming based pattern classification with feature space partitioning. Information Sciences, 2001, 131: 65-86.

[6] Koza J R. A genetic approach to econometric modeling. In: Sixth World Congress of the Econometric Society, Barcelona1, 990c.

[7] Jin Y Q, Wang Y. A genetic algorithm to simultaneously retrieve land surface roughness and soil wetness. International Journal of Remote Sensing, 2001, 22(16): 3093-3099.

[8] Holland J. Adaptation in Natural and Artificial System: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. Cambridge, Mass: MIT Press, 1992. 211.

[9] Zhang Liechao, Cai Zhihua, Chen Ansheng. A survey of SGA, GP, GEP. Control & Automation, 2006, 22(2): 185-187, 145. [张烈超, 蔡之华, 陈安升. SGA、GP、GEP 的研究概述. 微计算机信息, 2006, 22(2): 185-187, 145.]

[10] Yun Qingxia, Huang Guangqiu, Wang Zhanquan. Genetic Algorithms and Genetic Programming: An Approach for Search and Optimization. Beijing: Metallurgical Industry Press, 1997. [云庆夏, 黄光球, 王战权. 遗传算法和遗传规划: 一种搜索寻优技术 北京: 冶金工业出版社, 1997.]

[11] Koza J R. Genetic programming: A paradigm for genetically breeding populations of computer programs to solve problems. Stanford University Report STAN-CS-90-1394, 1990, http://www.genetic-programming.com/jkpubs72to93.ht-ml#anchor484765 .
[12] Deichmann U. A review of spatial population database design and modeling. Technical Report 96-3, National Center for Geographic Information and Analysis, USA. 1996.
[13] Yue T X, Wang Y A, Liu J Y. Surface modeling of human population distribution in China. Ecological Modeling, 2005, 181: 461-478.
[14] Li G Y, Weng Q H. Using Landsat ETM+ imagery to measure population density in Indianapolis, Indiana, USA. Photogrammetric Engineering & Remote Sensing, 2005, 71(8): 947-958.
[15] Balk D L, Deichmann U, Yetman G. Determining global population distribution: Methods, applications and data. Advances in Parasitology, 2006, 62: 120-154.
[16] Yue T X, Wang Y A, Chen S P. Numerical simulation of population distribution in China. Population and Environment, 2003, 25(2): 141-163.
[17] Lu Shaohua. Application of genetic programming in China's port throughput prediction. Journal of Wuhan University of Technology (Transportation Science & Engineering), 2006, 30(3): 520-523. [卢少华. 遗传规划在港口吞吐量预测中的应用. 武汉理工大学学报, 2006, 30(3): 520-523.]
[18] Wang Jiayao, Deng Hongyan. A model of cartographical generalization based on genetic algorithm. Geomatics and Information Science of Wuhan University, 2005, 30(7): 565-569. [王家耀, 邓红艳. 基于遗传算法的制图综合模型研究. 武汉大学学报·信息科学报, 2005, 30(7): 565-569.]
[19] Su M C, Chang H T, Application of neural networks incorporated with real-valued genetic algorithms in knowledge acquisition. Fuzzy Sets and Systems, 2000, 112: 85-97.
[20] Dobson J E, Bright E A, Coleman P R et al. Landscan: A global population database for estimating populations at risk. Photogrammetric Engineering and Remote Sensing, 2000, 66: 849-857.
[21] Nelson A, Deichmann U. The African Population Database, Version 4. New York: United Nations Environment Program (UNEP) and the Center for International Earth Science Information Network (CIESIN), Columbia University. 2004. http://www.na.unep.net/datasets/datalist.php3.

A Method of Spatialization of Statistical Population

LIAO Yilan[1,2], WANG Jinfeng[1], MENG Bin[3], LI Xinhu[4]
(1. *Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing* 100101, *China*;
2. *Graduate School of the Chinese Academy of Sciences, Beijing* 100039, *China*;
3. *College of Arts and Science of Beijing Union University, Beijing* 100083, *China*;
4. *Institute of Urban Environment, CAS, Xiamen* 361003, *China*)

Abstract: Mapping distribution of population has arisen as an important issue in the fields of geographical and relative researches, due to the necessity of combining with spatial data representing socio-graphic information across various spatial units, such as to evaluate the total numbers of people at environmental health risks or died in natural disasters. However, most existing solutions to this problem focus on selection and quantification of influencing factors and rarely take into account the correlation among selected factors. And much expertise is needed in modeling process to formulate the relationships between influencing factors and population data successfully. It usually not only produces information redundancy but increases the complexity of the problem. This paper explores a novel approach to transform population data from census to grid by integrating genetic programming (GP), Genetic Algorithms (GA) and Geographic Information Systems (GIS). A set of natural and socioeconomic factors which contribute to population distribution are identified and quantified under GIS environment. And then GP and GA are severally applied to build and optimize the population model in the hierarchical form, allowing for the computation of the relevant population data error. The experiment proves that the proposed method performs much better than stepwise regression analysis and adapted gravity model approaches. The GP/GA-based method is the first to introduce such computational intelligence techniques as GP and GA to generate gridded population maps, hence it is a methodological innovation in interpolation of population data.

Key words: spatial interpolation; surface modeling; GP; GA; GIS; Shanxi

以地理学家为主导编制的《东北地区振兴规划》发布实施

2007 年 8 月 2 日国务院对《东北地区振兴规划》进行了批复，8 月 20 日国务院振兴东北办在国务院新闻办公室举行新闻发布会，请国务院振兴东北办的主要领导介绍了有关情况。这标志着由国务院发布的第一个大区域规划正式实施。地理学家在这一规划的研究、编制过程中发挥了重要作用。

1. 中国科学院地理科学与资源研究所是这一规划的第一技术支撑单位

《东北地区振兴规划》是根据国务院 2005 年工作要点的要求，由国务院东北地区等老工业基地振兴办公室具体组织编制的具有全局性的规划，是“西部大开发、东北地区等老工业基地振兴、中部崛起、东部率先发展”战略的具体落实。在编制过程中成立了综合研究组、专题研究组、地方研究组和专家咨询组，组织了中国科学院、教育部、中国社会科学院、国家发展与改革委员会、环保总局等部委所属的科研院所参加了规划的研究和编制工作。组织单位经过对全国相关科研单位的筛选和评价，最终委托中国科学院地理科学与资源研究所作为规划的第一技术支撑单位，第二技术协调单位为东北师范大学，参加单位包括中国科学院东北地理与农业生态研究所、中国社会科学院工业经济研究所、吉林大学、国家发展与改革委员会宏观经济研究院国土开发与地区经济研究所等 30 多个单位。中国科学院地理科学与资源研究所的金凤君研究员是技术总协调人，陆大道院士被聘为规划编制专家咨询组组长。

2. 地理学家和区域经济学家在规划编制过程中发挥了重要作用

《东北地区振兴规划》是以国家战略需求为目标的综合性区域经济社会发展与振兴规划，涉及产业、区域、生态、环境、人才、体制、创新等多项内容，需要多学科多领域的综合性科研力量支撑，以及具有较强学科背景的学者进行系统设计。项目主要负责人在区域发展基础的科学分析与问题甄别、规划的顶层设计、规划实施内涵的设计等方面起到了核心设计者的作用。规划中目标的提出、产业调整方向、基础设施建设、生态建设、环境保护、能源开发、资源性城市可持续发展、人才建设、区域创新等方面的内容，均是由具有较强地理学背景和区域经济学背景的科研人员研究完成的。科学研究报告总字数达 150 万字，正在出版过程中。实践表明，地理学在国家经济社会建设中具有重要的应用基础。

3. 规划强调东北地区的振兴须以发展环境营造、产业优化升级、强化创新能力建设、促进机制变革创新、优化空间布局等为突破口

发展环境的营造是东北地区振兴的关键。突出的重点是：加大金融支持力度，再造信用东北；以转变政府职能和深化国企改革为重点，大力发展非公有制经济，加快完善市场体系，形成促进东北地区振兴的新机制；深化改革，扩大开放，打造全面发展的环境。

坚持“自主创新、重点跨越、支撑发展、引领未来”方针，构建以企业为主体、市场为导向、产学研相结合的技术创新体系。全面提高创新能力，为东北振兴提供强力支撑。

以整合和创新为主线，优化产业结构、组织结构和空间结构，建设具有竞争力的新型产业基地。基础原材料工业应本着“优化存量、深化加工、集聚发展”的思路，调整企业的规模结构与空间结构，形成具有全国意义精品原材料生产基地；装备制造业从产业结构和组织结构调整入手，提高自主开发能力和当地配套能力，形成具有国际竞争力的先进装备制造业基地；同时，积极扶持和发展高新技术产业和现代服务业，提升高新技术产业对东北地区产业结构调整的引领作用，以及现代服务业对第二产业和第三产业的带动和促进作用；继续巩固东北地区农业基础地位及国家重要商品粮基地的作用，推动农业产业化和农业基础设施的建设，加快东北地区新农村建设步伐。

主体功能区为支撑，强化空间布局，促进东北地区的快速振兴。重点地带、重点城市和重点产业的加快发展是振兴东北的重要措施之一。未来东北地区振兴以哈尔滨 (含齐齐哈尔和大庆) 一大连 (哈大轴线) 和东北沿海 (沿海轴线) 为一级发展轴线，优化空间发展格局。注重提高城市化质量，积极发展具有国际竞争力的大都市经济区，发挥沈阳、大连、长春和哈尔滨四城市的主体功能作用，强化合作与分工，共同带动东北地区振兴。大力推进资源型城市可持续发展，完善资源型城市的综合功能，积极发展接续产业，加强环境综合整治。

加强生态环境建设与保护。建成有效的水土保持综合防护体系，使适宜治理的水土流失地区基本得到整治，黑土区水土流失严重的状况得到基本控制。加强矿产资源的保护，建立矿产资源开发与环境保护协调机制，建立矿山环境恢复补偿机制和矿山环境监督管理机制。

(金凤君)