城市系统微观模拟中的个体数据获取新方法

龙 瀛[1,2], 沈振江[3], 毛其智[1]
(1. 清华大学建筑学院，北京 100084；2. 北京市城市规划设计研究院，北京 100045；
3. 日本金泽大学环境设计学院，日本金泽 920-1192)

摘要：目前自上而下的城市系统宏观模拟并不能解决城市这一复杂系统中出现的部分问题，城市系统微观模拟 (如多主体系统MAS) 已经成为城市系统模拟的新思路，其主要是基于个体数据 (如个人、家庭、公司或建筑物) 开展的。国际国内这方面的应用都受到个体样本稀缺的限制。微观模拟所需要的个体样本数据是原有的统计制度所不能适应的，尤其是在中国，个体样本在统计公报或年鉴中不公开，仅可通过典型调查来补充。本文旨在探索稀疏数据环境下构建城市系统微观模拟的个体样本数据的新方法。该方法基于已有的多源统计数据、典型调查数据以及个体的通用规则，反演出个体样本的属性信息和空间分布，进而可以以GIS图层的形式直接作为微观模拟的数据基础。通过本方法获取的样本，能够符合已有的统计资料，并遵照了样本的基本特征，可以作为现有数据条件下的微观模拟模型的数据输入。同时该方法的应用简单，统计意义上的准确度高，适合我国统计制度下的微观模拟模型的构建。
关键词：微观模拟；反演；多主体系统；统计数据；北京

1 引言

本文针对目前城市系统微观模拟方法在应用中面临稀疏数据环境制约的困境，试图提出一种基于现有的宏观统计资料、多个分散的典型调查数据、通用规则获得的样本属性间的依赖关系，反演个体样本的方法。通过本方法获取的样本，能够符合已有的统计资料，并遵照了样本属性之间的一般规律，可以以GIS图层的方式直接作为微观模拟模型的基本输入数据。该方法的应用简单，准确度高，适合中国现有统计制度下的微观模拟模型的构建，是对稀缺数据环境下微观模拟数据获取方法的有益探索。

由美国经济数学家Orcutt等创建的微观模拟模型[1] (Microanalytic Simulation Model，MSM)，在研究城市问题时能够较好地弥补宏观分析模拟模型的不足。与传统的自上而下的宏观分析模拟不同，城市系统微观模拟是典型的自下而上的过程，它以企业、家庭乃至个人等微观个体作为描述、分析和模拟的基本对象，每个微观个体都具有独有的自身特性与丰富的内部认知结构[2-4]。多主体系统 (Multi-agent System，MAS) 是目前城市系统微观模拟的典型研究方法之一，其所分析的基本对象是个体而非一类个体的集合 (如同一个行业的就业者、一个镇的居民)，因而它的原理与方法清晰直观，比较容易为人所接 受[5]。微观模拟的应用目前受到个体样本稀缺的限制。个体样本的属性特征是微观模拟的核心数据[6]，例如居民样本的属性主要包括年龄、性别、民族、婚姻、教育、工作等。这些原始数据的获取目前往往需要通过调查、搜集、整理以及正确地输入计算机这一系列的过程，需要大量的人力、物力和财力[7]。目前，中国多数统计部门的基本样本数据不公开，可获取的仅为统计层次上的数据，如某个行政区或某个行业的总体信息，适用于宏观分析模

收稿日期：2010-01-31; **修订日期：**2010-07-25
基金项目：国家自然科学基金项目 (51078213); 国家“十一五”科技支撑计划项目 (2006BAJ14B08) [**Foundation:** National Natural Science Foundation of China, No.51078213; Technical Supporting Programs Funded by Ministry of Science & Technology of China, No.2006BAJ14B08]
作者简介：龙瀛 (1980-), 男, 博士研究生, 高级工程师, 中国地理学会会员 (S110007674M), 主要研究方向为规划支持系统和城市系统微观模拟。E-mail: longying1980@gmail.com

拟，但并不适合微观模拟。国外的研究也大多不能采集到现时、可立刻使用的个体样本数据[8-9]。因而在国内外研究中，往往采取大量分散的、不同目的的调查形式，用于独立的不同微观模拟中，但这些数据在各个独立研究之外并没有得到系统的应用。因此微观模拟的个体样本数据的稀缺情况，在国内外都正制约着微观模拟进一步发展。

在数据稀缺的环境下，MAS的主要问题之一是如何考虑agent属性乃至偏好的异质性。很多反映土地使用动态的MAS，都不能很好地反映居民层次(actor-level)样本信息属性的异质性，这类模型倾向于使用集计 (aggregate) 数据作为agent的属性，如ILUTE[10]，或根据文献调研获得可接受的取值范围然后随机赋值给各个居民agent，如LUCITA模型[11]。在已有的MAS中，往往不能够做到1个agent对应1个居民个体，如张鸿辉等[12]将30 m×30 m的网格内的居民数量作为1个agent，陶海燕等[13]在居住区位选择的MAS中，1个网格对应1个居民agent，并不是网格空间所对应的实际居民数目；Shen等[14]分别尝试将1、2、3、5和10个居民作为1个agent，发现不同的比例对模拟结果具有较大的不确定性。因此可以看出，因为数据稀缺问题，多数MAS都不能实现1个agent对应真实城市的1个居民或家庭，同时对居民或家庭的集聚会带来模拟结果的不确定性，如果过于集聚也失去了微观模拟的精髓。Brown和Robinson[15]的研究也表明MAS中居民偏好的异质程度对模拟的土地使用形态具有较大的影响，因此个体样本信息对于MAS至关重要。

个体样本信息主要分为空间信息和属性信息，因此个体信息的反演分为空间反演和属性反演两类，其主要是基于宏观的统计数据实现。关于个体样本的空间反演，更多的研究侧重于人口数据空间化[16-20]，这方面的研究已较为成熟，其基于宏观的人口普查公报结合相关因素，利用GIS反演人口密度曲面，进而反演人口样本的空间分布，这一过程并不涉及个体样本属性数据的反演。基于人口数据空间化的反演结果，可以实现个体样本的空间定位，结合环境空间图层，可以识别个体样本所处的环境变量 (如Robinson和Brown[21]所提及的与学校或工作地的距离、邻里相似性、景观质量等)，进而作为agent的空间属性用于MAS模拟[8-9,12,14,22-23]，这种以环境变量作为居民agent主要属性的研究思路是当前数据稀缺环境下城市系统MAS的主流方法，这种情况下对个体样本的自身属性考虑的还比较少。

实际上很多情况下agent进行决策还是需要自身属性，如居民agent的收入、受教育程度等属性，对居住区位选择、就业区位选择等行为都有较大的影响，因此agent的自身属性在MAS中也需要重视。关于个体样本的自身属性的反演，Brown和Robinson[15] 基于592份居住地选择影响因素的调查问卷，分别采用随机设定、基于调查得到的概率分布等方法，分别设计了5种不同异质程度的居民agent居住区位选择偏好的实验，但实验中并没有考虑居民agent的自身属性，而是居民agent针对不同环境因子的偏好。Li和Liu[24]初步地指出了利用统计数据定义agent属性的可能性，其根据统计数据将所有城市居民根据有无子女、收入两个属性分为6类，每类具有不同的环境变量偏好，但其仅考虑两个自身属性并基于这两个自身属性将个体样本分为4类，并没有给出每个样本的反演的具体属性数值，也没有考虑样本属性间的关系。陶海燕等[13]根据统计年鉴采用概率分布的形式构建了居民agent的收入属性，并基于不同的收入区间设定了不同的环境因子偏好，但其并没有考虑多个属性及属性之间的关系。另外，Hynes等[25]利用模拟退火算法利用国家宏观农业普查数据修正典型调查数据，实现二者的匹配，与本文的研究内容并不相同。可以看出并没有文献重点针对个体样本的自身属性进行反演。

已有研究的不足之处在于，大多数MAS中的agent的属性多是根据个体样本所处的空间位置所确定的环境变量，agent的自身属性鉴于数据稀缺的原因考虑的较少。本研究旨在针对这一问题，国内外利用统计资料、典型调查、典型规则反演个体样本的空间和属性信息。

2 方法

2.1 已知信息

城市系统微观模拟的样本，根据模拟的对象和研究层次，可以为个人、家庭、建筑、车辆、地块等，也可以是子行政区域，样本是微观模拟中最为基本的研究对象。个体样本一般具有相应的社会、经济等方面的基本属性，例如人口样本的年龄、收入、教育程度、职业等属性，家庭样本的家庭成员数量、收入、地址等属性，地块样本的面积、产权单位、使用类型、高度等属性。这些属性又可以分为空间类属性和非空间类属性，如地块样本的面积、形状和位置，以及人口样本的所在街区等属于空间类属性。

已有的宏观信息 (aggregate data) 主要分为3个方面：①官方统计数据，其往往是对样本数据的统计描述，如样本的总个数，样本某属性的分布规律，如各个职业的人口分布，各收入区间的分布，样本属性与属性的相互关系 (如婚姻状况和年龄、收入水平和受教育程度的依赖关系)；②多个独立的典型调查数据，可以从中获得属性的分布信息和属性之间的依赖关系；③已有的通用规则，分为常识性 (如年龄低于18岁婚姻状态为未婚) 和科研成果，这些规则描述了样本的各属性之间的相互关系。因此，样本的已知信息可以概括为样本各属性的概率分布、样本各属性之间的关系。

2.2 数学建模

样本的属性分为连续类型和离散类型，在提出的方法中样本的所有属性都采用离散类型的方式进行建模，如婚姻状态属性分为已婚、未婚两个离散值，而年龄属性分为0~4、4~10、10~20等不同区间。因此每个可选数值均为字符串类型，无论原始是字符串型属性还是数值型属性。其目的一方面可以简化建模，另一方面可以简化样本反演的运算量。针对连续类型的属性，可在反演得到取值区间后从中随机选取数值来实现具体数值的确定。本文将每个属性的可选择数值称为域，每个数值称为域元素 (字符串类型)，样本中每个域元素出现的次数称为频数，占总样本的比例称为频率。例如，对于婚姻状态属性，其域为{已婚; 未婚; 离异}，其中“已婚”为域元素，相应的频数{45; 20; 35}表示45个样本为已婚，20个样本为未婚，35个样本为离异。这样就将反演样本的输入转变为域及其频数构成的域频表，及不同属性的域频表之间的关系表。

如下是反演过程中需要的变量名称。样本总数目(N)，属性总数目(M)，样本ID(i, $i \in \{1, 2, 3,\cdots,N\}$)，属性ID($j$, $j \in \{1, 2, 3,\cdots,M\}$)，属性类型($T_j$ 为字符类型则为STR，为数值类型则为NUM)，样本i的j属性反演数值($a_{i,j}$ 字符串类型)，反演样本值集合($A_{N\times M} = \{a_{i,j}\}$ 矩阵形式，二维数组)，样本i的j属性真实数值($a'_{i,j}$ 整数、小数或字符型)，真实样本值集合($A'_{N\times M} = \{a'_{i,j}\}$ 矩阵形式，二维数组)，i样本的所有属性值(a_i 行)，j属性的所有样本值(a_j 列)，b_j(属性j的域 ,集合)，k(域元素的ID)，$b_{j,k}$(属性j的第k个域元素)，K_j(属性j的域元素的总个数)，P_j(属性j的所有域元素的频数，即集合)，$P_{j,k}$(属性j的第k个域元素的频数)，$p_{j,k}$(属性j的第k个数值的概率，$P_{j,k} = P_{j,k}$/N)，f_j(属性j的概率密度分布函数)，h_j(与属性j有概率关系的属性ID)，H_j(与属性j有函数关系的属性ID集合)，g_j(属性j与属性H_j的函数关系)。

2.3 分布分析

鉴于每个属性的已知信息类型不同，使得每个属性的反演方式也不同。需要对样本的属性根据已知的分布特征进行分类，以N个样本的j属性为例：

(1) 频数已知的分布DB：文字类型，有K_j个数值可选，分别是$b_{j,k}$，每种的个数为$P_{j,k}$，则j属性反演后的数值为：

$$\boldsymbol{a}_j = \mathbf{rand}\mathbf{O}\left(\bigcup_{k=1}^{K_j}\{\underbrace{\boldsymbol{b}_{j,k}, \boldsymbol{b}_{j,k}, \cdots, \boldsymbol{b}_{j,k}}_{P_{j,k}}\}\right) \tag{1}$$

式中：*randO*函数表示对集合内容的随机排序 (这里假设集合为有序集合，不是无序的)。

(2) 概率密度已知的分布DA：即已知属性j符合概率密度分布函数$p(x = x_0) = f_j(x_0)$ (例如高斯分布、均一分布等)，从中共采集K_j个域元素 (用中值表示其数值)，$P_{j,k} = N \times p(x = b_{j,k}) = N \times f_j(b_{j,k})$，即数值等于$b_{j,k}$的样本数目为$N \times f_j(b_{j,k})$个，共选取$K_j$个数值，进而将问题转变为DB：

$$a_{,j} = randO\left(\bigcup_{k=1}^{K_j}\{\underbrace{b_{j,k}, b_{j,k}, \cdots, b_{j,k}}_{N^* f_j(b_{j,k})}\}\right) \tag{2}$$

(3) 无分布DC：不知道该属性的分布。

2.4 关系分析

利用不同来源的已知信息，如调查问卷、常识性知识、统计公报等，可以将这些已知信息转变为属性和属性之间的依赖关系，主要包括以下几种形式的建模规则：

(1) 函数关系RA：某样本i的某属性与本样本的其他属性 (一个或多个) 有函数关系，如$a_{i,j} = g_j(\bigcup_{h \in H_j} a_{i,h})$，$g_j$可以是线性关系，也可以是非线性关系 (如决策树形式的关系)，在其他属性反演的基础上，可以给出该属性的样本值$a_{,i}$。

(2) 概率关系RB：属性j与属性h_j (简化为h) 有概率关系 (频数上的关系)，并已知二者的频度，二者概率关系可以用关系矩阵$Q_{K_h * K_j}$表示，其元素q_{k_h,k_j}已知，表示关系概率，$q_{k_h,k_j} = P\{(a_{i,h} = b_{h,k_j}) \cap (a_{i,j} = b_{j,k_j})\}$，其中$P$为概率 (0~1，区别于变量$P_j$)。逐一对所有样本的属性$j$进行计算，给出每个样本对应的$K_j$个取值的概率，计算方法为$p\{a_{i,j} = b_{j,k_j}\} = q_{k_h,k_j} \times p_{j,k_j}$，其考虑了双重概率，然后利用蒙特卡洛的方法根据概率P可以得到属性j的所有样本值$a_{,j}$。$where\ a_{i,h} = b_{h,k_h}$

(3) 无关系RC：该属性不存在或不知道与其他属性的关系。

2.5 分布与关系的耦合分析

样本的不同属性的分布和关系的类型各不相同，这就需要对不同属性的数值反演有不同的方法。根据分布的类型和关系的类型，在反演样本时的耦合关系可以分为9种 (表1)。其中VBB和VAB都是用RB的结果后，再和DB去对比，修正规则，直到符合DB；VCB直接用RB的结果，VBC直接用DB的结果，VAC直接用DA的结果，VCA直接用RA的结果，而VAA、VBA、VCC不需要考虑。

表1　分布与关系的耦合分析表

Tab. 1　The coupled distribution and relation types

分布/规则	函数关系 RA	概率关系 RB	无关系 RC
概率密度已知的分布 DA	VAA (NA)	VAB	VAC
频数已知的分布 DB	VBA (NA)	VBB	VBC
无分布 (或未知) DC	VCA	VCB	VCC (NA)

注：NA 表示本文对这类关系不作分析，因为这种情况不存在或不必要进行分析

根据上面的分析，可以看出，9种情况都可以通过分布分析和关系分析中的算法实现。需要特别注意的是，对于VBB，即概率分布、概率关系的属性，需要对RB计算的结果进行验证，如果其结果的概率分布符合DB，则将其作为最终结果，否则调整RB的设置，使得最终RB计算后的结果的概率分布符合DB。

如果属性j的类型为数值型NUM，则所获得的属性数值a_j代表的是一个数值范围，a_j^L表示其最低值，a_j^H表示其最高值。则最终的样本值为

$$a'_{,j} = \begin{cases} a_j, & if\ T_j = "STR" \\ randV(a_j^L, a_j^H), & if\ T_j = "NUM" \end{cases}$$

式中：*randV*表示从数值范围中根据平均分布的规律选取随机值的函数。

2.6 反演结果的空间化

反演得到的个体样本数据一般是表格形式的个体属性数据，但城市系统微观模拟模型如MAS中，一般还需要个体样本的空间位置。为此，需要为每个样本增加一个属性字段FID，表示每个样本所对应的几何体 (点、线或多边形) 的唯一编号。所有个体样本所对应的整个空间由若干个几何体构成，每个几何体所包含的样本数量可以基于统计数据推测。将各个几何体的FID及其包括的样本数量作为一种概率分布，则可以在属性反演的过程中给出每个样本对应的几何体的FID属性。在此基础上，每个几何体采用随机的方式生成其所包括个体样本数量的点要素，作为个体样本对应的空间对象，进而将反演得到的个体样本与空间对象进行一一对应。在本文，个体样本反演的结果采用GIS点的空间图层形式存储，其在具备空间位置信息的同时还具备个体样本的所有属性信息，根据点所在的位置，还可以识别相关的空间属性 (如可达性、规划条件等)，进而为城市系统微观模拟提供更为完善的数据基础。但是本文注重探讨个体样本反演的方法，并没有充分考虑个体样本的空间属性与个体样本属性之间的关系，将来使用反演数据进行MAS应用时，还需深入探讨加入结合空间属性的个体样本反演。

3 应用

根据上述方法，采用Python脚本语言，基于Access数据库，在ESRI ArcGIS平台的Geoprocessing支持下开发了Agenter (Agent Producer) 模型，用于个体样本数据的反演。

要进行个体样本的反演，首先需要确定反演的样本总量，并根据相关的基础数据，构建相关属性的分布、关系及其耦合关系，将其以数据库的形式存储作为输入数据，并以数据库作为反演结果的存储方式。Agenter的数据库的构成如图1所示，主要是属性表的形式，其中“AgentAttrInfo”表存储了个体样本的每个属性的基本情况，如所属的分布类型、关系类型、耦合类型和数据类型等，“DA”表存储了概率密度分布的基本信息，如属性所对应的概率分布及其参数，“DB_”类型的表存储了相关属性的概率分布的基本信息，“RA”表存储了函数关系，“RB”表存储了两个属性之间的概率关系的总构成情况，“RB_”类型的表则存储了具体的各相关字段之间的概率关系。而“Agents”表存储的，是以上述表作为输入的Agenter模型反演得到的个体样本及其属性信息。“Agents_spt”为

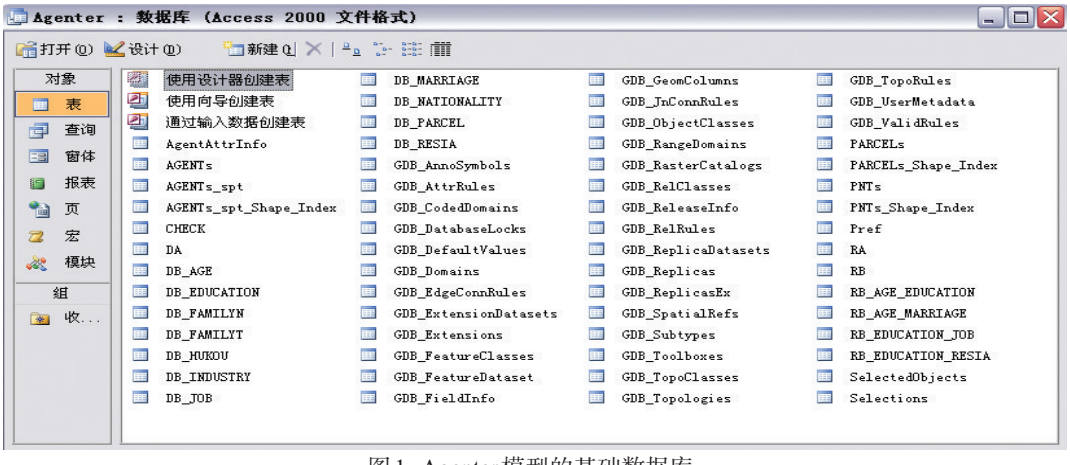

图1 Agenter模型的基础数据库

(GDB和Select类型的表格为ESRI的Personal Geodatabase的内置数据表，不是Agenter的输入数据)

Fig. 1 Datasets in the Agenter model

Geodatabase 格式的点状要素数据集，对应“Agents”表的空间化的结果。同时，在 VBB 耦合关系中，使用了 Access 的查询模块验证规则设置与概率分布的匹配情况，即通过反演频度与统计频度的对比，对 RB 关系进行修正。采用所开发的 Agenter 模型，可以对个体样本进行反演，这些样本可以是地块，也可以是个人或家庭。这里以基于北京市第五次人口普查的统计资料[26]，对人口样本数据的反演进行介绍。该统计资料为 2000 年 11 月 1 日零时为标准时间开展的第五次人口普查数据，其中有对人口的镇级总数统计、年龄、婚姻状态、职业、收入、民族、受教育程度、家庭、住房、死亡、迁移人口等统计信息，主要是各类别或各级别的分布。这些信息结合样本属性间的常规规则，可用于构建 Agenter 模型的输入数据。反演的样本总数为 10000 个 (实际上北京市的人口总数远超过这个数字，这里仅以 10000 个作为 Agenter 方法的测试)。

3.1　输入数据

Agenter 模型的输入数据主要是样本总数、属性设置、分布及关系等。样本属性及其参数表 (表 2)，其中考虑了人口样本的年龄、婚姻状况、月收入、受教育程度等 18 个属性信息，该表是样本属性信息的索引表，记录了每个属性的分布类别和关系类别，进而用于 Agenter 模型反演样本。表中的“PARCEL”属性表示人口样本所对应的地块编号，用于反演结果的空间化；“AID”属性表示人口样本所对应的空间点的编号。

分布类型为概率密度分布的属性的分布基本信息(表 3)记录了概率密度函数及相关参数，如月收入 (INCOME) 符合高斯分布，均值为 6000 元，标准差为 1000 元。这类属性与其他属性没有依赖关系。

DB 类型的表的基本形式参见表 4。其中 NAME 列对应 AGE 属性的域 (年龄段)，PERCENT 列对应域元素的频数。

RA 表用于存储函数关系，本案例中的出行方式“TRAVEL”属性属于这一类，其依赖于年龄 AGE、受教育程度 EDUCATION

表 2　AgentAttrInfo 表

Tab. 2　The AgentAtttrInfo table

编号	英文名	中文名	分布	关系	耦合类型	数据类型
1	KEY_ID	样本编号				INTEGER
2	AID	空间点的编号				INTEGER
3	AGE	年龄	DB		VBC	INTEGER
4	MARRIAGE	婚姻状况	DB	RB	VBB	BOOLEAN
5	INCOME	月收入	DA		VAC	LONG
6	JOB	职业	DB	RB	VBB	STRING
7	EDUCATION	受教育程度	DB	RB	VBB	STRING
8	RESIPLACE	居住地	DB		VBC	STRING
9	WORKPLACE	就业地	DB		VBC	STRING
10	SEX	性别	DA		VAC	BOOLEAN
11	INDUSTRY	行业	DB		VBC	STRING
12	RESIA	居住面积	DB	RB	VBB	FLOAT
13	NATIONALITY	民族	DB		VBC	STRING
14	FAMILYT	家庭类型	DB		VBC	BOOLEAN
15	FAMILYN	家庭成员数目	DB		VBC	STRING
16	HUKOU	户口情况	DB		VBC	STRING
17	TRAVEL	出行方式	DC	RA	VCA	STRING
18	PARCEL	所在地块	DB		VBC	STRING

表 3　DA 表

Tab. 3　The DA table

ID	FLD	DIS_NAME	P0	P1	P2	P3	P4	NOTE
1	INCOME	Gauss	6000	1000				高斯分布 P0 为均指 P1 为标准差
2	SEX	Uniform	男	女				

表 4　DB 表 (Name 列的单位为“岁”，PERCENT 列的单位为“人”)

Tab. 4　The DB table

ID	NAME	PERCENT	ID	NAME	PERCENT	ID	NAME	PERCENT
1	0 4	442578	8	35 39	1413872	15	70 74	327136
2	5 9	525033	9	40 44	1237967	16	75 79	187126
3	10 14	876157	10	45 49	1093304	17	80 84	86602
4	15 19	1234173	11	50 54	690082	18	85 89	34563
5	20 24	1259508	12	55 59	505715	19	90 94	9753
6	25 29	1271407	13	60 64	558646	20	95 99	1863
7	30 34	1317888	14	65 69	495674	21	100 150	147

注：该表的 PERCENT 一列的数据并没有进行归一化处理，对应北京市域的实际人口。

和月收入INCOME 3个属性，其函数关系为决策树，基于Python语言的具体形式如下 (本决策树仅供示意Agenter方法的应用)：

```
if AGE <= 4:
    TRAVEL = "无独立出行"
elif AGE >= 75:
    TRAVEL = "非机动车"
elif INCOME > 6000 and (EDUCATION =
  “大学本科” or EDUCATION = “研究生”):
    TRAVEL = "私家车"
elif INCOME <= 1000:
    TRAVEL = "非机动车"
else:
    TRAVEL = "公交车"
```

表5 (RB表) 为关系类型为概率关系的属性的关系基本信息，记录了样本属性间的关系索引，字段FLD中对应的属性的数值依赖于字段FLD_RB中对应的属性值，如婚姻状态(MARRIAGE) 依赖于年龄属性 (AGE)，职业属性 (JOB) 和居住面积属性 (RESIA)都依赖于受教育程度属性 (EDUCATION)。人口各个属性之间的相互依赖关系，可以从人口普查资料中获得，其中有多个表格描述两个属性间的关系，即将某一属性分为若干区间，针对每个区间给出另一属性的分布统计数据。

表5 RB表

Tab. 5 The RB table

ID	FLD	FLD_RB
1	MARRIAGE	AGE
2	JOB	EDUCATION
3	EDUCATION	AGE
4	RESIA	EDUCATION

表6 RB_AGE_MARRIAGE概率关系表

Tab. 6 The RB_AGE_MARRIAGE probability relationship table

ID	AGE(*h*)	MARRIAGE(*j*)
1	1	2 100%
2	2	2 100%
3	3	2 100%
4	4	1 0.3%; 2 99.7%
5	5	1 15.3%; 2 84.7%
6	6	1 71.9%; 2 27%; 3 0.6%; 4 0.5%
7	7	1 88.9%; 2 7.7%; 3 1.7%; 4 1.5%; 5 0.2%
8	8	1 91.4%; 2 3.1%; 3 2.5%; 4 2.7%; 5 0.3%
9	9	1 91.7%; 2 1.7%; 3 2.7%; 4 3.3%; 5 0.6%
10	10	1 92.2%; 2 1.5%; 3 2.1%; 4 3%; 5 1.2%
11	11	1 91%; 2 1.3%; 3 1.5%; 4 3.4%; 5 2.8%
12	12	1 88.2%; 2 1.1%; 3 1.3%; 4 3.9%; 5 5.5%
13	13	1 83.9%; 2 0.9%; 3 1%; 4 3.9%; 5 10.3%
14	14	1 80%; 2 0.8%; 3 0.9%; 4 3.9%; 5 14.4%
15	15	1 76%; 2 0.8%; 3 0.9%; 4 3.9%; 5 18.4%
16	16	1 72%; 2 0.7%; 3 0.8%; 4 4%; 5 22.5%
17	17	1 60%; 2 0.3%; 3 0.5%; 4 2%; 5 37.2%
18	18	1 50%; 3 0.2%; 4 2%; 5 47.8%
19	19	1 40%; 4 2%; 5 58%
20	20	1 30%; 4 2%; 5 68%
21	21	1 20%; 4 2%; 5 78%

具体的，对于AGE与MARRIAGE属性的关系表RB_AGE_MARRIAGE (表6)，AGE字段中的数值为年龄属性的ID (对应于表4中的ID列)，表示年龄阶段，MARRIAGE字段是相应年龄阶段的各种婚姻状况的概率。例如，对于25~29岁 (ID为6) 这一阶段的个人，根据DB_MARRIAGE表，初婚有配偶 (ID为1) 的概率为71.9%，未婚 (ID为2) 的概率为27%，离婚 (ID为3) 的概率为0.6%，再婚有配偶 (ID为4) 的概率则为0.5%，其他婚姻状况的可能均为0。

另外，为了将反演结果空间化，还需要将所有样本所对应的空间分布图层作为模型的输入，本文选取北京市的现状地块图层PARCELs作为样本的分布空间。鉴于本文反演的样本数量为10000个，而北京市域的人口数远超过该数目，因此仅选取中心城局部地区的若干地块作为PARCELs图层 (图2中的灰色多边形)。

3.2 反演结果及分析

10000个反演样本的前20项见表7，根据“PARCEL”属性空间化后的结果(图2)所示。最终的结果形式为ESRI Personal Geodatabase中的点状的要素数据集，每个点都位于某个地块内，具有相应的属性信息。该结果可以直接作为居民agent用于MAS模拟。基于空间化后的个体样本反演结果可以利用不同的属性进行空间分析和空间统计。

3.3 结果验证

不同分布与关系耦合类型的样本属性的反演方法不同，因此需要采用不同的方法进行反演结果的验证：

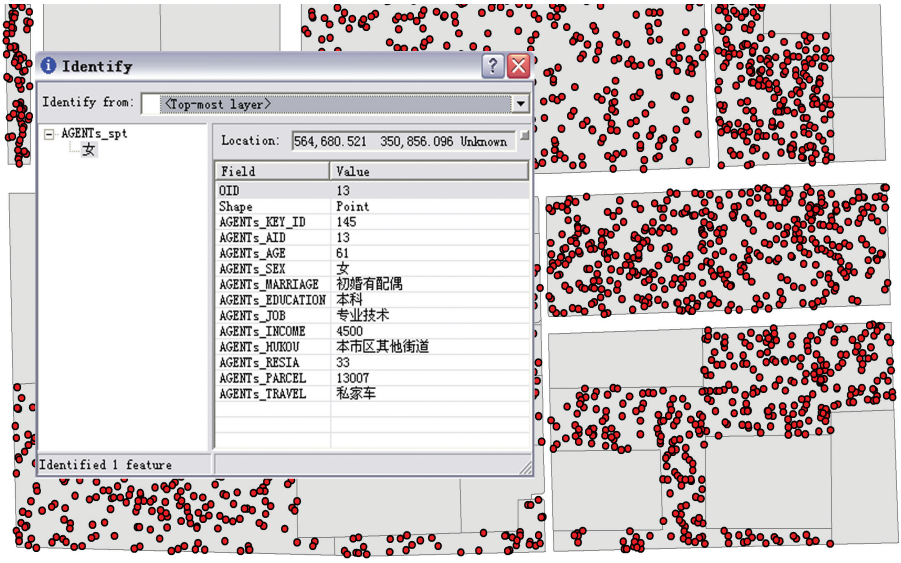

图2　反演的个体样本的空间分布图

Fig. 2　The spatial distribution of disaggregated individuals

表7　部分反演结果

Tab. 7　Part of the disaggregated results by Agenter

KEY_ID	AID	AGE	SEX	MARRIAGE	EDUCATION	INCOME	HUKOU	RESIA	PARCEL	TRAVEL
1	1376	52	男	初婚有配偶	初中	5948	省外	18	259	公交车
2	744	47	女	初婚有配偶	中专	6876	省外	14	197	私家车
3	1785	23	男	未婚	初中	6301	省外	27	338	私家车
4	984	53	男	初婚有配偶	初中	4981	本市区其他街道	20	211	公交车
5	1242	32	女	初婚有配偶	小学	6359	本省其他县(市)、市区	20	248	私家车
6	352	47	女	初婚有配偶	初中	5272	省外	25	90	公交车
7	209	19	女	未婚	小学	4699	省外	28	51	公交车
8	529	51	男	初婚有配偶	初中	5697	省外	23	117	公交车
9	1665	30	女	初婚有配偶	初中	6758	省外	34	320	私家车
10	1117	29	女	未婚	初中	6833	省外	2	237	私家车
11	236	48	男	初婚有配偶	初中	5050	省外	24	62	公交车
12	1168	51	女	初婚有配偶	初中	6964	省外	20	241	私家车
13	384	72	女	初婚有配偶	初中	3812	省外	24	92	公交车
14	1179	52	男	初婚有配偶	初中	5185	省外	21	242	公交车
15	1967	27	男	未婚	小学	3762	省外	4	358	公交车
16	1382	60	女	初婚有配偶	初中	6901	省外	13	259	私家车
17	1666	45	男	未婚	初中	6424	省外	26	320	私家车
18	1284	30	女	初婚有配偶	小学	6237	省外	14	251	私家车
19	1648	35	男	未婚	高中	7060	省外	80	319	私家车
20	1565	34	女	初婚有配偶	高中	6450	省外	17	305	私家车

(1) VAC类型，如“SEX”属性，反演结果中性别为“男”的样本数量为5004，“女”为4996，符合均一分布；对于“INCOME”属性，采用非参数检验 (单样本K-S) 的方法，假设反演得到的样本的“INCOME”属性符合正态分布，估计得到的均值为5993.93 (预先设定为6000)，标准差为989.99 (预先设定为1000)，双尾显著性 (Asymp. Sig. (2-tailed)) 为0.657，接受原假设，因此“INCOME”属性也符合预先设定的概率密度函数。

(2) VBB类型，如“EDUCATION”属性，其验证结果见表8，最大误差为受教育程度为“初中”的样本，误差在21.21%，其余受教育程度样本吻合情况较好，因此总体可以接受。为进一步提高原始分布与反演结果的吻合程度，鉴于“EDUCATION”属性是基于“AGE”属性计算得到，可以通过调整二者的关系表实现。VAB类型的验证方式与VBB类型相同。

表8 “EDUCATION”属性的原始分布与反演样本分布的对比

Tab. 8 The comparison table of the original and disaggregated distributions of the attribute of EDUCATION

ID	NAME	原始人数	原始比例 (%)	反演人数	反演比例 (%)	比例差异 (反演-原始, %)
1	未上过学	581639	4.47	416	4.16	-0.31
2	扫盲班	47253	0.36	0	0	-0.36
3	小学	2301749	17.67	1571	15.71	-1.96
4	***初中***	***4665146***	***35.82***	***5703***	***57.03***	***21.21***
5	高中	2197286	16.87	1342	13.42	-3.45
6	中专	946068	7.26	256	2.56	-4.70
7	大学专科	1029929	7.91	322	3.22	-4.69
8	大学本科	1082289	8.31	377	3.77	-4.54
9	研究生	172631	1.33	13	0.13	-1.20
总计		13023990	100.00	10000	100	

(3) VBC类型，如“AGE”属性，鉴于这类属性与其他属性没有依赖关系，是根据原始统计资料所描述的样本分布确定，因此该属性的统计分布符合表4的分布规律；

(4) VCA类型，如“TRAVEL”属性，反演得到的样本100%符合预先设定的决策树；

(5) VCB类型，本次试验没有涉及，但可以100%符合预先设定的字段之间的依赖关系。

通过对反演的10000个样本的各个属性的验证，其与预先设定的分布类型和关系类型基本一致，符合宏观统计信息，也符合样本各个属性之间的依赖关系。VBB类型的误差最大，但也仅局限于该类属性的个别取值 (如“EDUCATION”属性的“初中”)。

3.4 运行时间

模型运行时间主要依赖于需要反演的样本数量、属性数量及其类型。样本数量和属性数量越多，需要的运行时间越长；样本属性的分布和关系耦合类型越复杂 (如VBB类型)，运行时间越长；样本属性的域元素个数越多，运行时间越长。根据在CPU为3.0GHz*2、内存为4GB的工作站的实际测试，数据反演模型的运行时间总体上可以接受的，本文所开展的实验 (10000个样本、10个属性) 需时33 s，100万个样本需时3505 s (58.4 min)，模型运行时间与样本数量、属性数量的关系如图3所示。

4 结论与讨论

本文提出了一种利用统计数据和相关规则反演个体样本数据的方法 (Agenter)。该方法适合中国国情，充分利用已有的统计信息、典型调查和常规规则等多源数据，反演个体样本的属性信息，结合关于个体样本空间分布的统计性描述，进一步将个体样本体现在空间上。反演结果在统计层次具有较高的准确性，可以作为agent输入多主体系统 (MAS) 进行城市系统微观模拟，是对解决当前城市系统微观模拟中个体数据稀缺问题的有益探索，有望缓解微观模拟在我国乃至国际上应用的数据瓶颈问题。同时本文提出的方法计算的时间也可以接受，100万个样本 (每个样本10个属性字段) 的生成仅需要不足1小时的时间即可完成。

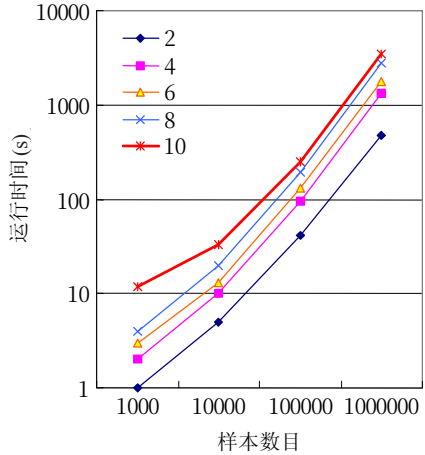

图3 计算时间与样本和字段数量的关系图

Fig. 3 The consumed computation time plot with various samples and the number of fields

采用本方法生成的个体样本数据，因为个体样本的不确定性，即使保持数据的统计特征，但并不是符合已知宏观数据的唯一的一组。利用反演结果作为微观模拟模型输入时，可以利用本方法生成多组的个体样本数据，分别输入模型进行模拟，取模型结果的平均值作为最终的模型输出，有望降低因为输入的数据问题带来的模拟结果的不确定性。

下一个阶段，还将通过数据挖掘的方法，使用实际统计数据对个体属性之间的概率关系，或函数关系进行挖掘，对生成的样本进行验证，进而修正样本，不断提高样本的真实性。同时，个体的自身属性和所对应的空间 (几何) 对象的空间属性 (如面积、周长、边数、朝向、地价、区位、规划条件等) 的关系也需要深入探讨，进而提高个体样本反演的精度。

致谢：本文在写作过程中得到了北京大学资源与环境地理系李昊的宝贵建议，在此表示感谢。

参考文献 (References)

[1] Orcutt G. A new type of socio-economic system. Review of Economics and Statistics, 1957, 58: 773-797.

[2] Ballas D, Clarke G. GIS and microsimulation for local labour market analysis. Computers, Environment and Urban Systems, 2000, 24: 305-330.

[3] Hanaoka K, Clarke G P. Spatial microsimulation modelling for retail market analysis at the small-area level. Computers, Environment and Urban Systems, 2007, 31: 162-187.

[4] Wu B M, Birkin M H, Rees P H. A spatial microsimulation model with student agents. Computers, Environment and Urban Systems, 2008, 32: 440-453.

[5] Parker D C, Manson S M, Janssen M A et al. Multi-agent systems for the simulation of land use and land cover change: A review. Annals of the Association of American Geographers, 2003, 93: 314-337.

[6] Pudney S, Sutherland H. How reliable are microsimulation results? An analysis of the role of sampling error in a U.K. tax-benefit model. Journal of Public Economics, 1994, 53: 327-365.

[7] van Sonsbeek J M, Gradus R H J M. A microsimulation analysis of the 2006 regime change in the Dutch disability scheme. Economic Modelling, 2006, 23: 427-456.

[8] Crooks A, Castle C, Batty M. Key challenges in agent-based modeling for geo-spatial simulation. Computers, Environment and Urban Systems, 2008, 32: 417-430.

[9] Crooks A. Constructing and implementing an agent-based model of residential segregation through vector GIS. CASA Working Paper No.133. Centre for Advanced Spatial Analysis, University College London, 2008.

[10] Miller J E, Hunt D J, Abraham J E et al. Microsimulating urban systems. Computers, Environment and Urban Systems, 2004, 28: 9-44.

[11] Deadman P J, Robinson D T, Moran E et al. Colonist household decision making and land-use change in the Amazon rainforest: An agent-based simulation" Environment and Planning B: Planning and Design, 2004, 31: 693-709.

[12] Zhang Honghui, Zeng Yongnian, Jin Xiaobin et al. Urban land expansion model based on multi-agent system and application. Acta Geographica Sinica, 2008, 63(8): 869-881. [张鸿辉, 曾永年, 金晓斌 等. 多智能体城市土地扩张模型及其应用. 地理学报, 2008, 63(8): 869-881.]

[13] Tao Haiyan, Li Xia, Chen Xiaoxiang. Simulation for evolvement of residential spatial patterns in real scene based on multi-agent. Acta Geographica Sinica, 2009, 64(6): 665-676. [陶海燕, 黎夏, 陈晓翔. 基于多智能体的居住空间格局演变的真实场景模拟. 地理学报, 2009, 64(6): 665-676.]

[14] Shen Z, Yao X, Kawakami M et al. Simulating the impact on downtown of large-scale shopping centre location: Integrating GIS dataset and MAS platform as a case study in Kanazawa city//Proceedings of the Conference of Computers in Urban Planning and Urban Management (Hong Kong), 2009.

[15] Brown D G, Robinson D T. Effects of heterogeneity in residential preferences on an agent-based model of urban sprawl. Ecology and Society, 2006, 11(1): 46. URL: http://www.ecologyandsociety.org/vol11/iss1/art46/

[16] Langford M, Unwin D J. Generating and mapping population density surfaces within a geographical information system. The Cartographic Journal, 1994, 31: 21-26.

[17] Jiang Dong, Wang Naibin, Liu Honghui. Method of pixelizing population data. Acta Geographica Sinica, 2002, 57 (suppl.): 70-75. [江东, 王乃斌, 刘红辉. 人口数据空间化的处理方法. 地理学报, 2002, 57(增刊): 70-75.]

[18] Mennis J. Generating surface models of population using dasymetric mapping. The Professional Geographer, 2003, 55 (1): 31-42.

[19] 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.]
[20] Liao Y, Wang J, Meng B et al. Integration of GP and GA for mapping population distribution. International Journal of Geographical Information Science, 2010, 24: 47-67.
[21] Robinson D T, Brown D. Evaluating the effects of land-use development policies on ex-urban forest cover: An integrated agent-based GIS approach. International Journal of Geographical Information Science, 2009, 23(9): 1211-1232.
[22] Crooks A. Exploring cities using agent-based models and GIS. CASA Working Paper No.109. Centre for Advanced Spatial Analysis, University College London, 2006.
[23] Li X, Liu X. Embedding sustainable development strategies in agent-based models for use as a planning tool. International Journal of Geographical Information Science, 2008, 22: 21-45.
[24] Li X, Liu X. Defining agents' behaviors to simulate complex residential development using multicriteria evaluation. Journal of Environmental Management, 2007, 85: 1063-1075.
[25] Hynes S, Farrelly N, Murphy E et al. Modelling habitat conservation and participation in agri-environmental schemes: A spatial microsimulation approach. Ecological Economics, 2008, 66: 258-269.
[26] Beijing Fifth Population Census Office, Beijing Statistical Bureau. Beijing Population Census of 2000. Beijing: China Statistics Press, 2002. [北京市第五次人口普查办公室, 北京市统计局. 北京市2000年人口普查资料. 北京: 中国统计出版社, 2002.]

Retrieving Individual Attributes from Aggregate Dataset for Urban Micro–simulation: A Preliminary Exploration

LONG Ying[1, 2], SHEN Zhenjiang[3], MAO Qizhi[1]
(1. *School of Architecture*, *Tsinghua University*, *Beijing* 100084, *China*;
2. *Beijing Institute of City Planning*, *Beijing* 100045, *China*;
3. *School of Environment Design*, *Kanazawa University*, *Kanazawa* 920-1192, *Japan*)

Abstract: As the traditional top-bottom based macro-simulation models can not properly adapt to the present research of spatiotemporal dynamic urban system, the bottom-up micro-simulation models using individual data have gradually become a novel perspective for investigating urban systems recently. However, one of the factors restraining the wide application of micro-simulation models is the limited individual data due to the difficulties of data fetching and processing. Such a situation is especially serious in China, where individual dataset is not available from the official census and can only be obtained via various surveys in small scale conducted by separate units. This paper proposes a new solution to retrieve individual dataset from aggregate dataset, e.g. statistical data, for urban micro-simulation under the current sparse-data environment. Based on the existing multi-resource official statistical data, non-official surveys, as well as general relationships among individual attributes, our approach can disaggregate the individual attributions and location obeying the input aggregation observations data. This approach has proved to be significantly accurate at the statistical level, and can be conveniently adopted for urban micro-simulation under the present statistical institutions of China.

Key words: micro-simulation; disaggregation; multi-agent system; statistical data