

人 博 局

申请学位级别: 工程硕士 领域名称: 计算机技术

校内导师姓名: 徐 泮 林 职 称: 教 授

校外导师姓名: 王 瑞 波 职 称: 高 工

山 东 科 技 大 学

二 零 零 四 年 五 月

The study of attribute data quality

A Dissertation submitted in fulfillment of the requirements of the degree of

MASTER OF ENGINEERING

from

Shandong University of Science and Technology

by

Guan Haiying

Supervisor: Professor: Panlin Xu

Associate researcher: Ruibo Wang

College of Information Science and Technology

May, 2004

声 明

本人呈交给山东科技大学的这篇硕士学位论文，除了所列参考文献和世所公认的文献外，全部是本人在导师指导下的研究成果。该论文资料尚没有呈交于其它任何学术机关作鉴定。

硕士生签名: 关海鹰

日 期: 2004.5.25

AFFIRMATION

I declare that this dissertation, submitted in fulfillment of the requirements for the award of Doctor of Philosophy in Shandong University of Science and Technology, is wholly my own work unless referenced or acknowledge. The document has not been submitted for qualification at any other academic institute.

Signature: Guanhaiying

Date: 2004.5.25

摘要

当前信息化社会正在深刻地影响人类生活的各个领域，而地理信息系统是信息化社会的骨架支撑技术之一，关系到国民经济建设、社会发展和国家安全。在地理信息系统中，数据质量的好坏都直接影响到地理信息系统的成败。

近年来，GIS 数据质量的研究集中在位置数据的误差上，而对属性数据误差的研究较少，多年来的研究主要集中在遥感分类属性数据的精度分析和一些新方法的引入上，而关系数据库存储的属性信息质量方面的研究相对比较薄弱。

为了给出一个 GIS 属性数据质量评价的基本模型，结合已有的数据质量理论，本文对属性数据的质量控制与评价进行了深入分析与研究，给出了 GIS 属性数据质量控制的有效措施，并提出了提高属性数据存储、操作精度的方法。在本文中，首次提出了根据评定项权分配评定项总分的属性数据质量综合评价模型。针对存储多类属性数据的大型数据库，本文首次提出了由各类的数据量在总数据量中所占比重与重要性来确定各类的权，然后根据各类的权分配各类所占总分及抽样数量的属性数据质量分类评价模型。

本文将上述理论应用到山东省 1:25 万地名数据库中，对该库属性数据进行了质量控制与评价，提出了一套切实可行的属性数据质量控制措施。应用上述评价理论对山东省 1:25 万地名数据库属性数据质量进行了分类评价，取得了良好的效果。

关键词：质量控制 综合评价 分类评价 权

ABSTRACT

Recently, information society is leading various directions of human life. Geographic information system is one of the framework technologies. It deals with national economy, development of society and security of our nation. In GIS, data quality is one of the key aspects.

The Recent studies of data quality in GIS most focus in the error of spatial data. The error study on attribute data is very weak. Most studies centralize in precision analysis of attribute data from remote sensing and introduction of new method. The studies of attribute information quality stored in RDB are very weak.

To design a basic evaluation model of attribute data in GIS, using various theory of data quality, this paper makes a deep analysis of quality control and evaluation of attribute data and brings forward a effective measure against error in attribute data. A good method is presented to improve precision of attribute data in this paper. In this paper, we present a compositive evaluation model that could distribute total score of each item according to its own power for the first time. For large database that can store various class data, this paper presents a an evaluation model that could compute power of each class according to its own proportion and weightiness in total data, and then distributes total score of each class according to its own power.

This paper applies above theory to 1/250000 toponym database of Shandong province, executes control and evaluation on attribute data and brings forward a feasible measure of quality control. After the evaluation in toponym database based on above theory, we get a good result.

Keywords: quality control, compositive evaluation, classification evaluation, power

目录

1 绪论	1
1.1 数据质量控制的重要性	1
1.2 属性数据质量控制的重要性	3
1.3 属性数据质量控制的研究现状与存在问题	4
1.4 课题的提出	5
1.5 论文的主要内容及结构	7
2 属性数据质量控制的特点与发展	9
2.1 GIS 属性数据管理	9
2.2 GIS 属性数据误差的起因与表示	13
2.3 研究 GIS 属性数据质量的主要理论	15
2.4 研究 GIS 属性数据质量的方法与标准	16
3 GIS 中属性数据质量控制方案设计	19
3.1 GIS 属性数据质量控制指标、过程与实施	19
3.2 GIS 属性数据精度评价模型设计	22
3.3 提高 GIS 属性数据精度的方法	30
4 山东省二十五万分之一地名库中属性数据的质量控制与评价	36
4.1 山东省 1:25 万地名数据库的设计	36
4.2 山东省 1:25 万地名数据库建立	37
4.3 山东省 1:25 万地名数据库属性数据质量控制	40
4.4 山东省 1:25 万地名数据库中属性数据质量评价	47
结论与建议	67

Contents

1 Introduction	1
1.1 The importance of data quality control	1
1.2 The importance of attribute data quality control	3
1.3 The study and problem of attribute data quality control	4
1.4 The question for discussion	5
1.5 The main content and configuration of this thesis	7
2 The development of attribute data quality control	9
2.1 The management of attribute data in GIS	9
2.2 The origin and expression of error of attribute data in GIS	13
2.3 The main theory of attribute data quality in GIS	15
2.4 The method and standard for the study of attribute data quality in GIS	16
3 The design of attribute data quality control scheme in GIS	19
3.1 The target, process, and implement of attribute data quality control	19
3.2 The design of evaluation model of attribute data precision	22
3.3 Some way to improve precision of attribute data	30
4 The control and evaluation of attribute data quality in 1/250000 toponym database of Shandong province	36
4.1 The design of 1/250000 toponym database of Shandong province	36
4.2 The establishment of 1/250000 toponym database of Shandong province	37
4.3 The quality control of 1/250000 toponym database of Shandong province	40
4.4 The quality evaluation of 1/250000 toponym database of Shandong province	47
Conclusion and advice	67

15
分类号: P208

U D C :

密 级: 公 开

单位代码: 10424

工 程 硕 士 学 位 论 文 详 细 摘 要

属性数据的质量问题探讨

关 海 鹰

申请学位级别: 工程硕士 领域名称: 计算机技术

校内导师姓名: 徐 泮 林 职 称: 教 授

校外导师姓名: 王 瑞 波 职 称: 高 工

山 东 科 技 大 学

二 零 零 四 年 五 月

论文题目：

属性数据的质量问题探讨

作者姓名：关海鹰

入学时间：2002年5月

领域名称：计算机技术

研究方向：地理信息系统

校内导师：徐 泮 林

职 称：教 授

校外导师：王 瑞 波

职 称：高 工

论文提交日期：2004年5月

论文答辩日期：2004年6月

授予学位日期：

The study of attribute data quality

A Dissertation submitted in fulfillment of the requirements of the degree of

MASTER OF ENGINEERING

from

Shandong University of Science and Technology

by

Guan Haiying

Supervisor: Professor: Panlin Xu

Associate researcher: Ruibo Wang

College of Information Science and Technology

May, 2004

摘要

当前信息化社会正在深刻地影响人类生活的各个领域，而地理信息系统是信息化社会的骨架支撑技术之一，关系到国民经济建设、社会发展和国家安全。在地理信息系统中，数据质量的好坏都直接影响到地理信息系统的成败。

近年来，GIS 数据质量的研究集中在位置数据的误差上，而对属性数据误差的研究较少，多年来的研究主要集中在遥感分类属性数据的精度分析和一些新方法的引入上，而关系数据库存储的属性信息质量方面的研究相对比较薄弱。

为了给出一个 GIS 属性数据质量评价的基本模型，结合已有的数据质量理论，本文对属性数据的质量控制与评价进行了深入分析与研究，给出了 GIS 属性数据质量控制的有效措施，并提出了提高属性数据存储、操作精度的方法。在本文中，首次提出了根据评定项权分配评定项总分的属性数据质量综合评价模型。针对存储多类属性数据的大型数据库，本文首次提出了由各类的数据量在总数据量中所占比重与重要性来确定各类的权，然后根据各类的权分配各类所占总分及抽样数量的属性数据质量分类评价模型。

本文将上述理论应用到山东省 1:25 万地名数据库中，对该库属性数据进行了质量控制与评价，提出了一套切实可行的属性数据质量控制措施。应用上述评价理论对山东省 1:25 万地名数据库属性数据质量进行了分类评价，取得了良好的效果。

关键词：质量控制 综合评价 分类评价 权

ABSTRACT

Recently, information society is leading various directions of human life. Geographic information system is one of the framework technologies. It deals with national economy, development of society and security of our nation. In GIS, data quality is one of the key aspects.

The Recent studies of data quality in GIS most focus in the error of spatial data. The error study on attribute data is very weak. Most studies centralize in precision analysis of attribute data from remote sensing and introduction of new method. The studies of attribute information quality stored in RDB are very weak.

To design a basic evaluation model of attribute data in GIS, using various theory of data quality, this paper makes a deep analysis of quality control and evaluation of attribute data and brings forward a effective measure against error in attribute data. A good method is presented to improve precision of attribute data in this paper. In this paper, we present a compositive evaluation model that could distribute total score of each item according to its own power for the first time. For large database that can store various class data, this paper presents a an evaluation model that could compute power of each class according to its own proportion and weightiness in total data, and then distributes total score of each class according to its own power.

This paper applies above theory to 1/250000 toponym database of Shandong province, executes control and evaluation on attribute data and brings forward a feasible measure of quality control. After the evaluation in toponym database based on above theory, we get a good result.

Keywords: quality control, compositive evaluation, classification evaluation, power

详细摘要

近年来, GIS 数据质量的研究多集中在位置数据的误差上, 而对属性数据误差的研究较少。为此, 本文对 GIS 属性数据质量控制措施与评价模型进行了深入探讨。

(一) 评价模型

(1) GIS 属性数据质量综合评价模型:

①将属性精度评价项构成集合:

$$C = \{C_1, C_2, C_3, \dots, C_n\}$$

②综合考虑各评价项反映属性数据质量的能力以及对总体的代表性, 确定各评价项权 $Q = \{Q_1, Q_2, Q_3, \dots, Q_n\}$

③对 C 的每一项按照②中的权进行所占总分 T 的分配:

$$\text{各评价项总分 } T = \{T_1, T_2, T_3, \dots, T_n\} = 100 * Q = 100 * \{Q_1, Q_2, Q_3, \dots, Q_n\}$$

④按照评价规则, 在总分 T 的基础上进行 C 中每一评价项的得分评价, 得出评价项集合 C 的每一项内容得分 S 为: $S = \{S_1, S_2, S_3, \dots, S_n\}$

⑤将每一项得分进行类加, 即得到百分制属性数据总体质量得分 B :

$$B = \sum_{i=1}^n S_i = S_1 + S_2 + S_3 + \dots + S_n$$

(2) GIS 属性数据质量分类评价模型

对于大批量的检验总体, 有时总体各类质量的实际分布差别很大, 采用不同的检验程序就比较符合实际。实现方法如下:

① 依据属性数据内容与表达方式将属性数据分成 n 类, 用集合 C 表示:

$$C = \{C_1, C_2, C_3, \dots, C_n\}$$

② 依据重要性、在总体中所占比例以及代表性等对各类属性数据赋予一个权 W :

$$W = (W_1, W_2, \dots, W_n) \quad \text{且} \quad \sum_{i=1}^n W_i = 1$$

③ 确定各类属性数据的抽样数量

为了获得各类属性数据合理的抽样数量，往往根据各类分配的权确定抽样比：若各类属性数据的权为 W ，则各类属性数据的抽样数量比 $\bar{S}$ 为：

$$\bar{S} = W_1 : W_2 : \cdots : W_n$$

然后根据评价费用以及结果可信度要求等确定样本总量 E ，按上述比例分配到各类属性数据。各类属性数据抽样数量为 $S = (S_1, S_2, \cdots, S_n)$ ，

$$S_i = E * \frac{W_i}{\sum_{k=1}^n W_k} = E * \frac{W_i}{W_1 + W_2 + \cdots + W_n} \quad i = (1, 2, \cdots, n)$$

④ 分别对 n 类属性数据依据相关评价标准进行属性数据质量评价，得到各类得分 U ：

$$U = (U_1, U_2, \cdots, U_n)$$

⑤ 对各类属性数据得分进行加权综合，得到总体质量得分为：

$$\begin{aligned} S &= W \times U^T = (W_1, W_2, \cdots, W_n) \times (U_1, U_2, \cdots, U_n)^T \\ &= \sum_{i=1}^n W_i * U_i = W_1 * U_1 + W_2 * U_2 + \cdots + W_n * U_n \end{aligned}$$

这里 W_i 是第 i 类的权：

(二) 属性数据精度提高

- (1) 采用定性概率表达方式提高属性数据精度。
- (2) 采用定性概率计算方式提高操作后的属性数据精度。

(三) 山东省二十五万分之一地名库中属性数据的质量控制与评价

(1) 属性数据质量控制

要保证属性数据的质量，必须从以下几方面考虑：数据源，数据格式，属性输出质量等近 12 项。

(2) 山东省 1:25 万地名数据库中属性数据的分类质量评价

评价方案：

① 确定属性数据分类

表 1 属性数据分类

Table 1 The classification of attribute data

类别	居民地	工业设施	农业设施	行政区划	道路	植被	水系	其他
----	-----	------	------	------	----	----	----	----

② 确定属性数据评价项及其在不同类中的权

表 2 属性数据评价项

Table 2 The inspection items of attribute data

评价项	数据 格式	完备性	唯一性	连接列 关键字	属性输 出质量
评价项	元数据 质量	数据转 换无损	属性数据 定义	描述信息 正确性	表格 正确性

由于在不同类的属性数据中，评价项的重要性、代表性和所处地位不同，依据各类特点与评价项对本类的重要性来确定评价项在各类属性数据中的权，更能提高评价质量。

③ 确定不同类中属性数据评价项所占分值

假设共有 m 类和 n 个评价项，则：

评价项 P_i 在类 C_j 中所占分值为 T_{C_j-i} ： $T_{C_j-i} = 100 * W_{C_j-i}$

$$W = \{\{W_{C1-1}, W_{C1-2}, \cdots, W_{C1-n}\}, \{W_{C2-1}, W_{C2-2}, \cdots, W_{C2-n}\}, \cdots, \{W_{Cm-1}, W_{Cm-2}, \cdots, W_{Cm-n}\}\}$$

W 为评价项在各类属性数据中的权集合，其中 W_{C_j-i} 为第 P_i 个评价项在 C_j 类中的权。

在山东省 1:25 万地名数据库中，依据上面分类，我们取 $m=18$ ， $n=10$ ，按照上面公式计算各评价项在不同类中所占分值。

④ 确定属性数据评价项的评分标准

对属性数据每一评价项进行评分时，必须具有一定的标准。对每一个评价项采用 A、B、C、D、E 五等进行评分。

⑤ 属性数据分类评价

根据各类属性数据的权，进行地名数据库属性数据质量分类评定，得分如下：

表 3 各类属性数据的权与得分

Table 3 The power and score of various attribute datas

类别	居民地	工业设施	农业设施	行政区划	道路	植被	水系	其他
权	0.15	0.05	0.05	0.25	0.15	0.1	0.15	0.1
得分	83	81	78	86	91	72	88	77

则地名数据库属性数据质量最终得分：

$$S = \sum_{i=1}^m W_i S_i = 0.15 \times 83 + 0.05 \times 81 + 0.05 \times 78 + 0.25 \times 86 + 0.15 \times 91 + 0.1 \times 72 + 0.15 \times 88 + 0.1 \times 7$$

7=83.65

其中， $W_i (i=1,2,3,\dots,m)$ 为各类属性数据的权， S_i 为各类属性数据的得分。

DETAIL ABSTRACT

Recently, most research on data quality concentrates in space. So the research on attribute data error is weak. For those, this paper carries out a deep study on data quality control and evaluation in GIS.

(一) Evaluation model

(1) A composite Evaluation model of attribute data quality in GIS

① Gather all items for quality evaluation of attribute data in GIS into C:

$$C = \{C_1, C_2, C_3, \dots, C_n\}$$

② Establish powers of all evaluation items according to representation ability to attribute data quality in GIS. We show it in $Q = \{Q_1, Q_2, Q_3, \dots, Q_n\}$

③ Distribute total score of each item in C according to its own power got in step two.

Total score of each item in C is T :

$$T = \{T_1, T_2, T_3, \dots, T_n\} = 100 * Q = 100 * \{Q_1, Q_2, Q_3, \dots, Q_n\}$$

④ Evaluate each item of C and get its score based on evaluation standard and T . Each item's score is S:

$$S = \{S_1, S_2, S_3, \dots, S_n\}$$

⑤ Sum on each score in S, so we get the finally score of attribute data quality counted in hundred. We show it in B:

$$B = \sum_{i=1}^n S_i = S_1 + S_2 + S_3 + \dots + S_n$$

(2) A classification evaluation model of data quality in GIS

For large quantities of test data, because each class' data quality is very different, the different evaluation process is necessary.

① Divide into n classes according to contents and expression of attribute data. We show it in C:

$$C = \{C_1, C_2, C_3, \dots, C_n\}$$

② Establish a power to each attribute data class according to its weightiness, scale and representation. We show it in W

$$W = (W_1, W_2, \dots, W_n) \quad \text{and} \quad \sum_{i=1}^n W_i = 1$$

③ Establish the inspection size of each attribute data class

To get appropriate inspection size of each class, we usually establish it according to the power of each class: If the power of each class is W , the rate of inspection size of each class is $\bar{S}$:

$$\bar{S} = W_1 : W_2 : \dots : W_n$$

Afterward, we establish total inspection size based on charge and reliability of this evaluation. The inspection size of each class is:

$$S = (S_1, S_2, \dots, S_n):$$

$$S_i = E * \frac{W_i}{\sum_{k=1}^n W_k} = E * \frac{W_i}{W_1 + W_2 + \dots + W_n} \quad i = (1, 2, \dots, n)$$

④ Respectively evaluate each attribute data class according to correlative rules and get its own score. We show it in U :

$$U = (U_1, U_2, \dots, U_n)$$

⑤ We get total score of all attribute data by summing on each score using power:

$$\begin{aligned} S &= W \times U^T = (W_1, W_2, \dots, W_n) \times (U_1, U_2, \dots, U_n)^T \\ &= \sum_{i=1}^n W_i * U_i = W_1 * U_1 + W_2 * U_2 + \dots + W_n * U_n \end{aligned}$$

There, W_i is the power of i th class.

(二) The improvement in precision of attribute data

(1) We can improve the precision of attribute data with non-quantity probability expression.

(2) We can improve the attribute data precision after operating with non-quantity

probability calculation.

(三) The control and evaluation of attribute data quality in 1/250000 toponym database of Shangdong province.

Evaluation scheme:

(1) The quality control of attribute data

To ensure the quality of attribute data, we must proceed with following aspect: The origin of data, the format of data, and the expression quality of attribute data,etc.

(2) The classified evaluation of attribute data quality in 1:250000 toponym database of Shangdong province:

① Classify all attribute data:

Table 1 The classification of attribute data

Class	Residential area	Industrial establishment	Agriculture establishment	Administration section
Class	Road	Vegetation	Water system	Others

② Get all evaluation items of attribute data and their power in different class.

evaluation items:

Table 2 The inspection items of attribute data

Connect field and keywords		Expression of attributes	
Data conversion		Definition of attributes	
Correctness of description		Correctness of table	
Data format	Maturity	Metadata quality	Exclusion

The weightiness, representation and status of evaluation items are different in various classes. The evaluation quality will be improved if we establish the power of each item in various class according to its own feature and weightiness.

③ Establish the score of each item in various classes

If there are m classes and n items, we limit:

The score of P_i in C_j is T_{C_j-i} . $T_{C_j-i} = 100 * W_{C_j-i}$

$$W = \{\{W_{C1-1}, W_{C1-2}, \dots, W_{C1-n}\}, \{W_{C2-1}, W_{C2-2}, \dots, W_{C2-n}\}, \dots, \{W_{Cm-1}, W_{Cm-2}, \dots, W_{Cm-n}\}\}$$

W is the power gather of items in various classes and W_{C_j-i} is the power of the i th item

in C_j .

In 1/250000 toponym database of Shandong province, based on above classification, we let $m=18$ and $n=10$ and compute the score of each item in various classes according to previous formula.

④ Establish the grade standard of attribute items.

There must have some standards when we grade each attribute item. We use A,B,C,D and E grading to evaluate each item.

⑤ Evaluate attributes in classification.

Evaluate attribute data in classification in toponym database according to its power. Its score show as following:

Table 3 The power and score of various attribute datas

Class	Residential area	Industrial establishment	Agriculture establishment	Administration section
	0.15	0.05	0.05	0.25
Class	Road	Vegetation	Water system	Others
	0.15	0.1	0.15	0.1

So the finally score of attribute data in toponym database is:

$$S = \sum_{i=1}^m W_i S_i = 0.15*83 + 0.05*81 + 0.05*78 + 0.25*86 + 0.15*91 + 0.1*72 + 0.15*88 + 0.1*77 = 8$$

3.65

There $W_i (i = 1, 2, 3, \dots, m)$ is the power of each attribute class and S_i is the score of each attribute class.

1 绪论

1.1 数据质量控制的重要性

在 GIS 诞生不久的二十世纪六七十年代,人们在开发 GIS 进行数字制图时,发现必须注意 GIS 数据的质量问题。尤其是在 1975 年,Macdougall 用令人信服的实例说明了不考虑数据质量信息所带来的严重后果。随后, GIS 数据质量研究,为 GIS 系统的开发者和用户所认可并得到真正的重视。GIS 应用技术的成熟使得 GIS 应用范围更加广阔。越来越多的专家和学者开始从事有关 GIS 数据质量的理论研究及其应用,并取得了许多显著成果。

当前的社会是一个信息化的社会,信息高速公路、互联网、数字地形等全新的信息化概念和重大的全球性信息工程的提出和实施,正在深刻地影响人类生活的各个领域。而地理信息系统是信息化社会的骨架支撑技术之一,关系到国民经济建设、社会发展和国家安全。而在地理信息系统中,数据质量的好坏和数据标准化程度的高低都直接影响到地理信息系统的经济效益和社会效益。因此, GIS 空间数据质量与属性数据质量已经引起了地理信息系统的设计者、建设者和最终用户的高度重视。

地理信息系统是一个有地理位置的信息系统,这是 GIS 区别于其他信息系统的一个主要特点。从 GIS 系统的 3 个基本内容,即具有地理位置的空间实体、描述实体的属性、空间实体之间的拓扑关系来看, GIS 系统的质量控制分别包含了空间数据的质量控制、属性数据的质量控制、空间实体之间关系和空间数据与属性数据之间关系的质量控制。GIS 系统的空间性决定了空间数据在 GIS 应用中所起作用的基础性和重要性,因此 GIS 空间数据的质量控制依然是 GIS 系统质量控制的重要要素。但是地理信息系统决不仅仅是一个数字化的地形图,因此 GIS 系统的质量控制就不能仅仅限于空间数据的质量,属性数据的质量控制和实体与属性之间关系的质量控制也是 GIS 系统质量控制中必不可少的内容。对 GIS 数据质量标准的论述归纳为以下六个方面,即:位置精度、属性精度、时间精度、数据情况说明、逻辑一致性、数据完整性^[1]。

我们知道,空间数据是有关空间位置、属性以及时间信息的符号记录,空间位置、属性(专题特征)以及时间是表达现实世界空间变化的三个基本要素,我们分别对其进行

以下解释和理解:

(1) 位置精度

是以具有三维坐标的点、线、面来表达实体并作为研究对象,以研究空间实体的坐标数据与真实位置的接近程度,即常表现为空间三维坐标数据的精度。它包括数学基础精度、平面精度、高程精度、接边精度等。

空间数据位置的精度将直接影响 GIS 的应用,在质量控制检查中,对数据的位置精度进行严格的检查和分析是十分重要的一环。一般来讲,定位精度由两部分组成,偏差与精度。偏差是描述真实位置与表达位置之系统差别,理想的偏差应当是零,表明图上位置与表达位置没有系统偏差。偏差通常由样本点的平均位置误差度量。精度是数据元素位置误差的描述,精度通常用对选择的测试点标准离散差的计算来估计。一个较小的离散差表明位置误差的取值范围紧凑,即误差的取值范围相对较小,量测的数据精度越高,使用的可信度也越高。位置精度的度量通常是采用在测量中的均方根误差,这种度量方法与系统偏差的度量方法没有明显区别,估计数据质量的均方根误差,对于有效使用数据是非常有价值的。因此,空间数据的质量控制可以从图廓点、平面精度实测数据、高程精度实测数据等方面加以考虑。

(2) 属性精度

指空间实体的属性值与其真值相符合的程度。通常用文字、数字、符号、注记及其组合等来表达实体的属性,如地形图中建筑物的结构、层数,要素的编码、层、色、线型等属性值,这些属性值的正确性和准确性即为属性精度。

(3) 时间精度

指数据的现势性。可以通过记录数据获取或更新的时间和频度来表现。

(4) 数据情况说明

要求用文字、数据或图表等形式对空间数据的来源、数据的内容及其处理过程等做出准确、全面和详尽的说明,便于使用。

(5) 逻辑一致性

是指地理数据关系上的可靠性。包括数据结构、数据内容(空间特征、专题特征、时间、数据情况说明和拓扑性质上的内在一致性。

(6) 数据的完整性

指地理数据在范围、内容、结构等方面满足所有要求的完整程度。包括数据范围、空间实体类型、空间关系分类、属性特征分类等方面的完整性。

数字产品按其点位精度、属性内容、应用范围不同大体上可以分为数字线划地图、数字栅格地图、数字高程模型、数字正射影像等。因为数据采集手段不同，其质量控制内容也有差异。这里要搞清楚的数字地图产品的质量与 GIS 的数据质量、数字产品的成图质量和数据质量概念不能混为一谈^[2]。

1.2 属性数据质量控制的重要性

地理信息系统决不仅仅是一个数字化的地形图，因此 GIS 系统的质量控制就不能仅仅限于空间数据的质量，属性数据的质量控制和实体与属性之间关系的质量控制也是 GIS 系统质量控制中必不可少的内容。地理信息系统对资源与环境空间数据的高度综合分析和快速处理计算的能力使之得到愈来愈广泛的应用。在 GIS 中，地理实体之间的错综复杂的关系是通过属性数据表示的。GIS 中地理数据包括了空间数据、非空间的专题属性数据、时间数据等，其用于描述自然地理现象、社会经济现象和人文地理现象等。其中属性数据又分为定性数据和定量数据。定性数据用于描述地理现象的质量特征，突出反映现象间质的差别，如描述土壤类型、行政区划名称等；定量数据描述地理现象的数量特征，如降雨量、人口数等。从地图学的角度讲，定性数据一般用于制作类型图或区域图，如土壤类型图、植被类型图、行政区划图等，许多学者称这种数据为类型数据或分类数据，而且利用矢量模型、栅格模型以及矢栅混合模型来实现这类数据的综合。定量数据一般用于制作直方图或对比图，如人口图，粮食产量对比图等。

属性数据的不确定性主要来自数据源的不确定性、数据建模的不确定性和分析过程中引入的不确定性等。其中数据源的不确定性又来源于数据采集过程中的测量、人为判断和假设等。例如，在 GIS 中将现实空间分为具有渐变特征的连续空间区域和具有跳跃特征的离散空间区域；对于后者，通常假定区域内部属性是均匀且同质的，其重要特征的变化发生在边界上，像土壤类型分布图。由于区域边界是经过人为判断和边界定位确定的，因此，区域多边形中属性数据的不确定性将主要取决于属性分类不确定性、边界定位误差和属性测量误差。这里，误差是指观测值与其真值间的差异，是一种具有统计意义的概念；而不确定性则是指被测量对象知识缺乏的程度，它表现为随机性和模糊性，而且受尺度、分辨率、抽样等许多因素影响。在矢量专题图的 GIS 应用分析中，属性数据的不确定性问题主要是属性不确定性的度量和在 GIS 分析过程中的传播。

在 GIS 的属性数据库设计中, 面对给定的数据对象, 应该构造一些什么样的属性数据结构, 应由哪些属性组成, 以及这些属性之间相互联系的方式如何决定等, 都是属性数据库应用设计所要解决的问题。地理实体之间相互的关系表示是十分复杂的, 如何使用简洁的属性表达复杂的地理实体之间的联系就成为 GIS 设计的核心问题。GIS 的属性数据库是以关系模型为基础的数据库, 属性数据库利用关系来描述客观世界^[3]。一个属性记录既可用于描述一个地理实体, 又可用于描述实体之间的联系。对于一个属性数据, 可以有多种不同的表示, 每个表示称为一条记录。对于一个属性数据库, 可以包含许多不同的记录。一个属性数据库的全部记录的集合称为属性数据库。因此, 属性数据库设计的优劣决定了属性数据库应用的性能。所以, GIS 空间数据的质量不仅涉及定位误差, 而且涉及属性数据的精度。目前, GIS 空间数据精度分析与质量控制主要在位置精度分析方面做了大量的工作, 并取得了一定的成果, 主要涉及 GIS 数据数字化误差分析与图形要素的几何纠正, 线元不确定性模型分析研究等。而在属性数据精度分析方面, 则主要集中在遥感分类属性数据的精度分析和一些新方法的引入上, 许多方面的研究相对比较薄弱。一个好的属性数据库设计, 要在设计时反复考察客观实体之间的联系, 仔细分解, 去其冗余, 保留本质属性, 进而确定其最优化的属性表达, 从而使设计出的属性数据库应用性能达到最优。

1.3 属性数据质量控制的研究现状与存在问题

GIS 已经发展成为一种产业, 与其它工业产品一样, GIS 产品也必须有质量管理和质量控制。十九世纪八十年代, 美国数字制图数据标准委员会 (NCDCDS) 制定了数字制图数据标准, 对数字制图的数据质量做了规定, 而后我国国家测绘局也制定了有关质量标准。

长期以来, 许多学者在数据质量与不确定性方面作了深入的研究并取得了许多成果。但多数都是对几何图形数据精度模型和应用进行研究和分析, 而对属性精度的研究仍然比较欠缺。目前对属性数据不确定性的研究已经越来越引起人们的关注。1999 年, 童小华、史文中、刘大杰等对 GIS 中地图扫描数字化所获得的属性数据从数据的完整性、数据的逻辑一致性、属性精度、位置精度等方面进行系统分析和分布拟合检验, 探讨了 GIS 属性数据质量评价的方法。Goodchild 把对数据误差的描述称为数据信息的一个必须的

功能,此外它也是 GIS 数据的一个基本元素。Unwin 随之提出了“误差敏感”GIS 的概念,这一概念已被广泛认可并为众多学者深入研究。此外,Goodchild 对属性数据质量的基本要素与处理方法做了详细的阐述。另外,在对属性数据的研究中很多成果是在遥感影像分类中探讨分层属性精度的度量与描述。如 Unwin 对遥感数据精度估计的方法和应用作了较为综合的论述,Card,Chrisman 和 Hay 探讨了遥感影像分层中误分层对面积估算的影响;此外,Stephen 讨论了分层随机抽样中分层属性精度的 Kappa 系数估计及其方差;Maand Roland 则讨论了遥感影像分层精度的误差矩阵,并采用 Kappa 系数、Tau 系数和概率估计方法对分层属性精度进行度量。近年来,在属性数据精度的研究方面又有了一些新的方法。例如:模糊集合理论被应用于对属性分层的不确定性进行度量,粗集理论被探讨应用于属性数据分类精度的描述和度量。事实上,属性数据在 GIS 系统中多用关系数据库以数据记录的形式存放,从而研究大量缺少相互数值关系的数据时,抽样方法是一个简单而实用的方法。抽样方法是获取统计资料的重要手段,已经成为世界各国社会经济基本情况调查和各项专题调查的一种方法。在 GIS 属性数据精度分析和度量中采用抽样技术,一方面符合数据作为产品进行质量检验的国际标准,另一方面也能以较低费用,较高效率对 GIS 属性数据质量进行推断和描述。1990 年,Caspary 讨论了统计质量控制方法和抽样方案主要参数的获取方法,并采用 OC 曲线体现抽样方案的优劣。刘春,刘大杰讨论了 GIS 数字图质量抽样检验的基本模型,分析了我国现行测绘产品质量标准中抽样检验方案理论上的不足之处,并提出了适合 GIS 数字图产品质量抽样检验的模型和方案,给出属性数据精度缺陷率的度量模型,又针对 GIS 属性数据分层存储和表达的特点,提出采用分层抽样检验的方法获得数据缺陷的调查数据,在该调查数据的基础上探讨了缺陷率对属性数据质量的度量。

GIS 属性数据的质量分析控制、误差传播与建模是当前 GIS 研究的重点和难点。属性数据是 GIS 的血液,其质量是 GIS 的生命,随着 GIS 在各行各业的推广应用,属性数据库日渐庞大,数据质量问题愈来愈受到人们的关注,其数据质量的好坏直接影响着 GIS 应用、分析、决策的正确性和可靠性^[4]。目前,GIS 属性数据采集方法可以分为直接法和间接法,从图形中自动提取是主要的间接采集方法,它比手工输入精度高、速度快、操作方便、受环境影响小,而 GIS 数据质量的研究集中在定位数据的误差上,对属性数据精度的研究较少。

1.4 课题的提出

当前信息化社会正在深刻地影响人类生活的各个领域，而地理信息系统是信息化社会的骨架支撑技术之一，关系到国民经济建设、社会发展和国家安全。在地理信息系统中，数据质量的好坏高低都直接影响到地理信息系统的成败。因此，有必要对地理信息系统的数据质量高度重视。作为一种产业，GIS 产品与其它工业产品一样也必须有质量管理和质量控制。随着 GIS 的大范围应用，属性数据库日渐庞大，其质量问题研究以刻不容缓。为保证 GIS 应用、分析、决策的正确性和可靠性，我们有必要对 GIS 属性数据的质量控制、误差传播与建模进行系统研究。

在 GIS 中，地理实体之间相互的关系表示是十分复杂的，如何使用简洁的属性表达复杂的地理实体之间的联系就成为 GIS 设计的核心问题。属性数据在 GIS 系统中多用关系数据库以数据记录的形式存放。而研究大量缺少相互数值关系的数据时，抽样方法是一个简单而实用的方法。它是获取统计资料的重要手段，已经成为世界各国社会经济基本情况调查和各项专题调查的一种方法。在 GIS 属性数据精度分析和度量中采用抽样技术，一方面符合数据作为产品进行质量检验的国际标准，另一方面也能以较低费用，较高效率对 GIS 数据质量进行推断和描述。由于地理实体之间的错综复杂的关系是通过属性数据表示的，所以属性数据库设计的优劣决定了属性数据库应用的性能，属性数据质量好坏对资源与环境空间数据的高度综合分析和快速处理计算的能力极为关键。

近年来，GIS 数据质量的研究集中在位置数据的误差上，而对属性数据误差的研究较少，多年来的研究主要集中在遥感分类属性数据的精度分析和一些新方法的引入上，关系数据库存储的属性信息质量方面的研究相对比较薄弱。在以属性数据分析为主的许多 GIS 应用领域中，例如地名信息库、土壤化学、环境科学和农林科学等，属性数据质量的重要性甚至远大于位置数据。

为了给出一个 GIS 属性数据质量抽样检验的基本模型，对我国现行 GIS 产品质量标准中抽样检验方案理论进行研究，提出一个适合 GIS 数字图产品质量抽样检验的模型和方案，给出属性数据精度缺陷率的度量标准已成为当务之急。针对 GIS 属性数据分类存储和表达的特点，采用分类抽样检验的方法获得属性数据精度，并在该基础上探讨缺陷率对属性数据质量的度量更适合当前 GIS 属性数据特点。为了从数据的完整性、数据的

逻辑一致性、属性精度等方面进行系统分析,我们必须提出一种新的 GIS 属性数据质量评价的方法。

国家测绘局 1998 年启动了全国 1:25 万数据库建设工程,包括数字高程模型数据库、数字栅格地图数据库、数字影像地图数据库、数字线划地图数据库、地名数据库、土地覆盖数据库、元数据库等 7 个子数据库。山东省 1:25 万地名数据库是国家基础地理信息系统全国 1:25 万地名数据库的重要组成部分,它是一个空间定位型的关系数据库,其数据内容、数据的组织、数据库结构、建库的工艺流程都精心设计,是对地理数据库应用一次成功的探索。其储存和管理着 1:25 万地形图上的各类地名信息与相关的信息,数据量大,信息复杂,收录了 1:25 万比例尺地形图上的全部地名和山东省所有建制村以上行政区划地名约 400 万条,乡镇以上行政区划地名还含有部分社会经济信息。它需要与山东省 1:25 万地形数据库连接使用,为山东省各部委、各专业部门提供准确的地名信息,特别是为发生各类紧急突发事件时能提供准确的空间定位信息,做到快速反应。

在如此庞大的属性数据库中,数据的质量是系统功能目标实现的关键,要保证数据库运行的可靠性、正确性,必须在建库过程中,将数据质量的检查与验收贯穿始终。为此,针对山东省 1:25 万地名数据库的特点,结合已有的数据质量控制评价理论,提出一套切实可行的属性数据质量控制与评价理论对地名数据库的建设至关重要。

1.5 论文的主要内容及结构

(一) 主要内容

结合已有的数据质量理论,本文重点对属性数据质量控制与评价进行了深入分析与研究,并应用在山东省 1:25 万地名数据库中,主要内容如下:

(1) 给出了 GIS 属性数据质量控制的有效措施,并提出了采用定性概率表达方式提高属性数据存储精度,采用定性概率计算方式提高操作后的属性数据精度。

(2) 提出了 GIS 属性数据质量综合评价模型,在此模型中首次提出根据评定项权分配评定项总分的构思。

(3) 针对存储多类属性数据的大型数据库,提出了 GIS 属性数据质量分类评价模型,在此模型中首次采用由各类数据量在总数据量中所占比重与重要性来确定各类的权,然后根据各类的权分配各类总分的评定方法。

(4) 在建立山东省 1:25 万地名数据库过程中, 将上述理论应用到属性数据的质量控制与评价中, 提出了一套切实可行的属性数据质量控制措施, 并应用上述评价理论对地名数据库数据质量进行了分类评价, 取得了良好的效果。

(二) 论文结构:

本文采用由浅入深, 由理论至应用的模式, 总体结构如下:

第一章: 给出了属性数据质量研究的现状与前景, 成果与不足。

第二章: 给出了质量管理和质量控制的一系列基本原理和概念, 常用的统计分析方法以及质量评价中一些基本理论, 重点对属性数据质量控制与评价做了深入分析。

第三章: 本章提出了一套极为有效的属性数据质量控制措施, 并结合属性数据的缺陷类型, 提出了一个可靠的质量评价模型: GIS 属性数据质量综合评价模型; 针对多类属性数据质量的缺陷率离散程度较大的情况, 本章提出了 GIS 属性数据质量分类评价模型, 提高了评价准确度。同时, 提出了采用定性概率表达方式提高属性数据精度和采用定性概率计算方式提高操作后的属性数据精度两种有效方法。

第四章: 将前述理论在山东省 1:25 万地名数据库中进行了检验, 提出一套属性数据质量控制措施与评分标准, 应用前述 GIS 属性数据质量分类评价模型对地名数据库中属性数据质量进行了分类评价。

2 属性数据质量控制的特点与发展

2.1 GIS 属性数据管理

2.1.1 GIS 应用

地理信息系统(Geographic Information System 简称 GIS)是一项以计算机为基础的新兴技术,围绕着这项技术的研究、开发和应用形成了一门交叉性、边缘性的学科,是管理和研究空间数据的技术系统。在计算机软硬件支持下,它可以对空间数据按地理坐标或空间位置进行各种处理、对数据有效管理、研究各种空间实体及相互关系。通过对多因素的综合分析,它可以迅速地获取满足应用需要的信息,并能以地图、图形或数据的形式表示处理的结果^[5]。

尽管现存的地理信息系统软件很多,但对于它的应用研究,归纳概括起来有二种情况。一是利用 GIS 系统来处理用户的数据;二是在 GIS 的基础上,利用它的开发函数库二次开发出用户的专用的地理信息系统软件。目前已成功地应用到了包括资源管理、自动制图、设施管理、城市和区域的规划、人口和商业管理、交通运输、石油和天然气、教育、军事等九大类别的一百多个领域。在美国及发达国家,地理信息系统的应用遍及环境保护、资源保护、灾害预测、投资评价、城市规划建设、政府管理等众多领域。近年来,随我国经济建设的迅速发展,加速了地理信息系统应用的进程,在城市规划管理、交通运输、测绘、环保、农业、制图等领域发挥了重要的作用,取得了良好的经济效益和社会效益。其主要有下面几方面应用:

(1) GIS 在地理空间数据管理中的应用

以多种方式录入地理数据,以有效的数据组织形式进行数据库管理、更新、维护、进行快速查询检索,以多种方式输出决策所需的地理空间信息。使其具有空间实体定义能力,空间关系查寻能力。

(2) GIS 在综合分析评价与模拟预测中的应用

GIS 不仅可以对地理空间数据进行编码、存储和提取,而且还是现实世界模型,可以将对现实世界各个侧面的思维评价结果作用其上,得到综合分析评价结果;也可以将自然过程、决策和倾向的发展结果以命令、函数和分析模拟程序作用上这些数据上,模

拟这些过程的发生发展,对未来的结果做出定量的和趋势预测,从而预知自然过程的结果,对比不同决策方案的效果以及特殊倾向可能产生的后果,以做出最优决策,避免和预防不良后果的发生。

(3) GIS 的空间查询和空间分析功能的应用

为了便于管理和开发地理信息,在建库时对空间信息分层处理。也就是说,根据数据的性质分类,性质相同或相近的归并一起,形成一个数据层。这样 GIS 容易对单幅或多幅图件及其属性数据进行分析 and 指标量算。这种应用以原始图为输入,而查询和分析结果则是以原始图经过空间操作后生成的新图件来表示,在空间定位上仍与原始图一致。因此,也可将其称为空间函数变换。这种空间变换包括叠置分析、缓冲区分析、拓扑空间查询、空集合分析(逻辑交运算、逻辑并运算、逻辑差运算)。

(4) GIS 的输出功能在地图制图中的应用。

地理信息系统的发展是从地图制图开始的,因而 GIS 的主要功能之一用于地图制图,建立地图数据库。与传统的、周期长、更新慢的手工制图方式相比,利用 GIS 建立起地图数据库,可以达到一次投入、多次产出的效果。它不仅可以为用户输出全要素地形图,而且可以根据用户需要分层输出各种专题,如行政区划图、土地利用图、道路交通图等等。更重要的是由于 GIS 是一种空间信息系统。它所制作的图也能够反映一种空间关系,可以制作多种立体图形,而制作立体图形的数据基础就是数字高程模型。

(5) 运用 GIS 系统,建立起专题信息系统和区域信息系统。

我们可以运用 GIS 系统,建立起多个专题信息系统和区域信息系统,如水资源管理信息系统、矿产资源信息系统、草场资源信息系统、水土流失信息系统和长途电信局 GIS 系统等等。这类信息系统具有有限目标和专业特点,系统数据项的选择和操作功能是为特定的专门目的服务。区域信息系统如加拿大国家信息系统、美国 Oakridge 地区模式信息系统等等。这类信息系统主要以区域综合研究和全面的信息服务为目标,可以有不同的规模,其特点是数据项多,功能齐全,通常具有较强的开放性。

(6) 地理信息系统与遥感图像处理系统的结合的应用。

遥感数据是地理信息系统重要信息源。其实目前大多数 GIS 系统已揉进图像处理功能,并把它作为其一个子模块。这种应用如海湾战争期间,美国国防制图局 GIS 实时服务,为战争需要在工作站上建立了 GIS 与遥感的集成系统,它能用自动影像匹配和自动目标识别技术处理,处理卫星和高低侦察机实时获得战场数字影像,及时地将反映战场现状的正射影影像叠加到数字地图上,数据直接传送到海湾前线指挥部和五角大楼,为

军事决策提供 24 小时的实时服务。

(7) GIS 中属性数据的综合及融合。

在现有的 GIS 中, 属性数据只是用于检索和查询, 或进行简单的统计, 难以深入的分析, 难以发掘隐含在其中的模式和规律。在众多项的属性数据中, 有时将几个属性项的属性数值加以综合, 构成一个具有某领域特定意义的新属性项新属性值, 这种综合不是综合前属性数据值的简单反映, 也不是它们的孤立集合, 而是经过某领域研究人员深思熟虑的综合分析, 用数量表示某领域问题的综合概念和结果特征。

数据融合的概念始于 70 年代, 但直接促其迅速发展的是进入 90 年代以后。利用 GIS 对空间数据强大的显示功能, 引进数据融合技术, 融合 GIS 的属性数据, 显示其结果。

总之, GIS 为人类由客观世界到信息世界的认识、抽象过程以及由信息世界返回客观世界的利用改造过程的发展和转化, 创造了空前良好的条件和环境。

2.1.2 GIS 数据库的发展

现实生活中, 大部分的数据(约 85%以上)都具有空间属性, 例如, 地址、电话号码、客户统计分布数据或者资产分布数据等等。利用这些数据信息的空间属性进行数据分析, 可以观察发展趋势, 帮助您掌握机遇。总而言之, 能够迅速有效地管理空间数据, 根据其空间属性进行分析, 对于当今的企业来说, 是势在必行的。特别是在金融、保险、市场营销、邮电通信、市政管理、城市规划、公安交通、石油地质、医疗保健等领域, 空间数据更是扮演着重要的角色。

空间数据的定义是多维的, 如经度、纬度、海拔和时间, 主要用来描述特殊区域或目标的地理或地面特征。例如在邮电系统中进行布线管理、故障分析、需求预测、场地安排、电话号码分配和紧急事件响应; 在市政管理中进行犯罪分析、交通安排、校园规划、社会服务、运输计划、环境保护、防灾减灾、建筑许可和税收管理; 在保险业中进行灾难管理、风险分析、投保人服务分析和市场分析; 在银行系统中进行财务管理、服务定位、抵押分析和贷款管理, 支行分合等方面, 空间数据的有效管理都是一个关键的组成部分。

为了对空间数据进行管理, 在 60 年代以后, 产生了 GIS 数据库, GIS 数据库是某区域内关于一定地理要素特征的数据集合, 主要涉及对图形和属性数据的管理和组织, 它有以下特点:

(1) GIS 数据库不仅有与一般数据库数据性质相似的地理要素的属性数据, 还有大量

的空间数据，即描述地理要素空间分布位置的数据，且这两种数据之间具有不可分割的联系。

(2) 用数据来描述资源环境，数据量大，即使是一个极小的区域。数据的应用相当广，如地理研究、环境保护、土地利用与规划、资源开发、生态环境、市政管理、道路建设等等。上述特点，决定了建立 GIS 数据库时，一方面应该遵循和应用通用的数据库的原理和方法，另一方面还必须采取一些特殊的技术和方法，来解决其它数据库所没有的管理空间数据的问题。由于 GIS 数据库具有明显的空间特性，所以有人又称它为空间数据库。

2.1.3 GIS 属性数据库的发展

空间地理数据的表示主要有栅格和矢量两种不同的形式。栅格形式是将地理表层空间划分为一系列网格，空间目标由这些网格的位置及其量化值来表示；矢量形式则是将地理空间的一切事物、概念进行抽象，形成点、线、面，由点、线、面来组成各类空间目标。由于数据获取、数据表达、地图投影等方面的原因，在计算机世界里，地理信息往往被表示成不连续的子空间，为了更好地模拟客观世界，人们需要针对这两种数据形式建立不同的属性数据格式，因而需要建立专门的 GIS 属性数据库。

地理信息系统 (GIS) 是在计算机硬、软件支持下对各种地图和空间地理分布信息进行数据采集、存储、管理、分析和输出的一个综合性空间型信息系统^[6]。GIS 中涉及的属性数据不同于常规管理信息中的数据，它是具有丰富地理特征的地理数据的集合，这些数据信息量大、来源复杂、原始数据不规范，因此如何将这些地理数据规范化、数字化，建立完整的 GIS 属性数据库是 GIS 建立过程中相当重要的基础工作。只有在建立属性数据库的基础上才能开展空间查询和空间分析。属性数据具有与一般事务类数据相同的特征。当前关系数据库技术日益成熟，适合组织与管理地理数据库中的属性数据，建立属性数据库。为了能对属性信息进行分析和处理，往往要根据一些特定的模式来建立属性信息，因而属性数据库的建立有其特定的模式。

2.1.4 属性数据和空间数据的统一

GIS 中的属性数据说明空间实体数据的几何类型、分类分级、数据特征、质量描述、地理名称、更新日期等等。按属性数据的性质、内容、存贮方式、与空间数据关系的密

切性, 属性数据又可以分为两类:

(1) 与空间实体数据紧密相关的、用于区分空间数据实体的基本属性类。它们可以是空间数据唯一标识符、分类代码、几何类型(点、线、面)、数据的所有者或采集者、数据来源、数据建立日期、数据精度、与其他空间实体之间的关系代码(点线面的组成关系、相邻关系等等)等信息, 这类属性的质量控制是 GIS 属性数据质量控制的重点。

(2) 描述空间实体专业特性的专题属性数据。这类属性数据因应用专业不同而异, 比如说在用于宗地管理的 GIS 系统中, 属性数据可以包含土地等级、土地权属性质、土地用途等。这类属性数据大都由专门的数据库管理系统(DBMS)存贮管理, 通过一定的方法与 GIS 数据库相联。

属性数据与空间数据之间无缝的实现属性数据与空间数据之间无缝要求将属性数据和空间数据完整地集成在一起。用全关系型或对象关系型数据库来存储和管理空间数据时, 可以将同一目标的属性数据和空间数据存放在同一条记录中, 空间数据用 BLOB 字段或空间类型字段来存放。这样, 属性数据和空间数据就可以完整地集成在一起。然而, 在全关系型或对象关系型数据库中, 属性数据和空间数据也可以分开用不同的表来存储(这样存储也带来了许多灵活性)。属性和空间数据即使存储在一个表中, 对关系表的操作(如投影操作)也可以对属性和空间数据分开使用, 因此, 此时属性数据和空间数据的一致性的维护也需要做特殊处理。更有效的属性数据与空间数据之间无缝的集成方法是建立纯的面向对象的数据库管理系统。面向对象的方法要求将对象的属性及其方法进行封装, 由对象本身来维护自身的状态。对于空间对象, 属性和空间数据一起封装在对象中, 系统对对象的操作和处理以对象为单位进行。因此, 属性和空间数据得到了很好的统一^[7]。

2.2 GIS 属性数据误差的起因与表示

地理要素应反映自然景观发展的客观规律, 真实再现该地貌特征, 无地理适应性矛盾, 同类实体和不同类实体间必须正确客观地反映地理实体间空间结构关系, 相关位置正确。在地理信息系统(GIS)中, 数据质量的好坏直接影响到系统的经济效益和社会效益。近年来, 中国在 GIS 的理论研究及产业化发展方面, 有许多成功的范例, 同时也存在一些问题。数据的质量欠佳, 是许多系统不成功的主要原因之一。

一、数据质量的基本概念与研究意义

(1) GIS 数据质量的定义

GIS 数据质量是指 GIS 数据产品满足一定使用要求的特性, 主要包括数据源、点位精度、属性精度、要素完整性和属性完整性、数据逻辑一致性、数据现势性等。GIS 的数据质量通常用数据的误差和正确率来度量。

(2) 研究 GIS 数据质量的重要意义

GIS 的数据质量问题是一个关系到数据可靠性和系统可信性的重要问题, 与系统的成败密切相关。由于现代 GIS 的先进技术, 用户可以不管比例尺的大小、图形的精度而较容易地把来源不同的数据进行综合、覆盖和分析, 如果 GIS 的数据质量不高, 其结果是误差增加, 导致不能决策, 系统失败。建立 GIS 的过程就是和误差作斗争的过程。GIS 产生的误差可以比喻成快速产生的垃圾, 如对误差处理不当, GIS 能以相当快的速度产生各种垃圾。如果不考虑 GIS 的数据质量, 那么当用户发现 GIS 的结论与实际的地理状况相差惊人时, GIS 产品就会在用户中立刻失去信誉。因此, GIS 若要生存发展下去, 必须花大力气从理论上研究 GIS 数据的质量问题。

(3) GIS 属性数据误差的起因与表示

属性数据质量是指关于属性类型和特征不能被准确确定的程度, 它是自然界各种现象自身固有的属性。在 GIS 图上存在大量的目标, 它不是随机的而是模糊的。例如无明确定义性质, 如可耕地、森林地带等, 它本身就带有属性不确定性。

根据数据性质的不同, GIS 中属性数据不确定性可分为由定位不确定引起的属性不确定性和由认知程度引起的属性不确定性。事实上, 在有些情况下, 这 2 种不确定性是可以相互替换的。例如土壤类别边界线确定的误差, 既可看由定位不确定引起, 也可作为由认知程度引起。

根据属性数据性质的不同, 属性数据质量总的来说可以用 2 种方法表示: 一种可以用点误差的分析方法, 而另一种则常用离散型的数据误差的评价方法来进行分析评价。前一种比较简单, 可用常规的标准差进行描述, 而后一种比较复杂, 尤其对那些定性属性数据的准确度的描述, 它将受到变量的离散值(如类型的个数), 每个属性值在空间上分布和每个属性的形态和大小, 检测样点的分布和选取以及不同属性值在特征上的相似程度等多种因素的影响。

(4) GIS 属性数据的误差类型

属性数据误差分为明显误差、源于属性获得的误差(又称源误差)、源于数据处理的

误差。这 3 组误差中, 第一组误差是十分明显而易于查出的; 第二组误差相对复杂而在与数据接触时可以查出; 第三组误差是执行某一处理时才会出现, 这种误差最难量测, 因为误差不仅与所使用的数据有关, 而且与所使用的算法、数据结构有关。三种误差所涉及的内容如下:

1) 明显误差主要包括:

①数据年龄。大多数系统使用已有的数据, 而这种数据往往是过时的。使用的数据越旧, 出现误差的可能性就越大。

②地图比例尺。地图比例尺越大, 表示的内容越详尽, 一般精度也越高。

③观测值的密度及其分布模式。有些专题图(如土壤图)的部分数据是对一系列观测值的内插得来的, 在建立数字地面模型时, 内插误差较量测误差要大得多, 有两个因素可影响其精度, 即观测值的密度及其分布模式。

2) 属性获得的误差包括

①属性数据采集仪器的误差, 由摄影平台、传感器的结构及稳定性、信号数字化、光电转换、分辨率等引起。

②属性数据读入误差, 由属性数据的录入、数据库的操作等引起。

③处理过程中产生的误差

指在对属性数据进行处理时产生的误差, 例如在下列处理中产生的误差: 属性纠正; 属性变换; 属性数据的编辑; 图形简化(如数据压缩); 属性数据格式转换;

2.3 研究 GIS 属性数据质量的主要理论

在 GIS 领域中人们已经研究了许多处理 GIS 属性数据误差的模型与方法。然而, 到目前为止还没有提出一个处理 GIS 属性数据质量问题的一般性理论。但是, 现有的若干理论可用于处理 GIS 属性数据质量的某些侧面的问题, 这些理论是概率论、模糊数学、证据数学理论以及空间统计理论。它们在处理 GIS 属性数据质量问题时具有互补性, 而且每一理论只可以处理某几方面的问题。

(1) 概率论

概率论可用于处理由于随机误差而产生的属性数据质量问题。在概率论中, 属性数据的误差被描述成给定某些统计值的条件下某一假设为真的条件概率。该假设的条件概

率表示了一个概率从 0 到 1 区间的定量描述。例如在遥感分类中,误差是由条件概率表示出来的,它指出在划分某一个像素为某一类别时的概率。

(2) 模糊数学

模糊数学是由 Zadeh(1965)提出的由计算机处理不精确概率的一种理论。其优点在于在进行自然语言有关属性查询方面具有较强的潜在应用可能。它的弱点是没有一个像概率论一样经过严格证明的过程。模糊性可能在属性域、时间域或(遥感影像的)频谱域,例如在空间域中的“城市区域”,属性域的“湿地”。对一个模糊集合,需给出一个隶属数值以表示该元素对于集合 A 的隶属度。

(3) 证据数学理论

证据数学理论是对传统概率论的一个扩展。在证据数学理论中,某一假设为真的概率量测值由支持某一假设的可获得的证据的概率来代替。证据理论是基于可信函数和可能函数所确定的一个区间,可信函数表示了证据支持假设的最低程度,而可能函数则表达了证据支持假设的最高程度。证据数学理论的优点在于它是概率论的更一般性表达,它的弱点是不能就矛盾证据或不同假设之间具有微弱的支持问题提出解决办法。

(4) 空间统计理论

统计学是用有序的模型描述无序事件的一门不确定性理论。Cressie 将 GIS 数据分为 3 大类:地理统计数据、栅格数据以及点数据。他提出了一个通用模型,该模型适用于上述 3 种数据类型。空间统计较之非空间统计是一个更一般性的理论。空间统计理论可应用于描述栅格属性数据,也可用于描述矢量属性数据的不确定性。

2.4 研究 GIS 属性数据质量的方法与标准

2.4.1 研究 GIS 属性数据质量的方法

(1) 建立误差分析体系

误差分析体系包括误差源的确定、误差的鉴别和度量方法、误差传播模型的建立,以及控制和削弱误差对 GIS 属性数据产品影响的方法。传统的概率统计仍是建立误差分析体系的理论基础。但是,必须根据 GIS 属性数据操作的特点对经典的概率统计理论进行扩展和补充。

(2) 敏感度分析法

一般而言,精确确定 GIS 属性数据的实际误差非常困难。为了从理论上了解输出结果如何随输入数据的变化而变化,可以通过人为地在输入数据中加上扰动值来检验输出结果对这些扰动值的敏感程度。然后根据适合度分析,由置信域来衡量由输入数据的误差所引起的输出数据的变化。为了确定置信域,需要进行地理敏感度测试,以便发现由输入数据的变化引起输出数据变化的程度,即敏感度。这种研究方法得到的并不是输出结果的真实误差,而是输出结果的变化范围。对于某些难以确定实际误差的情况,这种方法是行之有效的。

(3) 尺度不变空间分析法

地理数据的分析结果应与所采用的空间坐标系统无关,即为尺度不变空间分析,包括比例不变和平移不变。尺度不变是数理统计中常用的一个准则,一方面能保证在用不同的方法时得到一致的结果,另一方面又可在同一尺度下合理地衡量估值的精度。也就是说,尺度不变空间分析法使 GIS 的空间分析结果与空间位置的参考系无关,以防止由基准问题而引起属性数据分析结果的变化。

(4) Monte Carlo 试验仿真

Monte Carlo 试验仿真首先根据经验对数据误差的种类和分布模式进行假设,然后利用计算机进行模拟试验,将所得结果与实际结果进行比较,找出与实际结果最接近的模型。对于某些无法用数学公式描述的过程,用这种方法可以得到实用公式,也可检验理论研究的正确性。

(5) 空间滤波

获取 GIS 属性数据的方法可能是不同的,既可以采用连续方式也可采用离散方式采集。数据采集的过程可以看成是随机采样的,其中包含倾向性部分和随机性部分。前者代表所采集物体的实际信息,而后者是由观测噪声引起的。空间滤波可分为高通滤波和低通滤波。高通滤波是从含有噪声的数据中分离出噪声信息;低通滤波是从含有噪声的数据中提取信号。例如高通滤波后可得到一随机噪声场,然后用随机过程理论等方法求得属性数据的误差。

2.4.2 GIS 属性数据质量标准

鉴于 GIS 数据质量的重要性,国际标准化组织地理信息/地球信息业技术委员会(ISO/TC211)从 1995 年开始立项研制第一批国际地理信息标准,在其开展的 20 个项目中安排了 2 个与属性数据质量问题有关的项目:

(1) ISO 15046-13 地理信息—质量标准

该标准旨在建立描述地理数据质量的原理和提供有关数据质量的模型。该标准定义属性数据质量信息的组成部分,包括数据质量定量元素、数据质量定量子元素、数据质量度量方法、数据质量非定量元素等。地理数据的生产者和用户都使用这一标准,生产者按照该标准说明地理数据的质量,用户则由此确定地理数据是否满足其特定的应用需求。其中属性数据质量元素和子元素(即衡量 GIS 数据质量的标准)包括:

- ①完整性,如要素的完整性。
- ②逻辑一致性,如属性一致性、格式一致性、拓扑一致性等。
- ③精度,如绝对精度、像元精度、相对精度、形状的相似性等。
- ④时间精度,如与时间有关的属性或属性类型误差等。
- ⑤专题精度,如要素属性的连续值精度、有序值精度、额定值精度等。

(2) 地理信息—质量评价过程标准

该标准主要是提供数字形式地理数据质量的评价框架。评价过程根据地理信息—质量标准定义的数据质量模型确定数据质量评价内容,建立评价过程框图,将数据质量评价结果报告作为数据质量描述数据(即元数据)的一部分。按照数据生产者的生产规范和用户对数据质量的要求,确定 GIS 数据满足应用需求的质量等级。该标准提出的 GIS 数据质量评价过程简单可靠。数据生产者和用户在对 GIS 属性数据集进行质量评价时应遵循这一框架流程。

本章小结:

本章简单介绍了属性数据误差的起因与表示,研究 GIS 属性数据质量的主要理论以及研究 GIS 属性数据质量的方法与标准,分析了属性数据质量研究的现状与前景,成果与不足,给出了质量管理和质量控制的一系列基本原理和概念,常用的统计分析方法以及质量评价中一些基本理论,重点对属性数据质量控制与评价做了深入分析。

3 GIS 中属性数据质量控制方案设计

地理信息系统(GIS)区别于其它信息系统,主要在于它是一个有空间位置的信息系统, GIS 系统主要包含了具有地理位置的空间实体和描述实体的属性信息,由此可以看出, GIS 系统的质量控制实质上就是空间数据的质量控制、属性数据的质量控制,而属性数据的质量控制依然是 GIS 系统质量控制的重要要素。属性数据可以通过手扶跟踪数字化、扫描识别、实地获取、数字摄影测量、遥感等多种手段采集,所以属性数据质量控制的内容是丰富多彩的。数据采集手段不同,其质量控制内容也有差异。

3.1 GIS 属性数据质量控制指标、过程与实施

3.1.1 GIS 属性数据质量控制基本指标

属性数据质量可以分为对属性数据的表达和描述(属性数据的可视表现)和对属性数据的质量要求(质量标准)两个质量标准,保证了这两方面的质量,可使属性数据库的内容、格式、说明等符合规范和标准,利于属性数据的使用、交换、更新、检索、数据库集成以及数据的二次开发利用等。属性数据的质量还应该包括大量的引导信息以及由纯数据得到的推理、分析和总结等,这就是属性元数据,它是前述数据的描述性数据。因此,属性元数据也是属性数据可视表现的一部分。而精度、逻辑一致性和数据完整性则是对属性数据可视表现的质量要求。因此 GIS 属性数据的质量标准可这样表述:精度、逻辑一致性和完整性、时间精度、元数据精度。为了对属性数据进行质量评价,下面将表达和描述现实世界的属性数据应达到的质量要求:精度、逻辑一致性、数据完整性(即质量评价的指标):

(1) 精度:

描述空间实体的属性值(字段名,类别,字段长度等)与真值相符的程度。如类别的细化程度,地名的详细,准确性等。另外,数据采集更新的时间和频度,或者离当前最近的更新时间也反映了属性数据的精度。在 GIS 属性元数据中存放对图形,属性,时间及其相互关系或数据标识的参考信息,控制 GIS 属性元数据的质量也可以提高属性数

据的质量。

(2) 属性逻辑一致性:

属性值与真实地理世界之间数据关系上的可靠性称为属性逻辑一致性。包括数据结构, 属性编码, 线型, 颜色, 层次以及有关实体的数量, 质量, 性质, 名称等的注记, 说明, 在数据格式以及拓扑性质上的内在一致性, 与地理实体关系上的可靠性等。在具有时间度的 GIS 属性数据库中, 数据生产和更新的时间与真实世界变化的时间关系的正确性也反映了属性数据的逻辑一致性。在具有属性元数据的 GIS 属性数据库中, 对元数据内容的描述与真实地理世界数据关系上的可靠性和客观实际的一致性属性元数据的一个度量指标。

(3) 数据完整性:

地理数据在关系分类, 结构, 实体类型, 属性特征分类等方面的完整性。对于描述时间的属性数据, 必须完整记录数据生产或更新全过程。对元数据要求内容完整(现形元数据文件结构和内容完整)

通过上面的分析, 使我们很清楚 GIS 属性数据由三种数据表现形式来完整地描述, 它们是质量分析与控制研究的对象, 是数据表现形式的质量指标。属性数据这四种表现形式分别达到的精度、逻辑一致性和数据完整性以及这三者之间统一的程度, 便是属性数据的质量要求。使用 GIS 属性数据质量评价矩阵可以明确表明质量控制的对象、质量评价指标, 便于操作实施。

3.1.2 GIS 属性数据质量控制过程

质量是一切产品的生命, 它反映的是产品乃至生产产品的过程的优劣程度^[8]。地理信息系统数据库作为一种高科技产品, 为了能更好地发挥作用, 同样应研究并实施可行方案, 提高其质量, 并能对其质量进行总体评价。为此, 经过长期总结, 我们发现 GIS 属性数据库的质量控制分成前期控制和后期控制两个大部分最为理想。前期控制包括属性数据录入前的质量控制、属性数据录入过程中的实时质量控制, 后期控制为属性数据录入完成后的后处理质量控制与评价。

依据建库流程可分为下列几内容:

(1) 前期控制

前期控制是在提交成果(属性数据入库)之前对所获得的原始数据与完成的工作进行检查, 以进一步发现和改正错误。

(2) 过程控制

实施设法减少和消除误差和错误的实用技术和步骤，主要应用在建库过程中，用来对前期获得数据在录入过程中进行属性数据质量控制。

(3) 系统检测

在属性数据入库后进行系统检测，设计检测模板，利用检测程序进行系统自检。

(4) 精度评价

对入库属性数据用各种精度评价方法进行精度分析，为用户提供可靠属性数据质量信息。

3.1.3 GIS 属性数据质量控制具体实施

质量控制应贯穿在整个 GIS 属性数据采集、录入、存储与操作过程中，GIS 属性数据质量控制应分环节实施，具体实施方案如下：

(1) 设计全过程的质量控制

设计全过程的质量控制是指在系统调查和充分掌握资料 and 了解产品用途的条件下，制定属性数据质量目标，在产品设计的每个阶段，以质量目标为根据对设计进行评议审查；通过试制数据采集，对采集数据的质量做出技术鉴定和经济分析；保证设计生产过程符合规定的要求。

(2) 对基础属性数据资料的质量控制

根据产品的需求情况选择质量满足要求的数据源，这是决定数据产品质量的关键因素。原始资料的正确处理，不但可以减少误差，还可以提高效率，数据源的误差范围至少不能大于产品的质量要求范围。基础资料直接影响数据采集的质量，是数据质量的基础，因此必须对用于属性数据采集的基础资料进行严格的质量控制。

(3) 对属性数据采集手段的选择

根据数据产品的应用、用户的要求、精度高低的不同，合理选择不同的属性数据采集手段，满足质量及经济的双重需求。

(4) 对软、硬件配置的要求

用于内外业属性数据采集的各种软、硬件，其性能和技术指标必须满足属性数据采集的质量标准和技术设计书的要求。

(5) 属性数据采集前的准备工作

要做好数字产品过程的质量控制，首先必须从采集前的准备工作做起。准备工作包

括学习和理解有关技术文件,如技术设计书、属性数据产品质量标准、数据分类编码规定、操作技术规程等。

(6) 属性数据采集中的监控

指在属性数据采集的整个过程中,采取一些有效措施,实时地检验并预防和纠正误差和错误。该项质量控制是在建立属性数据文件和各项档案文件的制度下,按规定的各项质量指标抽取属性数据;利用各种统计图表工具经常而准确地掌握质量动态,以便发现问题,采取措施,防止或减少废次品的再生,使制造过程处于工艺稳定状态。过程表和进程跟踪就是典型的预防性质量控制的实时管理方式。过程表是对建立 GIS 属性数据文件的过程进行非常详细的规定。使用过程表可以确保所有操作员按照相同的程序进行。进程跟踪是实时质量控制中一种典型的预防性质量控制的管理方法。它利用计算机程序在属性数据采集过程中,通过属性数据之间应当具有的各种关系(数据取值范围、相联关系等等)来限制约束属性数据可能出现的质量问题^[9]。

(7) 结果控制

结果控制为属性数据录入完成后的后处理质量控制,给出 GIS 属性数据的质量评价。

3.2 GIS 属性数据精度评价模型设计

属性数据的不确定性主要来自数据源的不确定性、数据建模的不确定性和分析过程中引入的不确定性等。其中数据源的不确定性来源于数据采集过程中的测量、人为判断和假设。数据建模的不确定性主要与属性编码的合理性、完整性、包含性有关。属性精度评价是属性数据质量评价中最主要、最复杂、最困难的工作,原因是难以得到属性的真值(即参考数据)以及属性的组合模型相当复杂;自动判断其正确性涉及图形模式识别等复杂技术。属性组合的灵活性及属性值输入的复杂性使得数字化作业人员非常容易出错。

由于上述原因,必须在属性数据采集与处理过程中对其进行质量评价,一旦发现不合格产品,应立即采取相应措施。GIS 属性数据质量的评价内容包括数据集完整性、说明信息评价和地理目标数据评价几部分。数据集完整性包括要素分层的完整性、要素内容的完整性等;说明信息用于描述对图中各地理实体有共同影响的因素,包括原始资料的说明(如比例尺、生产单位、生产日期、地图投影和用于更新的辅助资料说明)、数据

获取的方法和日期、数据处理方法、作业人员、数据格式和各种文档资料；地理目标数据用于描述具体地理实体或其一部分的特征及其相互关系，评价内容包括属性精度、数据的完整性、逻辑一致性和拓扑表达正确性几个方面。

3.2.1 GIS 属性数据质量综合评价模型

为了从总体上对属性数据进行评价，必须设计出一套可行的属性数据质量综合评价模型。这些属性数据包括主码、识别码、描述码、参数、地名、地名指针、复合目标指针、拓扑关系等信息及其组合，对属性精度进行评价，首先要得到属性的真值（即参考数据），用扫描数字化地图底图作为判断属性正确性的标准，依据《数字地图要素属性编码》、《数字地图采集细则》对属性数据的值域、逻辑一致性进行检查，用图形符号、汉字、数字及色彩等配合坐标数据进行可视化表示，然后用自动检查和人机交互相结合的方法统计其出错率，根据错误种类、严重程度，利用属性数据质量综合评价模型评价其质量等级，具体实施步骤如下：

(1) 确定属性数据评价项

属性数据精度评价内容根据属性数据的表达内容不同而不同。从总体来讲，属性数据评价的内容可以参考以下几项：

①属性数据的完整性

属性数据要表达一定的信息，所表达的信息必须完整，才能使信息不致缺漏。属性数据的完整性包含形式的完整性，内容的完整性等几个方面。地理数据在空间关系分类，结构，空间实体类型，属性特征分类等方面必须完整，必须完整记录数据生产或更新全过程，元数据要求内容完整准确，保证现形元数据文件结构和内容完整。

②属性数据的逻辑一致性

属性的逻辑一致性是属性数据精度评价的主要内容之一，其反映了属性值与真实地理世界之间数据关系上的可靠性。逻辑一致包括数据结构合理，便于进行逻辑表达，能反映属性值与地理世界以及属性值与属性值之间的本质关系。属性编码简单清晰，便于查询，能反映属性类别等信息，编码必须具有统一标准，以利于以后的兼容。表达信息的线型，颜色，层次以及有关实体的数量，质量，性质，名称等的注记必须明确可靠，数据必须真实准确。属性说明必须简单易懂。拓扑性质反映属性数据的内在一致性，表达拓扑信息的属性数据必须准确，方便拓扑操作，关系查询，并且与地理实体关系可靠性。表达时间性的属性数据必须存储数据生产和更新的时间能反映真实世界的变化。属

性元数据必须一致可靠地反映属性数据之间的逻辑关系,元数据内容与客观实际一致性。

③属性定量精度

实体的属性值(字段名,类别,字段长度等)与真值必须相符,具有细化的类别,详细的地名。数据采集更新的时间必须能充分反映当前状况,当前最近的更新必须反映当前属性真值。在历史属性数据中,必须真实地反映当时的信息。属性元数据的属性,时间及其相互关系或数据标识的参考信息必须准确。

④属性定性精度

在没有办法定量表达的属性数据中,可以采用定性的方法进行表达。定性数据常常采用当前值反映真值的概率来表示。为了保证定性数据的可靠性,必须采用可靠的评估手段,确定量值。

⑤地名的正确性

地名是属性数据表达的重要内容,地名的正确与否直接关系到属性数据的其他信息的正确性。在属性数据中,地名的表达必须准确,不能有歧异。不同的地方,不能有相同的地名。同时,在不同的数据中,地名必须统一。

⑥关系的正确性

有一些表达地理关系的属性数据,其存储了地理实体之间的逻辑、等级、隶属等关系。正确表达他们的关系是属性数据质量的重要方面之一。通过详尽准确的关系表达,可以进行空间决策、空间推理、人工智能等系统设计。

(2) 确定属性数据评价项的权

在具体应用中,由于各评价项的重要性不同,所占比重不同,以及评价指标不同,评价项能反映属性数据质量的程度也不相同。为了最大消除评价误差,综合考虑各评价项反映属性数据质量的能力,采用权 Q 对各评价项在综合评价模型中的地位加以调整:

对应(1)中每一评价项,其权用 Q 表示:

$$Q = \{Q_1, Q_2, Q_3, \dots, Q_n\}, \text{ 且 } \sum_{i=1}^n Q_i = 1 \quad (3.1)$$

(3) 确定属性数据评价项所占分值

由于属性数据评价项在应用中所处地位不同,重要性不同,为了最大限度减少评价误差,必须合理分配各评价项所占分数。而评价项的权是根据其重要性与所占比重、评价指标和评价项对总体的代表性而确定的,为了计算方便,这里依据评价项的权来分配

量。因此，采用此种判分标准既是合理的，也是简单实用的。

(6) 属性数据精度综合评价

依据上面五步，我们得出对属性数据质量进行综合评价的方法如下：

①将(1)中属性精度评价项构成集合：

$$C = \{C_1, C_2, C_3, \dots, C_n\}$$

②综合考虑各评价项反映属性数据质量的能力以及对总体的代表性，确定各评价项权：

$$Q = \{Q_1, Q_2, Q_3, \dots, Q_n\}$$

③对 C 的每一项按照(2)中的权进行所占总分比例的分配：

$$T = \{T_1, T_2, T_3, \dots, T_n\}$$

④按照(5)中评价规则，在总分 T 的基础上进行 C 中每一评价项的得分评价，得出评价项集合 C 的每一项内容得分 S 为：

评价项 C_1 得分为： S_1

评价项 C_2 得分为： S_2

.....

评价项 C_n 得分为： S_n 即：

$$S = \{S_1, S_2, S_3, \dots, S_n\}$$

⑤将每一项得分进行累加，即得到百分制属性数据总体质量 B ：

$$B = \sum_{i=1}^n S_i = S_1 + S_2 + S_3 + \dots + S_n \quad (3.3)$$

3.2.2 GIS 属性数据质量分类评价模型

对于大批量的检验总体，有时总体各类质量的实际分布差别很大，对不同类采用不同的检验程序就比较符合实际。所以 GIS 数据质量的抽样检验中，引入分类概念是符合 GIS 数据组织的基本特点的。分类数据是属性数据的一个重要类型，它多用于描述不同类型地理现象的分布状况。这类地理现象一般存在从高级分类到低级分类，或者是从详细分类到粗略分类的分类体系。基于分类抽样的属性数据精度评价采用分类方法来度量

评价项所占分值：

评价项 C_1 所占总分 T_1 为： $T_1 = 100 * Q_1$

评价项 C_2 所占总分 T_2 为： $T_2 = 100 * Q_2$

.....

评价项 C_n 所占总分 T_n 为： $T_n = 100 * Q_n$

即：各评价项总分 $T = \{T_1, T_2, T_3, \dots, T_n\} = 100 * Q = 100 * \{Q_1, Q_2, Q_3, \dots, Q_n\}$ (3.2)

$Q = \{Q_1, Q_2, Q_3, \dots, Q_n\}$ 为评价项的权

(4) 确定属性数据抽样数量

度量属性数据精度，关键是获得属性数据的合理抽样数。不同的抽样方案总体获得的结果是不同的，合适的抽样方案应该是在某一费用、精度要求下获得属性数据的较为准确的调查数据。选择抽样精度高的抽样方法抽取样本，可以获得更好的说明总体数据质量的样本检验结果。确定抽样数量一方面要考虑到样本量的大小，另一方面也需要顾及抽样样本对总体的代表性。

(5) 属性数据评价项评分标准

属性数据评价项评分标准往往根据不同的目的与要求确定规则，最常用的是采用缺陷评价法。所谓缺陷评价法，简单地说就是估计评价项中包含的缺陷类型、性质等因素。这一评价法中的缺陷类型是属性数据评定内容的质量检验结果，可以被用来评价属性数据的质量。缺陷类型现有国家测绘数据质量检验标准，现已经采用，属性数据评价项的缺陷分为三种类型，即一般缺陷 (ligh-def)、重缺陷 (heavy-def) 和严重缺陷 (serious-def)。通常，当某一评价项属于一般缺陷时，其项得分应限定在此项总分的 80% 以上；而若此评价项带有重缺陷，其项得分应限定在此项总分的 60%-80%；若此评价项带有严重缺陷，则其得分应在此项总分的 60% 以下。

例如：对属性数据的完整性进行评价，若此项总分为 40 分，评价结果为此项属于一般缺陷，则此项得分应大于 $40 \times 80\% = 32$ 分。若此项评定结果发现带有重缺陷，则其得分应在 $40 \times 60\%$ 至 $40 \times 80\%$ 之间，既 24 分—32 分。如果其完整性属于严重缺陷类型，则其项判分应在 $40 \times 60\% = 24$ 分以下。

可以看到，属性数据缺陷法以已经颁布的测绘数据质量检验标准或已知值为依据来判断获得评价项得分，它们反映了属性数据评价项的可靠程度，也反映了属性数据的质

数据的精度，这在质量管理学中，被认为是研究大批量数据精度的一种简单且切实可行的方法，而它的前提是要分类合理。

在对 GIS 属性数据分类评价时，可考虑按照上面讲的 GIS 属性数据精度综合评价模型估计每一类的得分，类的得分通过适当加权就能得到对整个总体的属性数据质量的估计量，实现方法如下：

- (1) 依据属性数据内容与表达方式将属性数据分成 n 类，用集合 C 表示：

$$C = \{C_1, C_2, C_3, \dots, C_n\}$$

- (2) 依据重要性、在总体中所占比例以及代表性等对各类属性数据赋予一个权 W ：

$$W = (W_1, W_2, \dots, W_n) \quad \text{且} \quad \sum_{i=1}^n W_i = 1$$

- (3) 确定各类属性数据的抽样数量

采用分类度量属性数据精度，关键是获得各分类属性数据的合理抽样数。不同的抽样方案总体抽样获得的结果是不同的，合适的抽样方案应该是在某一费用精度要求下获得属性数据的较为准确的调查数据。确定抽样方案实际上是确定各类属性数据的抽样比，抽样比一方面涉及抽样样本量的大小，另一方面也需要顾及抽样样本对总体的代表性。由于 GIS 不同类属性数据的数据特性以及数据规模都有较大的差别，因此，分类评价比一般抽样能更有效地控制抽样误差。

为了获得各类属性数据合理的抽样数量，往往根据各类分配的权确定抽样比：

若各类属性数据的权为 W ：

$$W = (W_1, W_2, \dots, W_n) \quad \text{且} \quad \sum_{i=1}^n W_i = 1;$$

则各类属性数据的抽样数量比 $\bar{S}$ 为：

$$\bar{S} = W_1 : W_2 : \dots : W_n$$

然后根据评价费用以及结果可信度要求等确定样本总量 E ，按上述比例分配到各类属性数据。若样本总量为 E ，则各类属性数据抽样数量为 S ：

C_1 类抽样数量 S_1 ：

$$S_1 = E * \frac{W_1}{\sum_{i=1}^n W_i} = E * \frac{W_1}{W_1 + W_2 + \dots + W_n}$$

C_2 类抽样数量 S_2 ：

$$S_2 = E * \frac{W_2}{\sum_{i=1}^n W_i} = E * \frac{W_2}{W_1 + W_2 + \cdots + W_n}$$

.....

C_n 类抽样数量 S_n ：

$$S_n = E * \frac{W_n}{\sum_{i=1}^n W_i} = E * \frac{W_n}{W_1 + W_2 + \cdots + W_n} \quad (3.4)$$

$S = (S_1, S_2, \cdots, S_n)$ 为各类属性数据实际抽样量。

(4) 分别对 n 类属性数据依据相关评价标准进行属性数据质量评价，得到各类得分 U ：

$$U = (U_1, U_2, \cdots, U_n)$$

(5) 对各类属性数据得分进行加权综合，得到总体质量得分为：

$$\begin{aligned} S &= W \times U^T = (W_1, W_2, \cdots, W_n) \times (U_1, U_2, \cdots, U_n)^T \\ &= \sum_{i=1}^n W_i * U_i = W_1 * U_1 + W_2 * U_2 + \cdots + W_n * U_n \end{aligned} \quad (3.5)$$

这里 W_i 是每一层的权；

实例：

对某一城市交通图进行属性数据质量评价。由于该图是交通图，所以取其主要几类图形：道路、居民地、植被与其他类。例子中的土地植被边界首先是在 1：5000 比例尺地形图上勾绘出来，其植被类型的划分则是通过 1：5000 比例地形图的判读、航片的叠加以及外业调绘等方法获得，所以精度较差。在结束数据采集后，需要对属性数据的质量给予度量和评价，由于不同类的图形在整个图形中具有不同的地位与重要性，而且属性数据项的数据类型种类很多，所以对其质量进行检验和评价需要采用分类评价模型，可按四类进行分类评价。

(1) 将属性数据分成四类：道路类、居民地类、植被类和其他类。

(2) 各类属性数据的权依据重要性、各类数据量在总数据量中占的比重以及各类的代

表性来确定，对各类属性数据赋予权 W ($\sum_{i=1}^n W_i = 1$) 如下：

表 3.1 各类属性数据的权

Table 3.1 The power of various attribute datas

类别	道路($W_{\text{道路}}$)	居民地($W_{\text{居民地}}$)	植被($W_{\text{植被}}$)	其他类($W_{\text{其他}}$)
权	0.5	0.3	0.1	0.1

(3) 确定各类属性数据的的抽样数量

在此图中，因为权的确定考虑到了数据量、重要性等方面，所以可以按照权确定各类属性数据的抽样数量：

则四类的抽样数量比 C 为：

$$C = W_{\text{道路}} : W_{\text{居民地}} : W_{\text{植被}} : W_{\text{其他}} = 0.5 : 0.3 : 0.1 : 0.1 = 5 : 3 : 1 : 1$$

根据评价费用以及结果可信度要求等确定样本总量为 500，则按上述比例分配到各类属性数据为：

表 3.2 各类属性数据的抽样数量

Table 3.2 sample size of various attribute datas

类别	道路	居民地	植被	其他
抽样总量	250	150	50	50

分别对四类进行属性数据质量评分，最终各类属性数据得分为：

表 3.3 各类属性数据的得分

Table 3.3 Score of various attribute datas

类别	道路	居民地	植被	其他
属性数据质量评价得分	85	80	60	75

(4) 则总体属性数据质量评价得分为：

$$S = W \times U^T = \sum_{i=1}^n W_i * U_i = 85 * 0.5 + 80 * 0.3 + 60 * 0.1 + 75 * 0.1 = 80 \text{分}$$

通过与一般随机抽样方法在抽样精度上的比较，得出：当各分类属性数据质量的缺陷率离散程度较大，即各类数据质量的差别较大时，分类随机抽样有较高的精度。所以按“数据性质相近”的原则分类，可以提高抽样精度。同时得出当各分类缺陷率相差比较大，则采用优化设计的分类评价方案可以获得更好抽样精度的结论。

3.3 提高 GIS 属性数据精度的方法

3.3.1 采用定性概率表达方式提高属性数据精度

在地理信息系统空间数据库中的数据既包括位置数据又包括属性数据,近几年许多学者对定量数据精度问题做了较多的研究,而对定性 GIS 属性数据的精度问题研究甚少。此外,随着遥感(RS)技术的飞速发展,全球定位系统(GPS)、RS 与 GIS 的集成,遥感数据日益成为 GIS 的重要数据源。而 GIS 能有效地存储、管理、分析遥感数据,为遥感图像的解译提供了强有力的支持。众所周知,遥感数据在逻辑运算上具有算法简单、速度快等优点,这得宜于其属性数据大多是基于定性表达的,其随着遥感图像分类技术、图像边缘提取技术的发展,研究和分析定性属性数据的精度及其逻辑运算的精度十分必要。采用概率的方法对此进行分析,运用概率论对定性 GIS 属性数据及其逻辑运算的精度进行分析研究,成为一个研究热点。概率论是一种新型的处理数据的模糊性和不确定性的数学工具,自 1982 年以来,经过十几年的研究与发展,已经在理论和实用上取得了长足的发展。目前,它已经在人工智能、知识与数据发现、模式识别与分类、故障检测等方面得到较为成功的应用。

在图形中经常用多边形来表示空间地物或现象分布的数据组织,而用属性表达多边形所属区域的类型。事实上,一个多边形内经常包含多个类型属性值,而一个属性字段每次只能表示一个值。因此,根据多边形取值方法,如重要性法、面积占优法、中心点法等来确定多边形所属区域的类型,不管使用哪种方法,多边形内都含有了误差。

一般来讲,一个多边形内,不可能仅包含一个属性值,例如土地利用类型,其用地类型错综复杂。假定某一区域的土地类型分为森林、耕地、水面和空地四种,而土地类型图一般是通过解译航空像片或对卫星遥感数据进行计算机处理而得到,一个多边形区域经常会包括多种土地类型。例如森林中的水面,当其面积较小时,往往不单独设立多边形,而直接合并到森林这一土地类型中,这样就带来了属性误差。

若地图上共有 m 类,则需用 m 个字段,每个字段值为该类在相应多边形内所占的比例,即 0~100 范围内的概率值,这种方法叫概率法,能有效地反映多边形内属性的质量状况。

假定 P_{ij} 表示第 i 个多边形中 j 类的概率,若属性类型共分 m 类,则第 i 个多边形可用

概率矢量 $(P_{i1}, P_{i2}, P_{i3}, \dots, P_{im})$ 表示其不确定性。显然 $\sum_{j=1}^m P_{ij} = 1$ ，若 $P_{ij} = 0$ ，表示不含有 j 类，

若 $P_{ij} = 1$ ，则表示该多边形仅含有 j 类。在实际处理中，经常选其概率为最大的相应类别作为该多边形的属性值，显然这样丢失了部分属性信息，从而隐含了属性数据的不确定性。

假定现有一张含有森林、耕地和空地的地图用多边形表示，当多边形内部有零散的其他植被时，往往根据面积占优的原则，取面积最大的地类为多边形的植被，显然部分多边形内含有误差。如图 3.1 所示，在森林和空地多边形内含有耕地。传统的表示法采用面积占优的原则，表示如图 3.2 所示，其简单将区域分成森林，耕地和空地三部分。显然，各部分损失了内部包含的其它信息，如多边形 1 与多边形 2 损失了其内部的耕地信息，而外围耕地部分损失了一部分森林与空地信息。

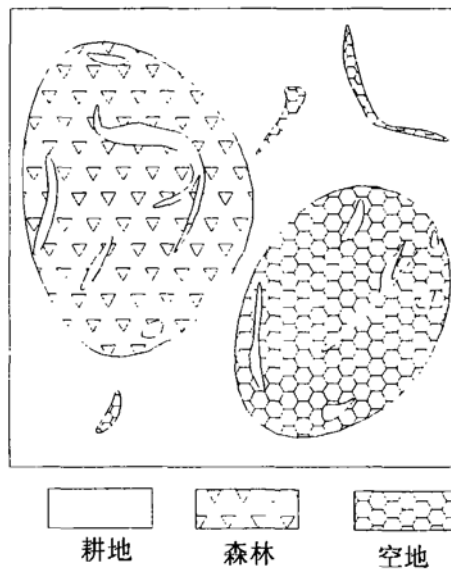


图 3.1 森林、耕地和空地分布图

Fig 3.1 The map of forest, plantation, and wild

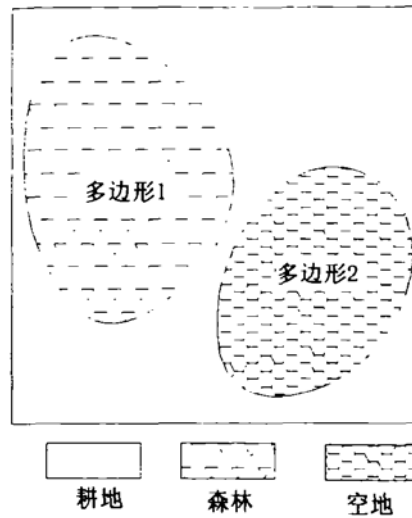


图 3.2 取舍后的森林、耕地和空地分布图

Fig 3.2 The disposed map of forest, plantation, and wild

为此，我们可以采用定性概率表达方式来表示此种类型的信息如下：

多边形 1：森林占 86%，耕地占 14%；

多边形 2：空地占 92%，耕地占 8%；

其余部分：耕地占 95%，空地占 5%；

采用此种属性表达方式，可以使属性数据存储更多的信息，同时数据质量也大为提高。

3.3.2 采用定性概率计算方式提高操作后的属性数据精度

GIS 产品若利用含源误差的数据，进行 GIS 分析操作，将产生新的误差，源误差和操作误差通过 GIS 操作经累计后传播到 GIS 产品中。下面就逻辑叠加关系讨论属性不确定性的传播问题。

假定需要进行 GIS 的空间操作 $y = f(x_1, x_2, \dots, x_m)$ ，其中 $x_i (i = 1, 2, \dots, m)$ 为描述空间数据的自变量，它含源误差； y 为描述 GIS 产品的因变量， $f(x)$ 为描述空间叠加过程的操作函数，则利用误差传播率可知 GIS 产品的相应误差 $D_y = F^T D_x F$ ，其中

$$F^T = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_m} \right], \text{ 为 } x \text{ 的协方差阵。}$$

示例：

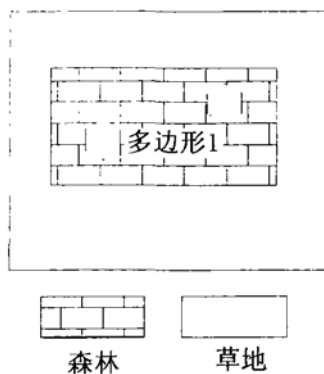


图 3.3 森林、草地分布图

Fig 3.3 The map of forest and meadow

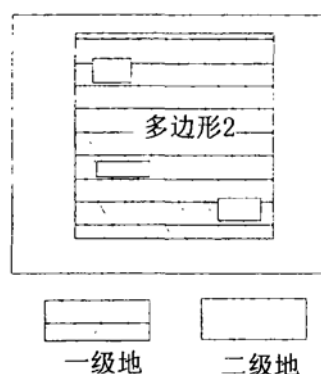


图 3.4 一、二级地分布图

Fig 3.4 The map of first and second grade land

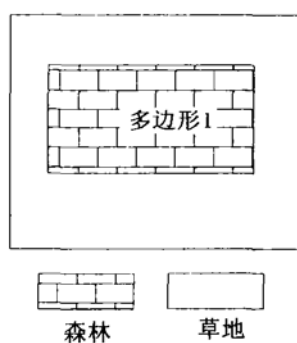


图 3.5 取舍后的森林、草地分布图

Fig 3.5 The disposed map of forest and meadow

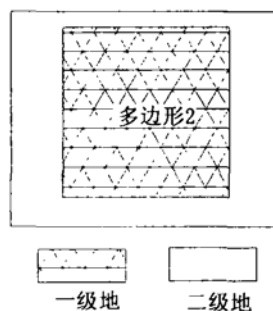


图 3.6 取舍后一、二级地分布图

Fig 3.6 The disposed map of first and second grade land

如图 3.3, 某地区有森林和草地两种植被, 森林被限定在多边形 1 内部。另外, 此地区土壤按肥沃程度高低被划分为一级与二级两种, 如图 3.4。其中, 一级地被限定在多边形 2 内部。但此地区草地与森林, 一级地与二级地交互复杂穿插 (如图 3.3, 3.4), 可以采用两种方法对此类信息加以表达计算:

(1) 面积占优法: (如图 3.5, 3.6)

多边形 1: 森林;

多边形 2：一级地；

当我们在此地区统计既属于森林又属于一级地的区域时，往往用到逻辑叠加操作。当用(1)中的面积占优法存储信息时，因为多边形 1 为森林，多边形 2 为一级地，所以我们将多边形 1 与多边形 2 进行叠加，得到下图 3.7 结果：

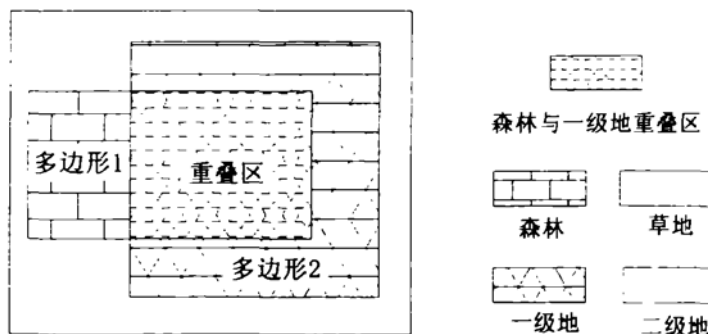


图 3.7 操作后的结果图

Fig 3.7 The result map after operation

从图上可以看出，两个多边形重叠区域为符合要求的查询区域。

(2) 概率法：（如图 3.3, 3.4）

多边形 1：森林占 80%，草地占 20%；

多边形 2：一级地占 60%，二级地占 40%；

当源层中含有误差时，则可用概率作相应的逻辑运算，即可获取经叠加后的结果层的概率，从而确定 GIS 产品的误差。若使用(2)中概率法进行属性信息记录时，我们可以将多边形 1 属于森林的概率设为 80%，而将多边形 2 属于一级地的概率看成 60%，当求二者的交时，由概率论知， $A \cap B = P(A) * P(B) = 80\% * 60\% = 48\%$ ，从操作结果来看，二者叠加后的结果仅以 48% 的概率满足要求，即被高估了 $(1 - 48\%) = 52\%$ 。所以，多边形 1 与多边形 2 的重叠区域不能符合查询要求。

本章小结：

本章提出了一套极为有效的属性数据质量控制措施，并结合属性数据的缺陷类型，提出了一个可靠的质量评价模型：GIS 属性数据质量综合评价模型；针对多类属性数据质量的缺陷率离散程度较大的情况，本章提出了 GIS 属性数据质量分类评价模型，提高了评价准确度。同时，提出了采用定性概率表达方式提高属性数据精度和采用定性概率计算方式提高操作后的属性数据精度两种有效方法。

4 山东省二十五万分之一地名库中属性数据的质量控制与评价

地名数据库是空间定位型的关系数据库,它是将国家基本比例尺地形图上各类地名注记包括居民地、河流、湖泊、山脉、山峰、海洋、岛屿、沙漠、盆地、自然保护区等的名称,连同其汉语拼音及属性特征如类别、政区代码、归属、网格号、交通代码、高程、图幅号、图名、图版年度、更新日期、X坐标、Y坐标、经度、纬度等录入计算机建成的数据库。它与地形数据库之间通过技术接口码连接,可以相互访问,也可以作为单独的关系型数据库运行,形成数字化并具备快速查询、动态管理、图表联动功能,高速保密的地名信息管理系统,达到能瞬间为社会各界提供标准齐全的地名信息功能。

国家基础地理信息系统 1:25 万数据库的重要组成部分——山东省 1:25 万地名数据库研建工作目前已全面展开。山东省 1:25 万地名数据库是一个空间定位与查询型的关系数据库,储存和管理 1:25 万地形图上的各类地名信息与其相关的信息,包括行政区划、坐标、高程和图幅信息、居民地、河流、湖泊、岛礁、山脉、沙漠、盆地、自然保护区的名称及属性等。它既可以与山东省 1:25 万地形数据库连接使用,又可以作为一个独立的专业数据库单独运行。它的建立可以为山东省各部委、各专业部门提供准确的地名信息,特别是为发生各类紧急突发事件的地区提供准确的空间定位信息,做到快速反应。该数据库建成后将收录 1:25 万比例尺地形图上的全部地名和山东省所有建制村以上行政区划地名约 400 万条,乡镇以上行政区划地名还含有部分社会经济信息。

山东省 1:25 万地名数据库建设是一项系统工程。建库工作采取各市(地级、县级市)测绘局(院)同步作业的规模化生产方式,由山东省测绘局牵头组织,中国测绘科学研究院负责技术支持。同时,按照分级建设和分级管理的原则,省级数据库实行集中建设和管理,市级数据库由各市级测绘单位建设和管理。数据库更新实行自下而上的原则,由市级测绘单位负责本区域地名的更新,定期向省级数据库管理单位汇总,进而形成地名数据采集、处理、分发、更新、维护的运行机制。

我省的地名数据库工作将以城市设标和建设数据库为重点,扎实推进地名管理规范化、信息化。目前工作重点是尽快完成城市标准地名设置和建成省市县(市、区)三级地名数据库工作,制定具体方案,尽快完成街道的命名更名,为设标奠定基础;对城市的街、路、巷、院等摸清底数,建立地名信息基础档案,为设标提供可靠依据;全面正

确地编制街路巷和楼门牌号码, 编列地名信息系统, 为设标和日后的地名管理工作提供保证。

建立山东省 1:25 万地名数据库, 既可以改变过去完全由手工制作地名注记的长周期、低效率的生产方式, 解决当前图上地名现势性差的问题, 又可以满足后续更新和 1:100 万、1:250 万地名数据库更新的需要。

4.1 山东省 1:25 万地名数据库的设计

4.1.1 数据库结构设计要点

(1). 数据库数据内容

数据库数据内容分为数据源、数据项与目标数据三种类型。

1) 数据源。数据库使用的数据源有:

①基础数据

基础数据内容包括 1:25 万地形图上的各类地名信息, 近 310 万条, 约 300MB;

②地名所辖空间范围界线。在全区外业采集, 并绘于 1:250000 分幅式图纸上;

③地名所辖空间范围内相关属性数据。按一定的编码记录到对应的调查表。

2) 数据项

数据项如下: 接口码、现图上名称、原图上名称、汉语拼音、语种、类别、政区代码、归属、驻地名、高程、图幅号、图名、出版年代、更新日期、成图单位、经度、纬度、采集单位、总接口码、层名等 30 项。这 30 个数据项不是每个地名都有, 有些地名只有其中几项, 如政区代码、归属、驻地名只在点状和面状居民地地名中有, 高程只在山峰名中有。经度、纬度对于以点状采集的地名要素为图上点状符号的定位点坐标; 对于以面状采集的地名要素为多边形标识点坐标; 当多个多边形为同一地名时, 则取其中主要的一个标识点为定位点。

以上 30 个数据项中 1~6 项直接从 1:25 万地形图抄表, 并使用统一的软件录入到 ORACLE 数据库中; 政区代码、高程、定位点经纬度、归属等数据项从地形数据库中获得, 其他数据项在数据组织和处理时添加。

3) 目标数据。

所有目标数据使用关系数据库管理: 地名数据库数据文件 = 目标空间数据 + 目标属性

数据。目标空间数据与目标属性数据的组织、结构的设计满足第三范式。

(2). 数据库数据模型

该库建立在 ORACLE 关系数据库上, 最终数据格式采用 ORACLE 的对象类型与关系表, 在现有的硬件设备及操作系统平台的基础上, 运用 Aro/INFO 的模块进行库管理, 并为全区建立一个数据库索引文件 INDEX。它具有拓扑结构, 库中数据可以根据单一文件访问, 也可以可作为一个整体进行访问。

(3). 数据库数据的组织

数据分多类, 数据在各类上表达的形式及其格式为点、线与多边形。该设计考虑空间数据与属性数据分别采集, 共同存储。依具体的物理条件, 采用空间数据直接加入属性数据并最终存储于 ORACLE 数据库中的方式, 使用 ARC/INFO 平台+RDBMS 软件实现。ARC/INFO 引入了新的对象关系模型, 可以建立一对多的复合关系。因此, 该设计中考虑单位名属性表(众多的单位或公司的名称)采用关联法与空间数据临时连接, 而不采用空间数据和属性数据 ID 编码的直接加入。关联法适合某一狭小的空间位置而拥有大量的属性值这一在地名数据库中常见的一对多现象, 可以保持空间数据和属性数据直接对接的相互均衡。

(4). 数据库的基本功能实现

使用 ORACLE 在建库时是分类设立的, 但在管理和查询统计时很容易实现多类合一, 也可以在矢量数据上叠加或再产生新类。该库的基本功能包括查询, 检索, 图形显示, 数据提取, 统计分析, 绘图输出等 GIS 常用功能。

4.1.2 主要实施步骤及关键技术

(1) 数据采集

1) 依据数据库的设计, 需要采集下列数据:

- ①基础空间数据; 通过解析地形纸图或模拟生产获得。
- ②地名空间范围数据; 依赖大量详细的外业现场调查记录。
- ③地名属性数据。依赖大量详细的外业现场调查记录。

2) 外业现场调查内容:

- ①地名空间范围调查内容: 院落、小区、独立楼座等构成的多边形; 边界线;
- ②地名属性调查内容: 区域建筑用途; 公共用途地区; 特殊用途区域。
- ③调查方法: 填信息调查表。

④调查表编码信息:分幅图编号+地名属性分类+地名空间顺序号。

(2) 空间数据提取

空间数据提取是实施的关键技术之一,使用 AutoCAD2002 软件,以相邻的矢量数据为背景,按照外业调查的边界,范围分别提取多边形,实体。依照建立的 GIS 空间数据层,分别存储为 DXF 格式。需要特别指出的是每一个提取出的多边形都对应一个惟一的 ID 编码(即该多边形的属性),在 AutoCAD2002 中,应建立文本(text)实体层专门存放提取出的多边形的 ID 编码(可作为点实体存储),为下一步转换成 ORACLE 格式数据做准备。

(3) 空间数据格式转换及数据处理

空间数据格式转换也是实施的关键技术,分下列几步操作:

首先,使用桌面 ARC/INFO 软件的 Arc Toolbox 工具包,打开向导,在向导器中选择 Textto Points,转换 DXF 的 text 实体成为属性表(ID 编码);第 2 步,转换 DXF 多边形实体成为多边形数据;最后使用 ORACLE 的整合操作使点和面层的属性合一,完成空间数据格式转换。

(4) 空间数据与属性数据的连接

使用桌面 ARC/INFO 的 Arc Map 数据处理工具,通过空间数据与属性数据的共同 ID 编码,将分幅图的空间数据与相应图幅的属性数据连接,从产生关系类的向导器进行关联操作,具体可参阅有关的指导书目。

(5) 分幅数据拼接

使用 Arc Toolcox 实现相邻图幅的合并,对每 4 幅数据进行拼接,使跨图幅多边形合并,跨图幅多边形的属性合并同类项,自动建立统一编码,以此,最终完成全省数据的整体拼装。

4.1.3 有关建库问题的探讨

(1) 规范化

地名数据库的组织目前尚无标准可循,需要这些数据的部门不得不根据自己的应用进行空间数据的采集。这种情况不仅造成大量的重复数据建设投资和数据质量低下的现状,而且各部门对信息需求的侧重面不一样,采集的数据内容及组织信息的方式都是自定义的,造成了数据共享和应用系统交互极其困难。因此,迫切需要提供规范化的地名信息以实现信息共享机制。

(2) 分布式设计与建设

建立整个地名数据库由于信息产生和更新的工作量十分巨大,一般需要若干部门同时参与进行这一工作,多个单位分片进行,各自负责每个市区数据的生产与更新,采用分布式地理信息数据库设计(DGDB),包括数据库分片技术,可以按照地域进行分片。又由于地名地理数据库的生产、组织和管理均以图幅为基本单位进行,因此参考地域(如各市区)分界按图幅进行分片应是一种切合实际的分片方案。

(3) 多比例尺数据融合

地名数据库应用是多方面的,多源多比例尺的基础数据和专题图数据的融合更适合城市电子政务发展的需要。关键是如何组织与管理这些海量的地理信息数据,并建立功能完善、反应快速的地理应用。

(4) 历史地名

地理数据与地名信息的变化日新月异,地名信息更新,历史地名的记录,新旧地名的对照,也应是地名信息库设计应该考虑的内容。街、路、巷、地址门牌号、小区、建筑物等一旦发生拆除或改变,在对地名进行维护的同时,要对已改变的历史数据按当时情况加以保留,以方便日后对历史数据的恢复和满足各不同用户对历史数据查询的需要。

(5) 空间数据提取和转换技术

空间数据提取是建库实施的关键步骤之一,使用 AutoCAD 软件和 Arc Toolbox 工具相结合提取和转换的方法,实现图形数据的直接编辑,显示,查询及数据格式转换,操作环节较少,也比较省时,可以提高建库实施的效率。

4.2 山东省 1:25 万地名数据库建立

4.2.1 数据库基本内容

地名数据库主要内容包括乡村、街道、居民院、交通地名及城市构筑物名。不同比例尺与不同地区的地理数据主体要素不尽一致,例如:1:25 万的地名数据库在城市建筑区主要准确地记录了建筑区要素信息。这些建筑区要素信息按建筑类型可划分为平房区和楼房区;按使用功能可划分为居住用途区和非居住用途区;按组合形式可划分为院落区和独立建筑等。与这些建筑区要素信息相联系的是大量的主体:城市居民的居住位置、居住类型信息;城市的功能体:商业,机关,工厂,医院,学校等的位置和类型信息。城市建筑区地名数据库的特点是以构筑物区为主,精度高,更新快,地名点密集,人文信息量大,

应用对象广泛和数据源多样性。

我国多数地区以院落和独立建筑物为基本格局,各基本单位互相连接或由街、巷分开。街路名称采用以条为主,条块结合的规律进行号码的编排,实现号码的序列化。所以,地名管理不能以地形图上的单体建筑为基本单元,而应以地址号区为管理的基本单元(即院落地址和独立建筑地址)。因此,按每个地址院落区或独立建筑的范围提取出边界线作为新图形目标并标注惟一代码,进行图形目标和地名属性的挂接,可以实现街、路、巷、院落、建筑物名称的层次化、群体化和管理的规范化。根据上述原则,地名数据库的基本信息应包括以下内容:

- (1) 空间范围。院落、建筑物、小区的边界线;
- (2) 空间位置。①分区, 邮政编码; ②地址名:街、路、胡同、小区;
- (3) 组合形式。院落与建筑物之间的隶属关系;
- (4) 名称。院落、建筑物、小区的标准名称;
- (5) 使用功能:院落、建筑物的用途;
- (6) 相关属性:
 - ① 居民数量:户数, 人口数量;
 - ② 院落或建筑物内单位名称;
 - ③ 行政管辖:街道办事处、居委会、社区;
 - ④ 图幅编码;
 - ⑤ 其他相关信息。

4.2.2 数据库数据采集

(1) 地名数据采集

地名数据采集工作由各市测绘局按照统一的技术规定分别承担,成果交省基础地理信息中心处理后入库。采用流程如图 4.1:

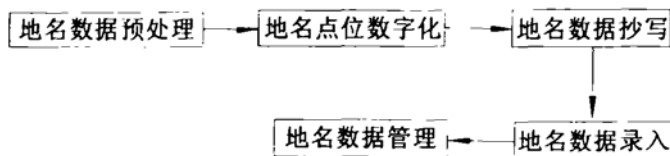


图 4.1 数据采集流程图

Fig 4.1 The flow chart of data collection

以上每一步骤都要进行严格的正确性检查，发现问题及时纠正，以保证整个数据库的优良品质和可信度。

(2) 数字化数据采集

地名数据库的关键是属性数据，这是一般地形图所不能比拟的，建立地名数据库不仅需要图形数据，还需要属性数据相匹配。它的属性覆盖面很广，在建立地名数据库时，地图手工数字化是常用的空间数据采集手段，在这一阶段，除了可以程序化的部分，人工干预的也比较多。为了保证数字化获得数据的质量，在数字化录入数据时必须有一定的规范，要求如下：

① 建筑物

在 1:25 万的地名数据库中，院落、建筑物、街路巷、地址、单元号、单位、小区等之间存在着包含关系，一对多，一对一的关系。这些关系相互制约，而且相互之间的引用是有效的。在建筑物要素采集中，由于建筑物在 GIS 系统中表示为面状，同时要求有相关的房屋属性输入，因此我们在数字化采集中要求作业员对建筑物属性数据的录入做到：

有用途的房屋数据必须有名称，

代码不为零

一般的建筑物楼层不能为零；

建筑物产权单位与使用单位必须明确；

必须存储建筑日期与使用年限等相关信息；

建筑物的用途必须加以区分；

如有资料，将建筑物设计单位，建设单位等信息一并存入；

② 道路

在道路要素采集中，道路在地名数据库系统中表示为线状或面状，要求有相关的属性输入，因此我们在数字化中要求作业员对道路属性数据的数字化采集做到：

道路名称必须准确；

道路要有等级信息；

道路边线的级别应与道路中心线同等级

同一条路接边的道路边线等级应相同

一条路赋予一个路宽，同时一个名称的道路应连续

如有资料，将道路建设日期，设计单位，建设单位等信息一并存入；

道路跨越地区，各段地质构造等信息放入相关信息段中；

路面材料等信息一并录入到相关字段;

属性信息与相关空间图形信息的连接必须正确;

③ 分幅

地名数据库的建立是在多种比例地形图的基础上的,对与每一种地形要素,其属性数据的质量控制措施也不尽相同。为了对地名数据库进行分区管理,我们对数据库进行分幅,而每一图幅必须保证:

采用的平面坐标系和高程基准符合要求;

图廓点、公里网、控制点等信息正确,

坐标、高程表示正确

④ 拓扑关系

空间对象的基本关系可分为点与点、点与线、点与面、线与线、线与面、面与面之间的关系,从空间对象关系的类型来看,可以成分离关系、相接关系、重叠关系、包含关系、等同关系等。在地名数据库中,对空间数据与属性数据之间的质量控制,要求两者之间进行质量检查和控制,以保证存放空间数据拓扑关系的属性数据的正确关联。为使后续系统分析和应用可靠,对存放空间数据拓扑关系的属性数据进行相应的质量控制,是保证拓扑关系正确的基础。由于人工操作的误差,不可避免地会破坏空间对象之间的正确关系,因此在数据进行入库时必须根据空间对象之间满足的一定关系对存放空间数据拓扑关系的属性数据进行拓扑检查,主要包括:重复点、重复线、包含线的删除、面状地物的严格闭合,合并交叉点、合并交叉线等。另外,对空间对象的关系必须注意的是:对于面状地物必须保证边界是封闭的,以免建立拓扑关系后丢失数据,对于各个属性的注记必须保证在对应地物的内部,对于各共享公共点和公共边的空间对象,应严格使公共边或点正确唯一。

4.2.3 地名数据组织与处理

(1) 地名数据的分项提交

1:25 万地名数据在数据采集方案制定时,为了各市测绘局作业方便,减少采集工作量,规定将能够从地形数据中获取的地名数据项暂不采集,各市测绘局上交的地名数据只包括地名库中的 ID 码,现图上名称、原图上名称、汉语拼音、语种、类别等 6 项,其他数据项需从地形库中直接或间接获取。因此,在建库前首先要对地名数据进行重新组织与处理,编制一些批处理程序,保证数据处理的质量与速度。可直接获取的地名数据项包括政

区代码、山峰高程、地名定位点的经纬度、层名等；还有一些地名数据项可通过地形数据与地名数据的关系建立一些表格文件,间接获取数据,包括居民地的归属、驻地名等；其他如图幅号、出版年代、更新日期、成图单位、采集单位等项,建立一个以图幅为单位的数据文件,不加在每个地名记录中,以减少数据冗余。按照 1:25 万地名数据库的设计,用于数据库管理方面和加快检索速度的一些辅助数据文件和数据项,也将从地形数据中获得,如图幅左下角经纬度、政区中心点坐标等。

从地形库中获取地名数据内容以图幅为单位进行,利用 Arc/Info 软件的 Info 属性数据库管理功能,将地名数据与地形数据结合,完成数据获取。

(2) 地名数据检查修改及格式转换

在数据入库前,必须对重新处理后的地名数据作全面检查。在地名数据重新组织与处理过程中,牵扯到的内容较多,数据量较大,稍有不慎容易出现错误,因此可以编一些检查程序进行自动检查。

另外,必须按照设计的数据结构将整理好的数据文件转换成 ORACLE 数据库可以接受的格式。

(3) 地名数据入库

地名数据入库的步骤是:生成数据库(名称为 NGIS25),生成表空间、定义表结构、定义视图、数据入库,建立索引等,数据入库主要采用文件插入方式。

数据入库后分别按照图幅号、政区代码、地名类别建立索引。

在已开发库外汉字库的基础上,对原采集的地名数据中用代码表示库外字的地方替换成汉字和相应的汉语拼音。

4.2.4 数据表结构设计

山东省 1:25 万地名库实体建立在 SUN1000E 服务器上,采用 Oracle9I 关系数据库进行管理。数据库应用管理系统采用 SQL 开发环境在 PC 机上开发,数据库和用户具有 Client/Server 体系结构,用户在 PC 机上实现对地名库的访问。在数据库中是以记录坐标的方式来表示点、线、面、体的位置及其空间关系。一般包括空间数据库和属性数据库,两者可通过内部惟一的标识码进行连接。

为将地名信息按照其结构进行设计组织,进行有效地管理、检索。地名库表结构设计如下:

建立图幅信息、行政区划信息、地名信息、图幅与政区关系、地名类别对照、行政

区划与政区代码对照六个表,前四个表为基本信息表,后两个表为辅助信息表。

第 1 个表——与图幅有关的信息。字段:图幅号、图名、出版年代、更新日期、成图单位、左下角经度、左下角纬度、采集单位。

第 2 个表——行政区划信息。字段:政区名称、原政区名称、政区代码、汉语拼音、语种、类别、驻地名称、驻地汉语拼音、驻地语种、驻地经度、驻地纬度。

第 3 个表——地名信息。字段:现图上名称、原图上名称、汉语拼音、语种、类别、高程、政区代码、经度、纬度、总接口码、层名。

第 4 个表——建立图幅与行政区域对应关系。字段:图幅号、政区名称、政区代码。

第 5 个表——地名类别对照。建立地名分类体系。

第 6 个表——行政区划与政区代码对照表。使行政区划与其政区代码相互对应,便于建立统一政区体系。

上述表与表之间都有直接或间接的关联项,为满足查询需要,方便使用,还应建立视图和索引,以及相关字段的数据字典。

4.2.5 管理系统设计与数据库的应用

根据山东省 1:25 万地名数据库的建库目标和要求,方便查询及结果输出,应用管理系统功能如下:

(1) 输入功能。可以从键盘或以数据文件形式输入各种地名信息,包括地名定位点坐标。

(2) 查询统计功能。可以对数据库中的任意数据项进行组合条件查询、模糊查询,并对查询结果进行统计。

(3) 输出功能。查询结果可以按照所需格式输出,如单条地名显示,打印地名表,地名卡片或输出数据文件。

(4) 数据库维护功能。

用户按照所授予的权限使用数据库,对数据库进行增、删、改、更新、备份等操作,以保证数据安全。地名数据库应用管理系统在 Oracle 提供的 PC 开发环境下按设计要求开发,建立弹出式菜单界面,界面友好,使用方便。

由于 1:25 万地名数据库检索速度快,又包含了空间定位信息,所以它在一些紧急突发事件的迅速定位方面发挥了很大作用,同时也弥补了现有 1:25 万地形数据库全库检索速度较慢的缺憾。我们已经为多家单位提供了山东省或各市的 1:25 万地名数据服务,在

数据提供过程中,我们也体会到目前 1:25 万地名数据库中包含的信息量是很大的,涉及到人口、工农业产值、城市建设、信息共享等。其在国民经济建设中发挥了巨大的作用,得到许多用户的青睐。

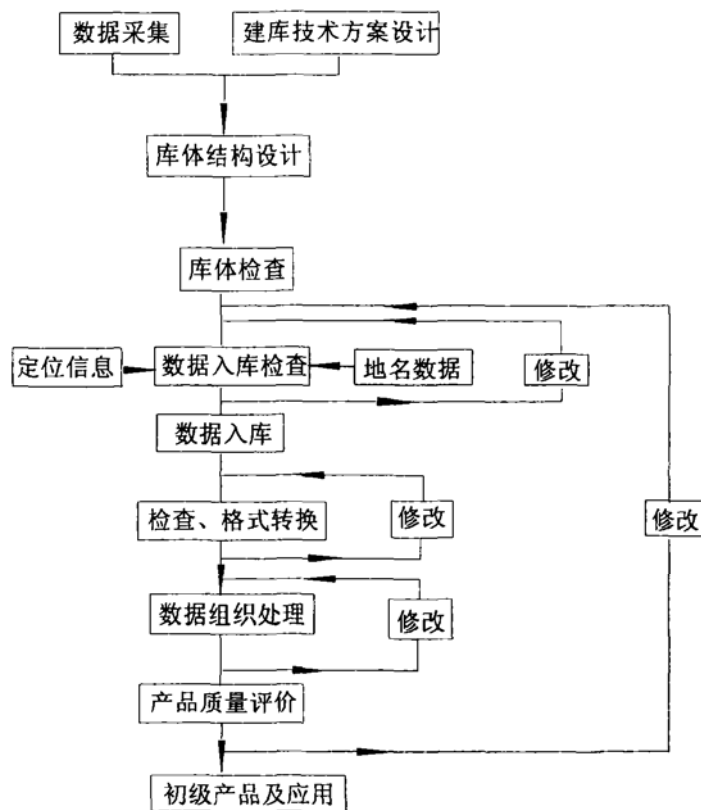


图 4.2 建库流程图

Fig 4.2 The flow chart of database establishment

由于关系型数据库的应用发展已经深入到广大的用户,其技术发展和广泛应用使其成为一种新的标准。使用关系型数据库技术解决空间地理数据库问题是当前 GIS 数据库发展的一个方向。山东省 1:25 万地名数据库的设计和建立,对将空间地理数据中的地名转到关系型数据库中的过程做了一次成功的探索,也是使用关系型数据库技术解决空间地理数据库问题做了一次有益的尝试,为将来大型空间地理数据库的建设积累了丰富的经验。

4.3 山东省 1:25 万地名数据库属性数据质量控制

山东省 1:25 万地名数据库是以 1:2000 地形图为主要的数据源,将地形图要素按照

一定的分类和编码系统分为十大类，并赋予相应的编码。属性数据以人工方式输入数据库中，并通过数据转换的方式进行 GIS 数据建库，在这一系列的过程中，数据质量的检查与验收贯穿始终。

4.3.1 误差来源

山东省 1:25 万地名数据库数据质量控制是指控制数据的可靠性和精度，因此有必要了解建库过程中的误差来源，在建库过程中数据的误差可分为源误差，处理误差和应用误差等三种类型，现分析如下：

(1) 源误差

源误差是指数据采集和录入中产生的误差，包括遥感数据、测量数据、属性记录、GPS 测量、制图、数字化等的误差。

(2) 处理误差

处理误差指数据录入后进行数据处理过程中产生的误差、包括几何改正、坐标变换和比例变换、几何数据的编辑、属性数据的编辑、空间分析、图形简化（数据压缩和曲线光滑）、数据格式转换、地形数据模型化、计算机截断误差等。

(3) 使用误差

使用误差是指数据被使用过程中出现的误差、包括数据的完备程度、时间的有效性（现势性）、拓扑关系的正确性、缺乏数据的质量报告等。

以上三种误差中，数据处理误差远远小于数据源的误差，使用误差属于数据本身的误差，但这些因素直接影响到应用的效果，所以也应列为数据误差的范畴。在数据质量中应考虑多方面的因素，对数据质量控制的主要任务是针对可能产生误差的各个环节进行必要的控制，尽量减弱这些误差的影响，同时对数据的质量进行检验和评价。

4.3.2 质量控制

(1) 空间数据质量控制

空间定位数据成果包括原始数字化定位数据、数字地图。原始数字化定位数据质量内容包括数据格式、定位精度及完整性。针对 GIS 系统中具有空间实体、描述属性、空间实体间拓扑关系三大特点的，山东省 1:25 万地名数据库的质量控制是贯穿于资料准备、仪器准备、采集、编辑、处理、输出全过程的，其中空间数据质量控制包括：

①空间实体的点线面类型定义必须正确。

②保证多边形空间实体的封闭性。

③保证线状空间实体的连接性。

④保证组合实体与基础图形要素之间的正确相关性。

⑤保证编码不重复。

⑥保证符合组合实体间和基础图形要素之间的关系原则或制约，其中包括：

连接(如墙可与墙连接，但不可与铁路线连接)、相交(如墙可与道路边线相交)、共享(如墙可与台阶、楼梯、道路边线共享)、落于(如建筑物中心点应落于墙内)、包含(如墙可包含建筑物中心线)

⑦保证不出现悬挂节点和伪节点，所有线状要素相交处，都应建立节点。

⑧每个线状要素必须既有从左侧相关于它的面域，又有从右侧相关于它的面域。

⑨数据格式必须符合“格式与代码”的规定；

⑩直线地物、独立地物的定位误差不得超过 0.2 米，曲线地物的误差不得超过 0.3 米。

保证空间数据之间拓扑关系的正确性是实现 GIS 系统数据处理和分析的基础，它的质量特性是 GIS 系统区别于 CAD 系统的重要内容之一。

(2) 属性数据质量控制

保证属性数据的正确性是实现山东省 1:25 万地名数据库的基础，它的质量特性是数据库评价的重要内容之一。从其编辑处理过程来看，要保证属性数据的质量，必须从以下几方面考虑：

1 数据源

被录入的信息与理论数据存在的误差进行校正。

2 采集方法和使用设备的有关参数

应描述清楚采集的过程，差错改正的情况，编辑处理的要素以及采用的扫描仪、手持数字化仪、绘图机、打印机等设备的型号、分辨率、参数设置等方面的情况。

3 分类属性数据设置

分类表示是对各种不同地物地貌要素进行组织、满足不同应用需求、甚至是输出不同图饰符号的手段。每个属性数据必须按定义赋予一定的值(或号)，同一类属性数据不能赋予一个以上图类，每一类只有一种属性数据。

4 完备性

GIS 系统中的属性数据，数据必须完备，每组数据文件应该完整，注记不得有错漏

和偏移。

5 数据格式

描述空间数据的属性项定义(包括名称、类型、长度等)必须正确;

属性表中各数据项的属性取值及其单位不得有异常;

数据项的定义、关键列、连接列的定义,地理视图的确定结构合理;

6 唯一性

标识码是区分标识空间数据的码,必须唯一有效、不重复。描述每个地理实体特征的各种属性数据应正确。

关键属性数据必须具有唯一性,不得有重复重叠。

7 连接无缝

属性数据可以按街坊、图幅、等进行采集,所有需要连接的数据不能出现逻辑裂隙,应满足相应规范的要求;

空间数据与描述它的属性数据之间一一对应的关系须正确,也即空间数据与属性数据必须正确相关;

具有一个以上属性表时,各属性表之间的相关性和网络层次应当正确描述和建立;

8 属性输出质量

地理信息系统的一个重要功能就是能输出模拟地形图。因此,输出的地形图必须清晰、易读,属性显示符合图式的规定,而且内容要完整。

9 元数据的质量控制

要保证属性元数据数据项齐全,内容正确、无遗漏。

10 数据转换无损

因数据来源不同,分辨率不同和时相不同,在属性数据入库过程中经常需进行不同数据之间的转换,转换过程中要保证数据精度、空间数据与属性数据的联接关系不会丢失。

11 数字化技术

地名数据库中的许多数据来自地图数字化,所以在地图数字化每一道工序应严格按照标准操作,实施检查验收,以确保数据的完整性、逻辑一致性及可靠性。

数字化技术报告是数据质量控制的主要手段,数字化工作者负责提供地图数字化的技术报告。内容包括被采集的数据源的说明,采集方法及使用设备的有关参数,生成最终数据文件的情况及格式转换方法的说明,资料清单等。

12 资料清单

包括原图、磁盘数据、检查图、图历薄, 以及检查验收结果和必须说明的情况。

4.3.3 已入库属性数据的二次匹配检验

对于已入库的属性数据, 我们必须进行二次检查, 确保整个地名数据库总体质量(如图 4.3)。属性数据是首先在图形数据采集时放入相关数据库中的, 然后再通过处理与图形数据库连结到一起。由于这一系列的处理是在 GIS 软件 ARC/INFO 中完成的, 因此我们对山东省 1:25 万地名数据库中属性数据的最终二次质量检查也将在 GIS 软件平台 ARC/INFO 上完成。

属性数据的质量控制应包括属性数据的逻辑检查和数据库记录的抽样检验。属性数据的逻辑检查是根据数据之间存在的逻辑关系对数据进行检查, 这些逻辑检查的方法是通过编制数据库检查程序, 自动进行报告检查情况, 进行交互式改正。

属性数据二次质量检查内容

(1) 属性值域的检查

用属性模板自动检查要素层中每个目标的主码、识别码、描述码、参数值的值域是否正确。对不符合属性模板的属性项在相应位置作错误标记, 并记入属性错误统计表。

(2) 属性值逻辑组合正确性检查

用属性值逻辑组合模板检查要素层中每个数字化目标的属性组合是否有逻辑错误, 是否按有关技术规定正确描述了目标的质量、数量及其它信息。

(3) 用符号化方法对各属性值进行详细检查

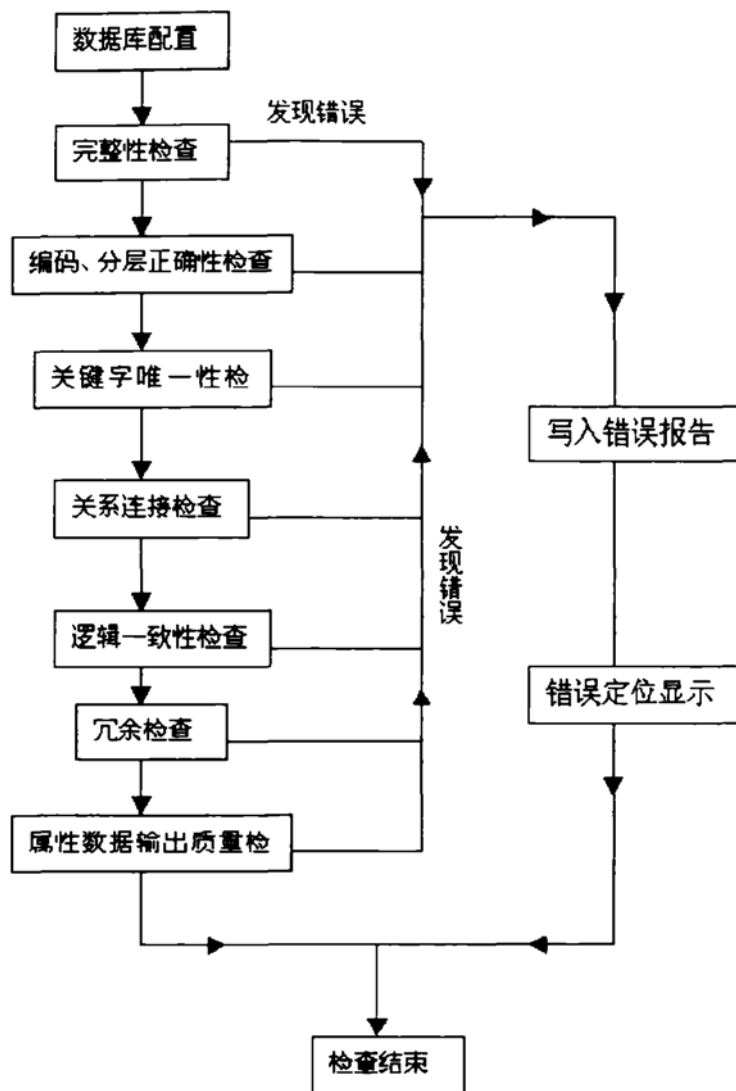


图 4.3 属性数据二次检查流程图

Fig 4.3 The flow chart of second check of attribute data

针对属性数据质量评价的特点，制定与图式规范尽量一致，又有利于目标识别和理解的符号化方案，可以较好满足属性数据评价的要求。符号化使图形相对定位简单易行，方便了人机交互检查作业。符号化表示时，属于同一主码的目标显示在同一层次上；把识别码分成点、线、面图形，分别对应点状、线状和面状符号库，用图式规定的符号及颜色，配合符号库解释规则把识别码解释成图形；描述码同识别码相结合，有些改变图形的表示方法，如建筑中的铁路用虚线符号表示；有些改变颜色，如不依比例圈形居民地用黑色表示，县级用绿色，省级用红色等；有些注记汉字，如时令河，在线划上注记“时令”两个汉字；要素所带参数用数字的形式注记出来，用颜色区分参数的类别，用红色表示宽度参数，用黑色表示相对高参数，用蓝色表示长度参数，用棕色表示其它参

数。对错误,用人机交互的方法在图上作标记,并记入属性错误统计表。对注记(包括名称注记和说明注记)的拼写、字体、角度、大小以及位置原则上要求与原来样图一致。当与相关要素发生矛盾时,按常规编图的等级顺序处理,可适当调整位置。整饰,增添符号或线型,增添部分按其定义的属性存储在相对应的层上。

4.4 山东省 1:25 万地名数据库中属性数据质量评价

在完成地名数据库所有工作以后,为了对整个地名数据库中的属性数据的质量进行度量,对属性数据的整体使用性进行评价,我们对属性数据进行抽样验收检查,以保证良好的数据质量,服务于各个信息系统的分析和统计。我们主要采用计数型分类抽样评价的方式,来评价山东省 1:25 万地名数据库中总体属性数据的质量。

4.4.1 确定属性数据分类

一般随机抽样是最基本的抽样方法,但是对于大批量的检验总体,有时总体各部分质量的实际分布差别很大,对不同部分采用不同的检验程序就比较符合实际。

由于山东省 1:25 万地名数据库中属性数据量非常大,同时具有区域的特点,各种地名在数据库中所占比例是不一样,可引入分类概念,对总体数据分类评价,评价每一类的质量,然后将各类的质量得分通过适当加权就能得到总体质量的估计值。

由于山东省地域广阔,既有沿海发达城市,又有内陆欠发达地区,所以在城市规划、行政区域开发等方面侧重点就不一样,有些区域注重开发工业用地,有些区域注重农业用地的利用,有些区域注重居住用地的开发等。以上情况造成一定区域所拥有的属性数据类型差别比较大,综合考虑各种差别后,本文将山东省 1:25 万地名数据库中属性数据分成如下几类:

表 4.1 属性数据分类

Table 4.1 The classification of attribute data

类别	居民地	工业设施	农业设施	行政区划	道路	植被	水系	其他
----	-----	------	------	------	----	----	----	----

4.4.2 确定属性数据评价项及其在不同类中的权

综合考虑山东省 1:25 万地名数据库中属性数据的采集、转换、存储和应用等方面因素,在此次评价中,采用如下十个评价项对属性数据质量进行总体评价:

表 4.2 属性数据评价项

Table 4.2 The inspection items of attribute data

评价项	数据	完备性	唯一性	连接列	属性输
	格式			关键字	出质量
评价项	元数据	数据转	属性数据	描述信息	表格
	质量	换无损	定义	正确性	正确性

由于在不同类的属性数据中,评价项的重要性、代表性和所处地位不同,依据各类特点与评价项对本类的重要性来确定评价项在各类属性数据中的权,更能提高评价质量。各评价项在不同类中的权如下表:

表 4.3 居民地类中各评价项的权

Table 4.3 The power of inspection item in residential area class

评价项	数据	完备性	唯一性	连接列	属性输
类别	格式			关键字	出质量
权	0.12	0.11	0.12	0.1	0.05
评价项	元数据	数据转	属性数据	描述信息	表格
类别	质量	换无损	定义	正确性	正确性
权	0.18	0.04	0.1	0.08	0.1

表 4.4 工业设施类中各评价项的权

Table 4.4 The power of inspection item in industrial establishment class

评价项	数据	完备性	唯一性	连接列	属性输
类别	格式			关键字	出质量
权	0.12	0.12	0.16	0.13	0.08
评价项	元数据	数据转	属性数据	描述信息	表格
类别	质量	换无损	定义	正确性	正确性
权	0.11	0.05	0.06	0.1	0.07

表 4.5 农业设施类中各评价项的权

Table 4.5 The power of inspection item in agriculture establishment class

评价项	数据	完备性	唯一性	连接列	属性输
类别	格式			关键字	出质量

权	0.12	0.12	0.15	0.13	0.08
评价项	元数据	数据转	属性数据	描述信息	表格
类别	质量	换无损	定义	正确性	正确性
权	0.11	0.05	0.07	0.1	0.07

表 4.6 行政区划类中各评价项的权

Table 4.6 The power of inspection item in administration section class

评价项	数据	完备性	唯一性	连接列	属性输
类别	格式			关键字	出质量
权	0.06	0.14	0.18	0.11	0.06
评价项	元数据	数据转	属性数据	描述信息	表格
类别	质量	换无损	定义	正确性	正确性
权	0.13	0.06	0.09	0.09	0.08

表 4.7 道路类中各评价项的权

Table 4.7 The power of inspection item in road class

评价项	数据	完备性	唯一性	连接列	属性输
类别	格式			关键字	出质量
权	0.14	0.15	0.2	0.14	0.07
评价项	元数据	数据转	属性数据	描述信息	表格
类别	质量	换无损	定义	正确性	正确性
权	0.07	0.06	0.04	0.1	0.03

表 4.8 植被类中各评价项的权

Table 4.8 The power of inspection item in vegetation class

评价项	数据	完备性	唯一性	连接列	属性输
类别	格式			关键字	出质量
权	0.16	0.14	0.03	0.14	0.17
评价项	元数据	数据转	属性数据	描述信息	表格
类别	质量	换无损	定义	正确性	正确性
权	0.1	0.1	0.03	0.11	0.02

表 4.9 水系类中各评价项的权

Table 4.9 The power of inspection item in water system class

评价项 类别	数据 格式	完备性	唯一性	连接列 关键字	属性输 出质量
权	0.16	0.18	0.03	0.11	0.14
评价项 类别	元数据 质量	数据转 换无损	属性数据 定义	描述信息 正确性	表格 正确性
权	0.12	0.04	0.06	0.14	0.02

表 4.10 其他类中各评价项的权

Table 4.10 The power of inspection item in other class

评价项 类别	数据 格式	完备性	唯一性	连接列 关键字	属性输 出质量
权	0.15	0.17	0.11	0.1	0.11
评价项 类别	元数据 质量	数据转 换无损	属性数据 定义	描述信息 正确性	表格 正确性
权	0.1	0.05	0.05	0.11	0.05

4.4.3 确定不同类中属性数据评价项所占分值

不同类中属性数据评价项所处地位不同,重要性不同,为了最大限度减少评价误差,必须合理分配不同类中评价项所占分值。而不同类中各评价项的权是根据其重要性、评价指标和评价项对本类的代表性而确定的,所以依据评价项的权来分配评价项所占分值是合理的,具体采用如下方法:

假设共有 m 类和 n 个评价项,则:

评价项 P_1 在类 C_1 中所占分值为 T_{C1-1} : $T_{C1-1} = 100 * W_{C1-1}$

评价项 P_2 在类 C_1 中所占分值为 T_{C1-2} : $T_{C1-2} = 100 * W_{C1-2}$

.....

评价项 P_i 在类 C_j 中所占分值为 T_{Cj-i} : $T_{Cj-i} = 100 * W_{Cj-i}$

$$W = \{\{W_{C1-1}, W_{C1-2}, \dots, W_{C1-n}\}, \{W_{C2-1}, W_{C2-2}, \dots, W_{C2-n}\}, \dots, \{W_{Cm-1}, W_{Cm-2}, \dots, W_{Cm-n}\}\}$$

W 为评价项在各类属性数据中的权集合,其中 W_{Cj-i} 为第 i 个评价项在 C_j 类中的权。

在山东省 1:25 万地名数据库中, 依据上面分类, 我们取 $m=18$, $n=10$, 则按照上面公式计算各评价项在不同类中所占分值如下:

表 4.11 各评价项在居民地类中所占分值

Table 4.11 The account of each item in residential area class

评价项 类别	数据 格式	完备性	唯一性	连接列 关键字	属性输 出质量
总分	12	11	12	10	5
评价项 类别	元数据 质量	数据转 换无损	属性数据 定义	描述信息 正确性	表格 正确性
总分	18	4	10	8	10

表 4.12 各评价项在工业设施类中所占分值

Table 4.12 The account of each item in industrial establishment class

评价项 类别	数据 格式	完备性	唯一性	连接列 关键字	属性输 出质量
总分	12	12	16	13	8
评价项 类别	元数据 质量	数据转 换无损	属性数据 定义	描述信息 正确性	表格 正确性
总分	11	5	6	10	7

表 4.13 各评价项在农业设施类中所占分值

Table 4.13 The account of each item in agriculture establishment class

评价项 类别	数据 格式	完备性	唯一性	连接列 关键字	属性输 出质量
总分	12	12	15	13	8
评价项 类别	元数据 质量	数据转 换无损	属性数据 定义	描述信息 正确性	表格 正确性
总分	11	5	7	10	7

表 4.14 各评价项在行政区划类中所占分值

Table 4.14 The account of each item in administration section class

评价项	数据	完备性	唯一性	连接列	属性输
-----	----	-----	-----	-----	-----

类别	格式			关键字	出质量
总分	6	14	18	11	6
评价项 类别	元数据 质量	数据转 换无损	属性数据 定义	描述信息 正确性	表格 正确性
总分	13	6	9	9	8

表 4.15 各评价项在道路类中所占分值

Table 4.15 The account of each item in road class

评价项 类别	数据 格式	完备性	唯一性	连接列 关键字	属性输 出质量
总分	14	15	20	14	7
评价项 类别	元数据 质量	数据转 换无损	属性数据 定义	描述信息 正确性	表格 正确性
总分	7	6	4	10	3

表 4.16 各评价项在植被类中所占分值

Table 4.16 The account of each item in vegetation class

评价项 类别	数据 格式	完备性	唯一性	连接列 关键字	属性输 出质量
总分	16	14	3	14	17
评价项 类别	元数据 质量	数据转 换无损	属性数据 定义	描述信息 正确性	表格 正确性
总分	10	10	3	11	2

表 4.17 各评价项在水系类中所占分值

Table 4.17 The account of each item in water system class

评价项 类别	数据 格式	完备性	唯一性	连接列 关键字	属性输 出质量
总分	16	18	3	11	14
评价项 类别	元数据 质量	数据转 换无损	属性数据 定义	描述信息 正确性	表格 正确性
总分	12	4	6	14	2

表 4.18 各评价项在其他类中所占分值

Table 4.18 The account of each item in other class

评价项 类别	数据 格式	完备性	唯一性	连接列 关键字	属性输 出质量
总分	15	17	11	10	11
评价项 类别	元数据 质量	数据转 换无损	属性数据 定义	描述信息 正确性	表格 正确性
总分	10	5	5	11	5

4.4.4 确定属性数据评价项的评分标准

对属性数据每一评价项进行评分时，必须具有一定的标准。属性数据具有的缺陷类型可以分为轻缺陷、重缺陷和严重缺陷 3 类。评价项具有的缺陷类型不同，对整体数据质量的影响也不同。缺陷类型的划分一方面根据国家相关数据质量标准，另一方面需要考虑系统对数据的特别要求。针对山东省 1:25 万地名数据库的特点，结合缺陷类型划分的国家相关数据质量标准，在此给出各评价项的评分标准，对每一个评价项采用 A、B、C、D、E 五等进行评分：

评价项被评为 A 等应满足的条件如下表：

表 4.19 评价项被评为 A 等应满足的条件

Table 4.19 The qualification of grade A to each item

评价项	评价标准
数据格式	1. 属性数据格式符合“格式与代码”的规定； 2. 属性数据格式符合主题、专题要求，且条理清楚； 3. 分类正确。每个属性数据归于一类，同一属性数据没有赋予一个以上的类； 4. 每一类只有一种属性数据； 5. 属性数据格式满足面向对象特性，具有良好的向后兼容性与扩充性；
完备性	1. 每组数据文件完整，没有错漏； 2. 最终数据文件名称清楚，分类合理，各类信息的统计结果以及采用的数据格式与设计要求格式无差异； 3. 属性数据库运行稳定可靠；

唯一性	1. 连接码唯一有效、不重复; 2. 标识码唯一有效、不重复; 3. 关键属性数据具有唯一性, 没有重复重叠; 4. 地名具有唯一性;
连接列关键字	1. 所有需要连接的数据没有出现逻辑裂隙; 2. 满足相应连接规范的要求; 3. 可以按图幅连接表达; 4. 属性数据与空间数据连接可靠;
属性输出质量	1. 能充分表示各种不同地物地貌要素, 满足应用需求, 能完整输出不同图饰符号; 2. 能完整正确输出模拟地形图; 3. 输出的地形图清晰、易读、完整; 4. 属性显示符合图式的规定, 而且内容完整;
元数据质量	1. 属性元数据数据项齐全, 内容正确、无遗漏; 2. 属性元数据数据项表达合理、清晰; 3. 元数据满足面向对象特性, 具有良好的向后兼容性; 4. 元数据能支持各项查询、修改等操作;
数据转换无损	1. 通过转换而来的数据没有信息损失, 现有数据可以无损地转换为其他类型数据; 2. 转换过程中数据精度、空间数据与属性数据的连接关系没有丢失; 3. 压缩与解压数据正确无损; 4. 数字化获得的数据完整可靠, 被录入的信息与理论数据存在的误差经过了严格校正; 5. 采用的扫描仪、手扶数字化仪、绘图机、打印机等设备的型号、分辨率、参数设置等方面符合规范要求;
属性数据定义	1. 属性数据定义简练, 且能充分表达出要表达的信息; 2. 属性数据的定义符合自然语言, 容易理解; 3. 属性数据的定义便于进行模糊查询;
描述信息正确性	1. 颜色分配合理、美观且能表达一定的分类信息;

	2. 线型符合“格式与规范”的要求，且可以表达相关信息； 3. 已有规定的符号严格按照图式绘制； 4. 需要自行设计的符号应与实际地物类似，且简单易绘； 5. 注记格式正确，大小合适，内容简练完整；
表格正确性	1. 各种表格字段清晰，内容正确； 2. 表格设计符合第一、二、三范式，没有冗余； 3. 表格之间连接可靠；

评价项被评为 C 等应满足的条件如下表：

表 4.20 评价项被评为 C 等应满足的条件

Table 4.20 The qualification of grade C to each item

评价项	评价标准
数据格式	1. 数据格式基本符合“格式与代码”的规定； 2. 属性数据格式基本符合主题、专题要求； 3. 分类正确。每个属性数据归于一类；
完备性	1. 每组数据文件基本没有错漏； 2. 最终数据文件名称清楚，分类合理，各类信息的统计结果以及采用的数据格式与设计要求格式基本无差异；
唯一性	1. 连接码、标识码基本唯一有效、不重复； 2. 关键属性数据具有唯一性，基本没有重复重叠； 3. 地名具有唯一性；
连接列关键字	1. 所有需要连接的数据基本没有出现逻辑裂隙； 2. 可以按图幅连接表达； 3. 属性数据与空间数据连接基本可靠；
属性输出质量	1. 能表示多种不同地物地貌要素，能完整输出不同图饰符号； 2. 输出的地形图清晰、易读； 3. 属性显示基本符合图式的规定；
元数据质量	1. 属性元数据数据项基本齐全，无遗漏； 2. 属性元数据数据项表达基本正确、合理； 3. 元数据能支持基本各项查询、修改等操作；

数据转换无损	1. 通过转换而来的数据基本没有信息损失; 2. 转换过程中空间数据与属性数据的连接关系没有丢失; 3. 数字化获得的数据完整可靠, 被录入的信息与理论数据存在的误差经过了初步校正;
属性数据定义	1. 属性数据定义简练, 且能基本表达出要表达的信息; 2. 属性数据的定义基于自然语言, 容易理解; 3. 属性数据的定义可以进行模糊查询;
描述信息正确性	1. 颜色分配合理且能表达一定的分类信息; 2. 线型基本符合“格式与规范”的要求, 且可以表达相关信息; 3. 已有规定的符号基本按照图式绘制; 4. 需要自行设计的符号应与实际地物类似;
表格正确性	1. 各种表格字段清晰, 内容正确; 2. 表格设计符合第一、二范式;

评价项被评为 E 等 (不合格) 应满足的条件如下表:

表 4.21 评价项被评为 E 等应满足的条件

Table 4.21 The qualification of grade E to each item

评价项	评价标准
数据格式	1. 性数据格式仅 85% 以下符合“格式与代码”的规定; 2. 属性数据格式 80% 以下符合主题、专题要求; 3. 仅有 83% 以下的属性数据能归于一类; 4. 仅有 87% 以下的类只有一种属性数据;
完备性	1. 85% 以下的数据文件完整, 没有错漏; 2. 各类信息的统计结果以及采用的数据格式与设计要求格式有较大差异;
唯一性	1. 90% 以下连接码、标识码唯一有效; 2. 93% 以下关键属性数据具有唯一性, 没有重复重叠;
连接列关键字	1. 87% 以下需要连接的数据没有出现逻辑裂隙; 2. 89% 以下需要连接的数据满足相应连接规范的要求;
属性输出质量	1. 仅能完整输出要求的 92% 以下地物地貌要素与图饰符号; 2. 仅 92% 以下属性显示符合图式的规定;

元数据质量	1. 91%以下属性元数据数据项齐全，内容正确； 2. 89%以下属性元数据数据项表达合理、清晰； 3. 89%以下元数据能支持各项查询、修改等操作；
数据转换无损	1. 通过转换而来的数据信息损失大于 10%； 2. 转换过程中空间数据与属性数据的连接关系丢失大于 8%； 3. 数字化获得的数据与理论数据存在的误差经没有校正；
属性数据定义	1. 属性数据定义不能表达出要表达的信息； 2. 属性数据的定义难以理解；
描述信息正确性	1. 颜色分配无序，不能表达分类信息； 2. 线型 6%以上不符合“格式与规范”的要求； 3. 有 6%以上的符号没有按照图式绘制； 5. 注记格式有 6%以上错误；
表格正确性	1. 各种表格字段含糊，难以理解； 2. 表格设计不符合第一、二、三范式，冗余较多； 3. 表格之间连接混乱；

评价项被评为 B 等与 D 等应满足的条件可以参考 A 等、C 等与 E 等的评价标准取中灵活掌握，这里就不再罗列。

针对 A、B、C、D、E 五等，给出了各等得分限定：

表 4.22 各等得分限定

Table 4.22 The score limits of all grades

等级	A 等	B 等	C 等	D 等	E 等
评价项	得分为总分	得分为总分	得分为总分	得分为总分	得分为总分的
得分	的 90%以上	的 81%-90%	的 66%-80%	的 50%-65%	50%以下

可以看出，采用此种评分方式，有利于对属性数据质量进行更全面的打分。

4.4.5 对属性数据进行分类评分

依据上述评分标准，对山东省 1:25 万地名数据库进行分类评分，各类属性数据中评价项得分情况如下：

表 4.23 居民地类中各评价项得分

Table 4.23 The score of each item in residential area class

评价项 类别	数据 格式	完备性	唯一性	连接列 关键字	属性输 出质量
得分	10	9	9	9	4
评价项 类别	元数据 质量	数据转 换无损	属性数据 定义	描述信息 正确性	表格 正确性
得分	16	3	7	7	9
合计: 83					

表 4.24 工业设施类中各评价项得分

Table 4.24 The score of each item in industrial establishment class

评价项 类别	数据 格式	完备性	唯一性	连接列 关键字	属性输 出质量
得分	12	9	14	11	6
评价项 类别	元数据 质量	数据转 换无损	属性数据 定义	描述信息 正确性	表格 正确性
得分	8	4	4	8	5
合计: 81					

表 4.25 农业设施类中各评价项得分

Table 4.25 The score of each item in agriculture establishment class

评价项 类别	数据 格式	完备性	唯一性	连接列 关键字	属性输 出质量
得分	8	9	11	11	7
评价项 类别	元数据 质量	数据转 换无损	属性数据 定义	描述信息 正确性	表格 正确性
得分	9	3	6	8	6
合计: 78					

表 4.26 行政区划类中各评价项得分

Table 4.26 The score of each item in administration section class

评价项	数据	完备性	唯一性	连接列	属性输
-----	----	-----	-----	-----	-----

类别	格式			关键字	出质量
得分	5	12	16	9	5
评价项	元数据	数据转	属性数据	描述信息	表格
类别	质量	换无损	定义	正确性	正确性
得分	12	5	8	8	6
合计：86					

表 4.27 道路类中各评价项得分

Table 4.27 The score of each item in road class

评价项	数据	完备性	唯一性	连接列	属性输
类别	格式			关键字	出质量
得分	14	12	19	13	6
评价项	元数据	数据转	属性数据	描述信息	表格
类别	质量	换无损	定义	正确性	正确性
得分	6	5	4	9	3
合计：91					

表 4.28 植被类中各评价项得分

Table 4.28 The score of each item in vegetation class

评价项	数据	完备性	唯一性	连接列	属性输
类别	格式			关键字	出质量
得分	8	10	2	9	16
评价项	元数据	数据转	属性数据	描述信息	表格
类别	质量	换无损	定义	正确性	正确性
得分	9	7	3	7	1
合计：72					

表 4.29 水系类中各评价项得分

Table 4.29 The score of each item in water system class

评价项	数据	完备性	唯一性	连接列	属性输
类别	格式			关键字	出质量
得分	17	19	2	10	12

评价项	元数据	数据转	属性数据	描述信息	表格
类别	质量	换无损	定义	正确性	正确性
得分	7	4	3	12	2
合计：88					

表 4.30 其他类中各评价项得分

Table 4.30 The score of each item in other class

评价项	数据	完备性	唯一性	连接列	属性输
类别	格式			关键字	出质量
得分	12	15	8	7	8
评价项	元数据	数据转	属性数据	描述信息	表格
类别	质量	换无损	定义	正确性	正确性
得分	7	4	4	9	3
合计：77					

4.4.6 确定各类属性数据的权

依据各类数据量在总数据量中占的比重与重要性、代表性来确定各类属性数据的权。

表 4.31 各类属性数据的权

Table 4.31 The power of various attribute datas

类别	居民地	工业设施	农业设施	行政区划	道路	植被	水系	其他
权	0.15	0.05	0.05	0.25	0.15	0.1	0.15	0.1

4.4.7 属性数据分类评价

根据各类属性数据的权，进行地名数据库属性数据质量的分类评定：

表 4.32 各类属性数据的权与得分

Table 4.32 The power and score of various attribute datas

类别	居民地	工业设施	农业设施	行政区划	道路	植被	水系	其他
权	0.15	0.05	0.05	0.25	0.15	0.1	0.15	0.1
得分	83	81	78	86	91	72	88	77

地名数据库属性数据质量最终得分：

$$S = \sum_{i=1}^m W_i S_i = 0.15 \times 83 + 0.05 \times 81 + 0.05 \times 78 + 0.25 \times 86 + 0.15 \times 91 + 0.1 \times 72 + 0.15 \times 88 + 0.1 \times 77 = 83.65$$

7=83.65

其中, $W_i (i=1,2,3,\dots,m)$ 为各类属性数据的权, S_i 为各类属性数据的得分。

分类评价方法比一般评价方法具有更好的抽样精度。若数据层间的质量差异不大,一般评价方法和分类评价的抽样精度差别不是很大。但是分类评价方法抽样量少于一般方法,采用分类评价方法另一个优点是可以获得各类属性数据的质量抽样结果。所以分类评价方法能以比一般评价方法少的抽样费用而得到更好的抽样精度,因此分类评价方法是切实可行的。

本章小结:

本章将前述理论在山东省 1:25 万地名数据库中进行了检验,提出一套属性数据质量控制措施与评分标准,应用前述 GIS 属性数据质量分类评价模型对地名数据库中属性数据质量进行了分类评价,并取得了良好的效果。

结论与展望

(一) 主要结论:

(1) 由于采用由权确定属性数据评价项所占分值的方法, GIS 属性数据质量综合评价模型可以较好地反映分类较少的属性数据质量。

(2) 对于大批量、属性数据总体各类质量的实际分布差别很大的情况, 由于采用由权确定各类属性数据抽样数量的方法, GIS 属性数据质量分类评价模型可以以最小的检验成本获得最高的评价效果。

(3) 采用定性概率表达方式可以大大提高属性数据的精度。

(4) 采用定性概率计算方式可以大幅提高操作后的属性数据精度。

(二) 笔者建议:

由于各种原因, GIS 的属性数据库中存在着大量不确定性。笔者建议, 根据性质的不同及数据结构的不同, 可以分别用不同的方式表示属性数据误差。对于具有定量性质的数据可用标准差来表示, 而对于定性的数据则可用概率来表示。然而, 在属性数据库中, 由于增加了描述属性不确定性的指标, 这会大大增加数据的存储量, 尤其是栅格数据。因此如何合理地组织这些数据, 还值得进一步研究。另外, 经空间操作分析后, 源误差也已经积累传播到操作目标中, 如何合理有效地估计这些误差的传播, 也是一个值得深入分析的方面。

致谢

本文是在导师徐泮林教授,王瑞波高工的悉心指导下完成的。在数年的学习期间,导师严谨求实的治学作风,勤奋敬业的献身精神,一直鼓舞着我努力学习、奋发图强。在论文的撰写阶段,导师倾注了大量的精力,花费了许多宝贵时间,给予了严肃、认真的检查和指导。值此论文完成之际,特向尊敬的导师表示真挚的感谢和由衷的敬意!

感谢以卢秀山院长为核心的各位学院领导、院办的全体老师所给予的生活及学习上的关心和帮助。

感谢测绘院领导的支持与帮助。

感谢成枢老师、崔先国老师、李云岭老师、张连蓬老师、季民老师给予的学术思想的启迪。

感谢陶华学老师、靳奉祥老师、郭金运老师、刘国林老师、韩晓冬老师、于胜文老师、张纯连老师、郑文华老师、赵会兴老师等测绘工程与地理信息系统专业各位老师的关心和支持。

感谢各位师兄、师姐、师弟、师妹的热情帮助。

参考文献

1. 贾冰媛、王学军. 统计地图的制作技术与实现[J], 计算机应用, 1991, 4: 16-20
2. 彭望碌. 遥感数据的计算机处理与地理信息系统[M]. 北京: 北京师范大学出版社, 1991.
3. 周心铁. 地理信息系统中曲线段数字记录重建多边形和方法[J], 环境遥感, 1988, 1(3): 55-62
4. 童小华, 刘大杰, 杨东援. 数字化数据的平差模型及软件实现[J], 同济大学学报, 1998, 26(6): 77-70.
5. 田铮. 动态数据处理的理论与方法[M], 西安: 西北工业大学出版社, 1995.
6. 周概容. 概率论与数理统计[M]. 北京: 高等教育出版社, 1988.
7. 黄幼才. GIS 空间数据误差分析和处理[M], 武汉: 中国地质大学出版社, 1995.
8. 史文中. 空间数据误差处理的理论与方法[M], 北京: 科学出版社, 1998.
9. 赵耀龙等. 模糊信息处理在测绘学中的应用探讨[J], 测绘通报, 1998, (8): 37-42
10. 黄幼才, 刘文宝: 数字化误差建模中粗差探测和抗差估计. 武汉测绘科技大学学报. 1995 Vol. 20 No. 2 pp151~156..
11. 刘文宝, 黄幼才: 地图曲线数字化误差估计, 武汉测绘科技大学学报: 1994, 19(4): 352~358
12. 刘智敏. 不确定度原理[M], 北京: 中国计量出版社, 1993.
13. 黄崇福, 王家鼎. 模糊信息优化处理技术及其应用[M], 北京: 北京航空航天大学出版社, 1995.
14. 边馥苓. 地理信息系统原理和方法[M], 北京: 测绘出版社, 1996.
15. 叶德培. 测量不确定度[M], 北京: 国防工业出版社, 1996.
16. 赵耀龙, 赵俊三, 侯至群. 略论云南省地理信息产业发展战略[D], 全国现代管理科学学术研讨会论文集, 昆明: 云南科技出版社, 2000.
17. 史文中, 王树良. GIS 中属性不确定性综述[D], 中国地理信息系统协会 2001 年年会论文集, 成都, 2001.
18. 刘大杰, 刘春. GIS 空间数据不确定性与质量控制的研究现状[J], 测绘工程, 2001, 10(1): 54-59.
19. 徐绍铨, 张海, 等. GPS 测量原理及应用[M], 武汉: 武汉测绘科技大学出版社, 1998.
20. 宫同森. 地形、地籍测量精度[M]. 北京: 测绘出版社, 1992.

21. 华慧等. 数字化地图位置精度的统计分析[J]. 测绘通报, 1998(3): 78-82.
22. P.J.G. Teunissen. 非线性最小二乘法/欧吉坤译. 测绘译丛[M], 1991.
23. 刘国林等. 非线性平差模型强度的加权曲率度量[J]. 山东矿业学院学报, 1998(3): 45-49.
24. 周江文, 黄幼才, 杨元喜等. 抗差最小二乘法[M]. 武汉: 华中理工大学出版社, 1997.
25. 杨元喜. 抗差估计理论及其应用[M]. 北京: 八一出版社, 1993.
26. 周江文. 经典误差理论与抗差估计[J]. 测绘学报, 1989(2): 12-17.
27. 欧吉坤. 一种检测粗差的新方法—拟准检定法[J]. 科学通报, 1999(16): 78-82.
28. 欧吉坤. 论真误差拟准解的基本特性[J]. 测绘通报, 1999(2): 34-38.
29. 埃替斯, 霍罗维茨等. 用 C++ 描述数据结构[M]. 北京: 国防工业出版社, 1997.
30. 龚洁辉, 白玲, 高建美. 最短路径算法的改进与实现方法[J]. 解放军测绘学院学报, 1998, (2): 56-59.
31. 王家耀, 武芳, 吴升. 发展我国数字地图生产若干问题的思考[J]. 数字制图技术与数字地图生产学术讨论会论文集[D]. 长沙: 湖南地图出版社, 1999.
32. 杨春成. 基于分布微机数据库的电子地图制作系统[J]. 测绘科技, 1997(2): 56-59.
33. 刘光运, 韩丽斌. 电子地图技术与应用[M]. 北京: 测绘出版社, 1997.
34. 杨春成. 拓扑自动生成系统的设计与实现[J]. 解放军测绘研究所学报, 1999(1): 41-44.
35. 黄杏元, 汤勤. 地理信息系统概论[M]. 北京: 高等教育出版社, 1989.
36. 齐华, 刘文熙. 建立结点-结点拓扑关系的 Qi 算法[J]. 测绘学报, 1996(3): 21-24.
37. 孟晓林, 刘大杰. 地图数字化数据误差的 NL 分布检验, 同济大学学报, 1996, 10: 60-66.
38. 李德仁, 彭美云, 张菊清: 估计 GIS 中面要素点位误差的精度指标, 武汉测绘科技大学学报 1996 Vol, 21, No, 2, pp. 134~138,.
39. Prunt B F. Hidden-line Removal from Three Dimensional Maps and Diagrams[J], Display and Analysis of the Spatial Data. London: Wiley, 1973, 118-209.
40. Cole R, Sharir M. Visibility Problems for Polyhedral Terrains[J]. Technical Report, Courant Institute, New York University, 1986.
41. Goodchild M F, Lee J. Coverage Problems and Visibility Regions on Topography [J], Surfaces. Annals of Operations Research, 1989(20): 175-186.
42. Berry J. PMAP. The Professional Map Analysis Package[J], Papers in Spatial Information Systems., 1986(3): 22-26.
43. Marzano P, Puppo E. Line-of-sight Communication on Terrain Models[J]. IJGIS, 1994,

8(4):329-342.

44.Anderson T, Zube E, Connell W. Studies in Landscape Perception[J], Institute for Man and Environment, Amherst,1976. 6-49.

45.A. J. Pope. Two approaches to nonlinear least-squares adjustment[J], The Canadian Surveyor,1974(5).

46.J.L Nugent ,Quality Control Techniques for a GIS [J].Database Development Project PE &RS, Vol,61,No,5,pp.523-527,1995.

47.Shi Wenhong :Modeling and Thematic Uncertainties in Integration of Remote Sensing and GIS[M]. Holland, ITCpublication.1994.

48. Pavel Hruby: Specification of Workflow Management Systems with UML:

<http://www.umlchina.com/>

49. Sjaal Brinkkemper, Shuguang Hong: Object-Oriented Analysis and design methods comparativeReview:<http://www.is.cs.utwente.nl:8080/dmrg/OODOC/oodoc/oo.html>.

50.Martin Fowler:Techniques for Object Oriented Analysis and Design:

<http://www.martinfowler.com/>.