浙 江 林 学 院 学 报 2009，**26**(4)：587－591
Journal of Zhejiang Forestry College

基于 Map X 的空间数据不一致性问题处理

张广群[1]，王保平[2]，汪杭军[1]

（1. 浙江林学院 信息工程学院，浙江 临安 311300；2. 南阳师范学院 计算机与信息技术学院，河南 南阳 473061）

摘要：数据预处理是数据挖掘过程中很重要的环节。由于复杂的空间数据更易造成数据的不一致，加上不同应用空间数据的特殊性，对空间数据的预处理往往需要采用特殊的方法。针对林业小班空间数据在数字化过程中产生的空间对象位置的不一致性问题，给出了一种解决的方法：以基准面为基础，通过最外 2 个交点，在所有的交点上进行修正。实际空间数据对比实验结果显示该方法比节点抓取具有更好的处理效果。图 7 参 10

关键词：森林经理学；空间数据预处理；图形修正；小班；节点抓取

中图分类号：S757　　**文献标志码**：A　　**文章编号**：1000-5692(2009)04-0587-05

MapX-based processing of spatial data inconsistent problems

ZHANG Guang-qun[1]，WANG Bao-ping[2]，WANG Hang-jun[1]

（1. School of Information Engineering，Zhejiang Forestry College，Lin'an 311300，Zhejiang，China；2. School of Computer and Information Technology，Nanyang Normal University，Nanyang 473061，Henan，China）

Abstract：Data pre-processing is an important step in data mining. Special method is usually used in spatial data pre-processing because complex spatial data easily bring about the case of inconformity and spatial data in different application has its own characteristic. In this paper，an effective method was presented to solve the problem of inconsistent spatial object's location which was produced in the process of sub-compartment digitalization. The method was based on a referent graphic surface through the two outermost points of intersection，and the subcompartment could be amended at the all intersection points. Compared with conventional methods (point capture)，the proposed approach had a better effect. [Ch，7 fig. 10 ref.]

Key words：forest management；spatial data pre-processing；figure adjustment；subcompartment；point capture

滥用缩写词、数据输入错误、数据中的内嵌控制信息、不同的惯用语、重复记录、丢失值、拼写变化、不同的计量单位和过时的编码等原因会造成数据库中数据出现不完整，并含有噪声和不一致的情况。数据预处理就是要解决这些问题，它是数据挖掘过程中一个很重要的环节[1]。空间对象包括非空间和空间 2 类属性，其中非空间属性刻画对象的非空间特性；空间属性是定义空间对象的位置和范围的[2]。由于空间对象包含点、线和面，其空间属性经常包括与空间位置有关的信息，如经度、纬度、海拔和形状等，空间数据比经典数据挖掘中的数据要复杂的多[3-4]。因此，空间数据中经常会出现数据不一致的现象。数据预处理一般分为 4 个步骤：数据选取、数据表属性一致化、数据清理和数据离散化[5]。但是在不同的应用中，产生数据不一致的原因各有不同，往往需要针对具体情况作一些特殊处理[6-9]。笔者首先提出针对林业小班空间数据的数字化过程中存在不同空间对象位置的不一致

收稿日期：2008-09-03；修回日期：2008-12-26

基金项目：浙江省自然科学基金资助项目(Y3080457)；浙江省科技计划重点项目(2008C21087)；浙江省林业厅资助项目(07A14)

作者简介：张广群，讲师，硕士，从事数据仓库与数据挖掘方向研究。E-mail：gloria@zjfc.edu.cn。通信作者：汪杭军，副教授，硕士，从事数据挖掘和智能信息处理等研究。E-mail：whj@zjfc.edu.cn

性问题，然后给出了一种解决的方案。通过对实际的空间数据进行对比实验。结果表明，该方案比节点抓取具有更好的结果。

1 林业空间数据不一致问题

图 1 是采用 MapInfo Professional 管理的某市林业小班数据。该图是通过森林分布图采用手工方式数字化后录入到 MapInfo 系统中的。从图中圈内所标记的小班部分可以发现，原本相切的小班由于数字化不当的原因造成了相交。这必将影响到今后基于该小班地图进行各种计算和分析的正确性和可靠性。因此，必须对空间数据进行预处理，解决空间数据的不一致性问题。经过对原始图观察和分析，数据不一致的主要是图 2 所示的 2 种情况，其中上图是下图的一种简单情况，左图是右图的数字化后的原始图，右图是左图的实际小班情况。

图 1 数字化后原始图

Figure 1 Digital original picture

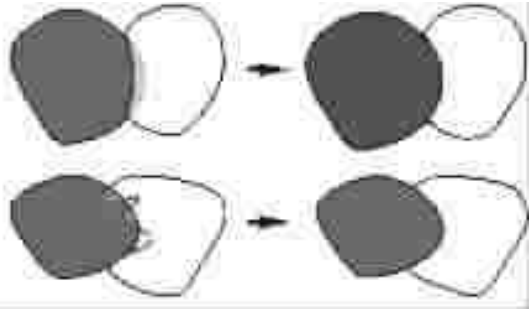

图 2 不一致小班

Figure 2 Inconsistent subcompartment

在 MapInfo Professional 6.5 后的版本中新增加了节点抓取功能，它能从不同对象中抓取节点集合在一起，并在保持对象的图形形状时，减少一个对象中节点的数量[10]。利用节点抓取可以较好解决简单情况的不一致小班问题，但对于具有多个相交点的相邻小班则会出现。在图 3 中，右图是节点抓取处理后的结果，这里 2 个小班出现了重叠和空隙区域。另外，应用节点抓取需要设置节点间的公差，这对给定的区域中存在面积大小不均的小班设置统一的公差是非常困难的。由于小班实际的数字化过程中经常会产生第 2 种不一致的复杂情形，一个小班对象的边界要根据相交的另一个小班对象进行调整。我们需要采用新的空间数据预处理方法，解决这一类不一致小班问题，以提高后续分析的正确性。

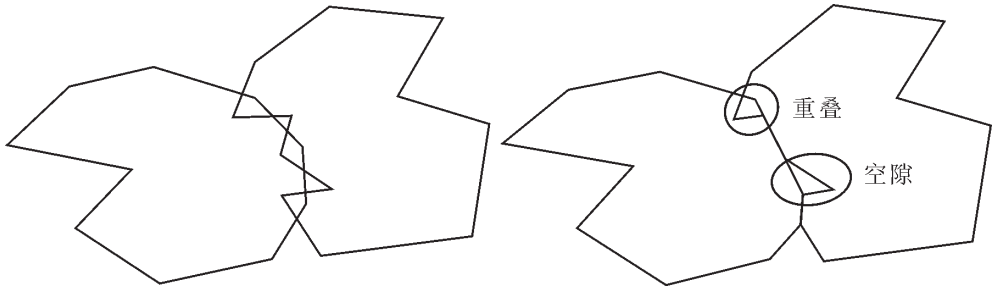

图 3 节点抓取处理效果图

Figure 3 Drawing of point capture processed

2 基于 Map X 的解决方法

在 MapInfo 中每个小班是由若干个点的集合所组成的面对象，每个相邻的点间由一条直线相连。根据这个特点，我们用一个包含小班所有线段的集合来标识小班(以下称图形)。我们提出 2 个图形之

间的修正算法如下：

输入：2 个图形的线段集合 *Sa* 和 *Sb*。

输出：被修正后的图形 *Sb*。

Step 1：求出 2 个图形的所有交点。

Step 2：求出“最外面”2 个交点(如图 4 中的 *a* 和 *b* 2 个点)。

Step 3：以某一个小班为母版，修正另一个图形。

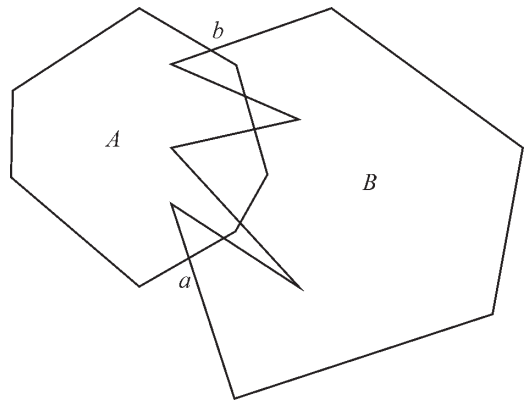

图 4　2 个图形间交点

Figure 4　Intersecting point of two picture

2.1　求给定 2 个图形的所有交点

输入：2 个图形的线段集合 *Sa* 和 *Sb*。

输出：2 个图形的所有交点集合 Point。

Step 1：求出 *Sa* 集合中所有线段的定义域，并取它们的并集构成 *Sa* 的定义域 *Da*，同理求出 *Sb* 的定义域 *Db*；这样可以求出 *Sa* 和 *Sb* 的公共定义域 $Dab = Da \cap Db$。

Step 2：根据它们的公共定义域分别求出 *Sa*，*Sb* 中定义域与 *Dab* 的交集不为空的线段的集合 *Ssa* 和 *Ssb*。

Step 3：从 *Ssa*(或者 *Ssb*)选择一条线段 *La*，假设其定义域为 *DLa*，从 *Ssb* 中选择出定义域与 *DLa* 有交集的线段，求出交点并放入交点集合 Point 中。

Step 4：重复 Step 3 直到 *Ssa*(或者 *Ssb*)中的所有线段都被测试过。

Step 5：经过上述步骤得到点集合 Point 就是 2 个图形的所有交点的集合。

上面给定的求 2 个相交图形所有交点的算法，考虑到 2 个图形相交大多数的线段根本不相交或者只与其中的一部分线段相交，因此若将不可能相交的线排除就可大大减少求 2 条根本不相交直线交点所作的额外工作，提高算法的效率。

2.2　求“最外面”交点

得到了 2 个图形的所有交点后，接下来我们要获取“最外面”的 2 个交点。这里通过以下 4 个步骤来完成：

输入：2 个图形的线段集合 Sa 和 Sb；2 个图形的所有交点集合 Point。

输出：2 个图形最外面的 2 个交点 Point 0 和 Point 1。

Step 1：从点集合 Point 中任意选取一点赋值给 *p*0(如图 5 中的 *c* 点)。

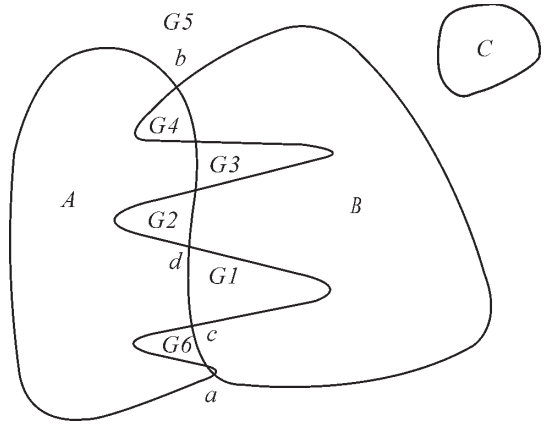

图 5　“最外面”交点示意图

Figure 5　Picture of outermost intersecting point

Step 2：从 *p*0 点开始按 *A* 图形逆时针方向，*B* 图形顺时针方向选取一个相邻的交点 *p*1(图 5 中的 *d* 点)，构造图形 *G*（如图 5 中的图形 *G*1，由 *c* 和 *d* 间 2 个小班边界所围封闭区域，下同），并检测图形是否包含不属于 *A*，*B* 的另外一个图形(如图 5 中的 *C* 图形)。

Step 3：如果包含则将构成图形 *G* 的 2 个交点取出这 2 个点即为需要寻找的最外面的 2 个交点。如果没有包含，则从下一个点(如图 5 中的 *d* 点)重复 Step 2，直到找到一个包含图形 *C* 的图形 *G*(如图 5 中的 *G*5)。

Step 4：将找到的两点放到 Point 0 和 Point 1 中。

2.3　图形修正

通过上面 2 步，我们得到了图形 *A*，*B* 最外面的 2 个交点 Point 0 和 Point 1 和以这 2 个交点为起

点和终点的图形 G(构造方法见 2.2)，下面就是对图形进行修正的基本算法。

输入：原始的图形 Sa 和 Sb；2 个最外面的交点 Point 1 和 Point 2 和由这 2 个点所组成的图形 G；

输出：被修正后的图形 Sb.

Step 1：从 Sa 中将属于 G 的线段删除并将剩余的线段集合存入新的线段集合 TA。

Step 2：从 Sb 中将属于 G 的线段选出并存入新的线段集合 $TB = SB \cap G$。

Step 3：根据 Point 1 和 Point 2 的相交情况，相应的边放入到 TB 中。

3 实验结果与分析

根据上面提出的算法，我们采用 VC++6.0 和 Map X 工具进行了实现。图 6 是利用我们提出的新方法对图 3 重新进行处理的结果。从图中可以发现，该算法对不一致小班能进行较好的修正，解决了相邻小班间重叠和空隙的问题。

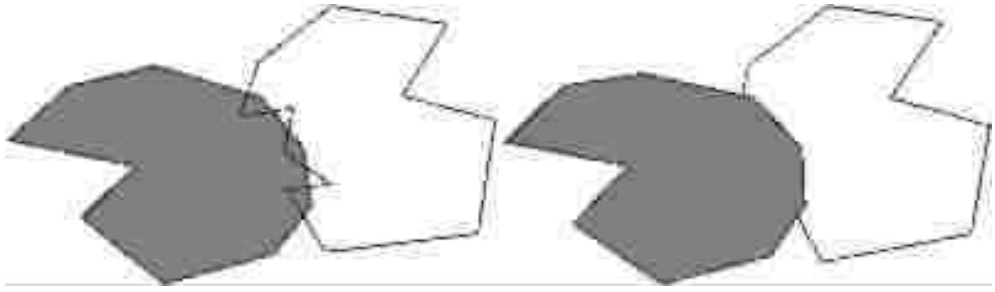

图 6 新的不一致小班处理结果图

Figure 6 New processed drawing of inconsistent subcompartment

为了将该算法应用于小班空间数据预处理，需要对 3 个及以上的相交图形进行修正。本算法在处理更多的图形情况下具有通用性。在 3 个图形相切的时候，可取 2 个图形作为参考面，另外的剩下的那个作为修正面。依次类推，当有 N 个图形相切的时候，取 N–1 个图形作为参考面并把它们的所有线段并到集合 Sa 中，而剩下的那个图形作为修正面将其线段集合作为 Sb，进行修正。应用该方法，我们对图 1 的原始小班图进行了处理，结果如图 7 所示。从图中可以发现，对于原始图中小班相切都已经进行了有效的处理，特别注意图 1 中圈起的部分。

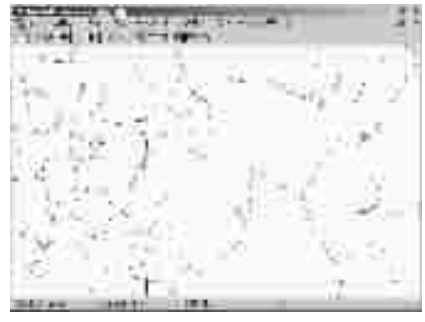

图 7 对图 1 小班处理图

Figure 7 processed picture to Figure 1

4 总结

小班数据的质量将直接影响到基于该空间数据的进一步处理和分析，特别是空间数据挖掘。由于在小班数字化过程中存在着空间问题的不一致性，尤其是相切的小班被数字化为相交的情况，因此，需要在建模前进行数据准备工作——空间数据预处理，一方面保证建模数据的正确性和有效性，另一方面通过对数据格式和内容的调整，使数据更符合建模的需要。

本文针对小班数字化过程中存在的不一致问题提出了一种解决方案，通过与节点抓取方法的实验对比效果来看，该方法能够在小班间复杂的相交情况下比节点抓取方法能得到更好的预处理效果。使用本文提出的算法对某市的森林小班数字化原始图进行处理后得到的效果也是比较理想。

参考文献：

[1] 菅志刚，金旭. 数据挖掘中数据预处理的研究与实现[J]. 计算机应用研究，2004，**21**（7）：117 – 118.
JIAN Zhigang，JIN Xu. Research on data preprocess in data mining and its application [J]. *Appl Res Comp*，2004，**21**（7）：117 – 118.

[2] 余慧，张曙光，刘英，等. 空间对象及其拓扑关系[J]. 计算机工程与应用，2004（6）：77 – 79.
YU Hui，ZHANG Shuguang，LIU Ying，*et al*. Spatial object and topological relation [J]. *Comp Eng Appl*，2004（6）：77 – 79.

[3] 汪杭军，方陆明，张广群. 空间分析中的空间数据结构[J]. 浙江林学院学报，2007，**24**（3）：363 – 368.
WANG Hangjun，FANG Luming，ZHANG Guangqun. Spatial data structure in spatial analysis [J]. *J Zhejiang For Coll*，2007，**24**（3）：363 – 368.

[4] 巩华荣，何佳. 空间数据挖掘技术的研究与发展[J]. 测绘与空间地理信息，2007，**30**（5）：81 – 84.
GONG Huarong，HE Jia. Research and development of spatial data mining technique [J]. *Geomat & Spat Inform Technol*，2007，**30**（5）：81 – 84.

[5] 韩家炜. 数据挖掘：概念与技术[M]. 北京：机械工业出版社，2001.

[6] 刘博，彭宏，郑启伦. 一种新的数据预处理算法——NLCA[J]. 计算机应用，2006，**26**（6）：1406 – 1408.
LIU Bo，PENG Hong，ZHENG Qilun. Pretreatment method of data mining based on non-linear correlation analysis [J]. *J Comp Appl*，2006，**26**（6）：1406 – 1408.

[7] 李芳玉，潘懋. GIS 数据预处理分析及若干算法的研究[J]. 计算机工程与应用，2004，**40**（1）：54 – 55.
LI Fangyu，PAN Mao. Research on some algorithms in data preprocessing of GIS [J]. *Comp Eng & Appl*，2004，**40**（1）：54 – 55.

[8] 孙庆辉，池天河，赵军喜，等. 空间数据处理模型误差和不确定性分析[J]. 测绘科学技术学报，2007，**24**（1）：33 – 36.
SUN Qinghui，CHI Tianhe，ZHAO Junxi，*et al*. Errors and uncertainties analysis of spatial data processing model [J]. *J Zhengzhou Inst Surv Mapp*，2007，**24**（1）：33 – 36.

[9] 付建德，高莉，张海印. 城市基础地理空间数据建库中的质量控制研究[J]. 测绘科学，2005，**30**（6）：67 – 73.
FU Jiande，GAO Li，ZHANG Haiyin. Quality control methods on building city spatial databases [J]. *Sci Surv Mapp*，2005，**30**（6）：67 – 73.

[10] 罗云启，曾琨. GIS 数字化地理信息系统建设与 MapInfo 高级应用[M]. 北京：清华大学出版社，2003.