地理学报 2003年第58卷第1期

基于GIS的青藏高原人口统计数据空间化

作者: 廖顺宝 孙九林

根据2000年第5次全国人口普查数据分析，西藏、青海2省区各市县平均人口密度与海拔高度、土地利用、主要道路有较强的相关关系，河流水系对居民点分布的影响较为明显，而居民点是人口分布的重要指示因子。以GIS软件为工具，通过较为客观的方式赋予各影响因子人口分布影响权重，运用多源数据融合技术进行了人口统计数据的空间化。结果显示，通过数据融合产生的人口密度与各市县实际人口密度的相关系数大于0.80，与试验区各乡镇的实际人口密度的相关系数大于0.75。最终生成的栅格人口密度数据既与各市县统计型人口数据保持一致，又反映了各市县内部人口分布的空间变化。

基于GIS的青藏高原人口统计数据空间化 廖顺宝，孙九林 (中国科学院地理科学与资源研究所，北京 100101) 1 引言人口数据通常是以行政区为单元，通过统计、普查、逐级汇总获得。因此，利用GIS表现人口数据的常规方法是：为统计单元建立多边形边界数据库，把人口作为这些多边形的属性数据进行存储，各种分析和操作均基于一系列统计单元[2]。这种方法给人的感觉是，一个区域内人口是均匀分布的，而实际情况并非如此。这种人口数据处理方法在理论上和实际应用中均存在问题[3]。1994年全球人口制图研讨会达成共识，认为统一的全球栅格人口数据库对于跨学科的研究具有重要价值[4]，这次会议的重要成果是导致了全球栅格人口数据库GPW V1在1995年诞生。但GPW V1和后来的V2在实现人口数据栅格化的过程中，均没有考虑对人口分布产生影响的环境因素，所使用的数据仅包括人口数据和行政边界数据。其它的有关栅格人口密度的研究，在方法上与GPW类似。显然，这种简单的栅格化方法对小区域、样点密集的情况是适合的，当区域扩大，采样稀疏时 (如地广人稀的青藏高原)，计算结果的精度将很难保证。 J. Sweitzer 和 S. Langaas[5]在计算波罗的海 3 国人口密度时，虽然考虑了影响人口分布的地理要素，但在确定各因子对人口分布的影响权重时采用的是主观打分的方式，而且对模型的计算结果没有进行检验。本文选择土地利用、海拔高度、主要道路、河流数据作为人口分布的影响因子，以居民点作为人口分布的重要指示因子，应用多源数据融合技术进行西藏、青海2省区人口数据的空间化，使最终生成的1 km×1 km分辨率栅格人口密度既保持市县层次上与统计数据一致，又反映各市县内部人口分布的空间变化。 2 人口分布与环境关系的宏观分析 2.1 数据源本次研究使用的数据包括：(1) 青藏高原30 ” 分辨率高程数据、(2) 青藏高原1 km×1 km土地利用数据(2000)、(3) 青藏高原1:100万比例尺地形图、(4) 青海、西藏2省区第5次人口普查数据 (分市县) 和 (5) 青海、西藏2省区县级行政区划图及部分乡镇界线图。 2.2 数据预处理数据预处理主要包括以下内容：①在ARC/INFO中将所有空间数据转换成统一的双纬线等面积圆锥投影；②在ENVI中对DEM数据进行投影转换，并重采样成1 km×1 km的空间分辨率；③在ARC/INFO中进行县级行政区划图与土地利用图叠加，提取每个县的各类土地利用面积；④计算各市县不同土地利用类型面积占全县土地总面积的百分比；⑤用人口统计数据和各市县面积计算各县平均人口密度；⑥在ARC/INFO将行政区划图栅格化成1 km×1 km的格网数据，并与数字高程数据进行叠加，计算各县的平均海拔高程；⑦从道路图中提取主要公路、一般公路和乡村道路数据，并与县界图在ARC/INFO中叠加计算各县道路总里程，最后求出各县道路网密度、公路网密度和主要公路网密度；⑧计算居民点与主要道路的距离；⑨计算居民点与河流的距离。 2.3 人口分布与环境因素之间的关系 (1) 人口密度与海拔高度的关系 以 DEM 逐点计算出的各市县平均高程与平均人口密度的相关系数为 -0.33，当把人口密度取对数后，相关系数为-0.53。而在青海，取对数后的相关系数达到-0.86 (图1)。 (2) 人口密度与土地利用之间的关系 据1 km×1 km土地利用图统计出各市县的耕地面积和城镇－工矿用地面积，除拉萨和西宁2个省会城市以外，其余各市县人口密度与耕地比重 (耕地占全县土地总面积百分比) 的相关系数为0.90 (图2)；在2省区113个市县中，人口密度与城镇－工矿用地比重的相关系数为0.85。 (3) 人口密度与主要道路网之间的关系 各市县人口密度与1:100万地形图上全部道路密度、一般公路网密度和主要公路网密度的相关系数分别为0.48、0.66和 0.82。 可见，人口密度与主要公路网密度之间也存在较为密切的关系。 (4) 人口密度与河流的关系 分析发现，各市县平均人口密度与1:100万地形图中的河网密度的关系不明显，但居民点的分布与河流的关系十分密切。 (5) 人口密度与居民点分布密度的关系 除西宁外，在有城市和建制镇分布的市县中，人口密度与市镇居民点密度的相关系数为0.82；在除拉萨和西宁以外的所有市县中，人口密度与乡镇级居民点分布密度的相关系数为0.87，而与村级居民点密度的相关系数则高达0.92。 3 人口统计数据空间化 3.1 人口数据空间化的基本思路从上一节的分析可以看出，各市县的平均人口密度与海拔高程、土地利用、居民点分布、主要道路均有不同程度的相关关系，表明它们是影响该地区人口分布的主要因素，在人口数据空间化的过程中必须加以考虑。但这种相关关系难以定量、定位地在人口数据空间化的过程中直接运用。基于多源数据融合的思想，本文通过如下的思路进行市县级人

口统计数据的空间化： (1) 首先计算出海拔高度、土地利用、道路系统、河流水系决定的居民点密度。 (2) 确定不同类型居民点对人口分布的影响权值，并与第一步的结果进行融合，得到整个区域的人口分布系数图。人口分布系数是空间上人口密度的相对值。 (3) 将人口分布系数图与县界数据融合，得到各市县平均人口分布系数，进而求得各市县的平均人口密度 (基于各影响因子和全区总人口)，并将之与用统计数据计算获得的各市县的平均人口密度进行相关性分析。根据分析结果调整居民点的人口权重系数并重新计算，当二者的相关性达到最高时计算停止。 (4) 将相关性最好 (在市县和乡镇2级) 的融合算法及结果与市县级人口统计数据和市县级行政区划数据进行融合，得到整个区域的人口密度图。 3.2 各影响因子对人口分布影响权值的确定 (1) 海拔高度对居民点分布影响权值的确定 根据DEM的高程值对海拔高度进行分级，每100 m作为一个级差。把分级后的高程数据与居民点分布图叠加，得到不同高程带内居民点的分布数量、面积，从而计算出居民点密度 (表1)。以该值作为海拔高度对居民点分布的影响权值。 (2) 土地利用对居民点分布影响权值的确定 将土地利用图与居民点分布图叠加，统计出各类土地利用类型的面积、分布在其上面的居民点总数，计算出各类土地利用类型中的居民点密度 (表2)。以表2中的居民点密度作为土地利用对居民点分布的影响权值。由于数据的误差，使本不应有居民点分布的水域中出现少量居民点，因此，应用时将水域的权值改为0。 (3) 主要公路对居民点分布影响权重的确定 对研究区主要道路系统每隔10 km建立缓冲区，生成道路缓冲区分布图，把道路缓冲区图与居民点分布图叠加，得到不同距离缓冲区中的居民点数量和缓冲区面积，算出各缓冲区中居民点的密度 (表3)。 (4) 河流水系对居民点分布影响权重的确定 对研究区河流每隔1 km建立缓冲区，生成河流缓冲区分布图，把河流缓冲区图与居民点分布图叠加，得到不同距离缓冲区中居民点的数量，计算出各缓冲区中的居民点密度 (表4)。 (5) 居民点人口权值的确定 (a) 居民点面积的确定 在人口地理学中，城镇人口被看成是点状分布的，乡村人口被看成是面状分布的。进行人口数据的空间化，既要考虑这种特点，但又不能完全被这种思想所束缚。在考虑城镇居民点人口密度时，城镇居民点的面积要适当扩大，农村居民点的面积要适当缩小，这样，既可以避免在城乡结合部出现过大的人口密度差异，又可体现农村居民点对人口分布的影响。因为根据常识，农村居民点及其附近地区的人口密度正常情况下应比两个居民点之间地带的人口密度高。具体计算时，以2 km为半径画圆作为建制镇的面积；以3 km为半径画圆作为县级市的面积；以4 km为半径作为地级市的面积 (居民点面积大小对计算结果无原则上的影响)。对于农村居民点，如果面积过大，将会使大部分居民点连片，难以突出居民点对人口分布的影响，面积过小，又难以突出农村人口分散的特点。通过在ARC/INFO中对农村居民点进行从1 km到5 km的不同半径建立缓冲区发现，随着居民点面积的扩大，连片的居民点数量越来越多。具体变化情况为：当半径为1 km、2 km、3 km、4 km、5 km时，被连片的居民点个数分别为居民点总数的7 %、15 %、25 %、39 %和55 %。由此可见，当半径为5 km时，已有超过一半的居民点与其他居民点合并在了一起。为了既突出农村人口的面状分布、又考虑居民点与非居民点地区人口分布差异的特点，农村居民点的半径定为3 km。 (b) 居民点对人口分布影响权值的确定 ① 农村居民点及远离居民点地区人口分布权值的确定 用2省区总人口减去城镇居民点非农业人口的总和得到2省区的农村人口数，农村人口主要分布在农村居民点及其附近地区，在居民点以外的地区虽然也有分布但分布数量较少。用Ptotal、Ptown、Arural、Drural、Arest、Drest分别表示2省区总人口、城镇非农业人口、农村居民点的总面积、平均人口密度、远离居民点地区的总面积和平均人口密度，则有如下关系式： Ptotal - Ptown = Arural ×Drural + Arest ×Drest (1) 代入有关数据，可求得农村居民点的平均人口密度与远离居民点地区平均人口密度的关系式： Drural = 40.96?C12.27Drest (2) 当Drural = Drest = 3.09 时，农村居民点和远离居民点地区的人口密度相等，根据常识，一般情况下，居民点地区的人口密度要比远离居民点地区的人口密度大，因此Drest 应取一组小于3.09的数据，并求出相应的Drural (表5)。 ② 城镇居民点对人口分布影响权值的确定 用2省区城镇居民点非农业人口除以城镇居民点的面积，得到城镇居民点的人口密度Dtown，Dtown 是随城镇的不同而变化的。用表5中的数据作为农村地区的人口分布权重，用城镇人口密度作为城镇居民点的人口权值，在ARC/INFO中对居民点缓冲区矢量图进行栅格化，并分别用Dtown、Drural、Drest作为城镇居民点、农村居民点和远离居民点区域的栅格属性值，从而得到整个区域的居民点人口分布影响权值图，共8种情况。 3.3 各影响因子与居民点权重数据融合 (1) 单要素与居民点权值的融合 首先分别计算出土地利用类型、海拔高程、主要道路系统、河流水系等因素决定的 (基于1km ×1km栅格) 居民点分布密度相对值rd，然后分别与居民点人口权重指数数据ri叠加，得到人口密度的相对值pdr (= rd ×ri)，将人口密度相对值pdr与县界数据叠加，计算出各市县平均人口密度相对值p。根据研究区的总人口和各市县平均人口密度相对值p，可计算出各市县的平均人口密度d，计算公式为： D = (3) 式中：Di 为第i个市县的平均人口密度，Pt为研究区的总人口，n为研究区市县总数，Pi 为第i 个市县的平均人口密度相对值，Ai为第i个市县的面积。根据上述算法得到的基于土地利用数据、基于数字高程数据、基于主要道路缓冲区数据和基于主要河流缓冲区数据决定的人口密度 (注：调整前均基于方案5的居民点权值，见表5。下同) 与基于统计数据计算出的各市县的实际人口密度的相关系数分别为0.818、0.704、0.790和0.775 (不包括拉萨和西宁)。 (2) 多因子加权融合法 将上述方法得到的4种人口密度 (Pd1，Pd2，Pd3，Pd4) 进行加权融合，如果权重相等，即： Pd = (Pd1 + Pd2 + Pd3 + Pd4)/4 (4) 得到的结果与实际人口密度的相关系数为0.813。如将权重系数分别调整为0.5，0，0.3，0.2或0.6，0，0.2，0.2 则相关系数进一步提高为0.822。由此可见，加权融合可以在一定的程度上提高融合结果与实际人口密度的相关性。 (3) 多因子乘积融合 将基于土地利用数据、数字高程数据、道路缓冲区数据和河流缓冲区数据决定的居民点密度指数LRP、DRP、HRP和RRP之间进行相乘 (本文选择了LRP ×DRP、LRP ×DRP ×HRP、LRP ×DRP ×HRP ×RRP、DRP ×HRP、DRP ×HRP ×RRP和LRP × DRP ×RRP 6种情况)，然后再分别与居民点人口权重相乘，又得到6套人口密度数据，它们与各市县的实际人口密度的相关系数分别为0.784，0.776，0.791，0.729，0.704和0.797。由此可见，乘积融合并没有提高融合结果与实际人口密度的相关性。 (4) 融合结果的进一步验证 上述各种融合结果是在市县级水平上与实际人口密度进行分析和比较，相关系数均在0.7以上。那么，在乡镇级，其相关程度又如何呢？在研究区选择了5个县，共包括68个乡镇。用各乡镇的人口统计数据除以各乡镇的面积得到各乡镇的平均人口密度，形成序列1；通过数据融合的方法计算得到各乡镇的人口密度，形成序列2，序列2共有10组不同的序列 (4种单要素、6种组合因素，排列顺序与计算市县级人口密度时相同)。需要说明的是，在通过数据融合方法计算市县级平均人口密

度时，使用的人口数据是研究区的总人口，而在计算各乡镇的平均人口密度时，所使用的人口数据是乡镇所在县的总人口。通过对序列1和序列2的相关性分析发现，基于统计数据得到的乡镇人口密度与用不同数据融合方法得到的乡镇人口密度的相关系数分别为0.719，0.780，0.700，0.685，0.632，0.618，0.358，0.766，0.582，0.370。 3.4 调整居民点对人口分布的影响权值共设计了8种不同的居民点权值方案（表5），除已经用过的第5种方案外，还有7种方案可供选择。根据调整后的居民点人口权值，再次进行单要素融合和加权融合。每一种居民点权值方案与一种影响因子融合，生成一套人口密度数据，将它们与实际人口密度数据（市县级）进行相关分析（表6）。对每一种居民点权值方案与单要素决定的人口密度进行加权融合（表7）。表7中的第1行是4种单要素决定的市县级人口密度的最优加权融合系数，所谓最优，是指在融合后产生的人口密度与实际人口密度有最高的相关系数（市县级）；第2行是最优加权融合生成的人口密度与实际人口密度（市县级）的相关系数；第3行是按照该加权方案融合生成的乡镇级人口密度与实际乡镇人口密度的相关系数。从表7可以看出，虽然对于全部8种居民点权值方案，4种单要素决定的人口密度加权后与实际人口密度在市县级的相关系数均能达到0.80以上，但前3种方案产生的人口密度在乡镇级与实际人口密度的相关系数均在0.70以下，第4种也刚刚大于0.70，因此，实际上只有后4种居民点人口权值方案可以使用。 3.5 研究区人口密度图的生成 (1) 人口密度的计算 设空间任意一个栅格i的人口密度系数为Wi，X为一常数（对同一个市县），WiX为该栅格点的人口密度，Tp为某市县的总人口，n为某市县的栅格数，则人口密度为：WX = (5) 式中的Wi，Tp可在ARC/INFO的GRID模块中通过一系列运算求得，从而计算出整区域的栅格人口密度。 (2) 人口密度图的平滑 所求得的人口密度图在相邻栅格间的值相差很大，几乎为完全离散状态，因此需要对其进行平滑处理。图像的平滑可在ARC/INFO的GRID模块中进行，也可在专门的图像处理软件（如：ENVI）中完成。图3是经过平滑和分级处理以后的西藏青海2省区1 km ×1 km分辨率的人口密度图。 4 结论本文以土地利用、海拔高度、主要道路和河流作为影响青藏高原人口分布的主要环境因子，以居民点信息作为人口分布的指示因子，通过较为客观的方式赋予各影响因子人口分布影响权重，运用多源数据融合技术进行了人口统计数据的空间化。在设计的8种居民点人口权值方案中，有4种方案使数据融合产生的人口密度与实际人口密度在市县级的相关系数大于0.80，在乡镇级的相关系数大于0.75，而其中又以第8种方案最好（2个级别上的相关系数均最高），说明在青藏高原地区，居民点对人口分布有重要影响。 高程、主要河流对人口的分布具有重要影响。在海拔2000~5000 m范围内分布的人口为661×104人，占2省区总人口的88.90 %，而在2000 m以下和5000 m以上地区分布的人口仅占总人口的11.10 %；在距河流0~1 km、1~2 km、2~3 km的缓冲区内分布的人口分别为324×104人、204×104人和118×104人，分别占2省区总人口的43.50 %、27.45 % 和15.80 %，即：在距河流3 km以内的缓冲区中分布的人口占总人口的86.75 %，在3 km以外地区分布的人口仅占总人口的13.25 %。土地利用、主要公路与人口的分布也密切相关。分布在耕地、林地和草地3种土地利用类型中的人口分别为181×104人、43×104人和433×104人，分别占总人口的24.35 %、5.80 % 和58.17 %，大量人口分布在草地类型中，说明研究区以畜牧业为主。但就人口密度而言，耕地中的人口密度分别为林地、草地中人口密度的68倍和63.7倍；在距主要公路0~10 km、10~50 km、大于50 km的缓冲区中分布的人口为470×104人、187×104人和87×104人，分别占总人口的63.19 %、25.12 % 和11.69 %。 致谢：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了1km×1km分辨率土地利用数据，刘闯研究员提供了研究所需要的部分空间数据，李泽辉副研究员提供了研究中使用的人口数据，郭连保高级工程师完成了部分乡镇界线数据的采集，向世芳女士完成了部分乡镇人口数据的录入，对他们提供的支持和帮助表示衷心的感谢！ GIS Based Spatialization of Population Census Data in Qinghai-Tibet Plateau LIAO Shunbao, SUN Jiulin (Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China) Abstract: In the Qinghai-Tibet Plateau, correlation ratios between population density and percentages of arable land and city and town used land at county level reach 0.90 and 0.85 respectively. In Qinghai province, there exists a logarithmic correlation of ratio = -0.86 between population density and average territorial elevation at county level. There is a correlation of ratio = 0.82 between population density and main highway density at county level. Correlation ratios between population density and densities of city and town residential areas, township residential areas and village residential areas reach 0.82, 0.87 and 0.92 respectively. Density of residential areas drops along with increasing of distance to rivers. Therefore, territorial elevation, land use, road and river system are the main factors affecting distribution of population in the Qinghai-Tibet Plateau. Residential areas are an important indicator to distribution of population. Weight values of affecting factors are assigned objectively and multiple sources data fusion technology is applied to spatialize population census data. There is a correlation of ratio > 0.80 between the population density generated by data fusion and actual population at county level, and ratio > 0.75 at township level. The finally generated grid population density not only keeps consistence with statistical population data at county level but also reflects changes of population distribution inside each county. Key words: Qinghai-Tibet Plateau; population; spatialization

关键词： 青藏高原；人口；统计数据；空间化

