海洋渔业数据库质量控制研究

杜云艳 周成虎 邵全琴 苏奋振
(中国科学院地理科学与资源研究所资源与环境国家重点实验室, 北京 100101)

摘 要 为了利用可视化技术来对海洋渔业数据库进行质量控制,在综述了近年来 GIS 数据库质量控制的基础上,首先提出了 GIS 数据库质量控制的多层次概念模型;然后针对海洋渔业地理信息系统的特点,给出了海洋渔业数据库质量控制的体系结构和具体的控制方法,并在现有的元数据和数据仓库技术的基础上,提出了基于元数据的智能化控制方法原型;最后结合前人工作和国家 863 计划 818 主题的海洋渔业 GIS 课题研究成果,给出该质量控制的概念模型应用于海洋渔业具体的控制体系和方法.

关键词 质量控制 多层次概念模型 智能化数据质量控制
中图法分类号: TP311.13 P208 S951.2 **文献标识码**: A **文章编号**: 1006-8961(2002)03-0276-06

Data Quality Research on Ocean Fishery Database

DU Yun-yan, ZHOU Cheng-hu, SHAO Quan-qin, SU Fen-zhen
(*LREIS, Institute of Geographic Sciences and Natural Resources, Beijing* 100101)

Abstract Data quality has been the vital factor on which the reliability of the result of GIS analysis and application depends with the development of GIS. Based on the review of GIS database quality control recently research, a multi-hierarchy conceptual model of data quality control is given in this paper. We can accomplish data quality control beginning with data source and then database proceeding, finally data using by this model, and then aiming at the characteristics of Marine Fishery GIS data, the detail structure of database quality control is schemed, at the same time several practical approaches of data quality control are discussed thoroughly. At last, based on the techniques of metadata and data warehouse, a prototype of intelligent data quality control is provided.

Keywords Data quality control, Multi-stage conceptual model, Intellectualized data quality control

0 引 言

从数据作为产品的角度来看,数据质量评价是必须具备的. GIS 作为一个面向特定应用目标的用于采集、管理、分析、显示地理信息的信息系统,在初期发展中,其数据质量控制相对于系统的数据库建设来讲,显得薄弱,也不成体系. 由于随着应用的深入,系统的建设者和用户越来越感觉到数据质量的好坏已成为影响系统应用分析结果的可靠性以及系统真正实用性的关键因素,因此全面开展了 GIS 的数据质量控制理论和方法研究. 早在 1995 年 7 月,在葡萄牙的里斯本就专门举行了 GIS 数据质量的专家会议,与会代表来自计算机和 GIS 领域,在该会议上,就当前数据质量急需解决的问题提出了数据质量的概念、数据质量的机理、数据质量控制和应用 4 个大的研究方向[1]. 之后,对这 4 大方面,国内外均开展了很多深入的研究,尤其在 GIS 空间数据质量传输机理和数据质量控制及其应用方面. 其中,前者主要是从模型的角度研究空间数据的不确定性和误差传递,该方向已经延伸到 GIS 空间分析、决策和应用的不确定性研究;后者主要从软件角度探讨常规空间数据质量控制功能模块和特定应用系统质量控制体系的应用,如现有的 Arc/Info 软件,其强大的图形编辑和图形属性完善的连接方式等功能都用于质量的控制. 近年来,随着城市地理信息系统研究与开发的成熟,阎正给出了城市地理信息系统的数据质量概念及其影响因素和误差传递规律[2].

基金项目:国家高科技 863 计划海洋领域监测主题(863-818-11-03)
收稿日期:2000-10-24;改回日期:2001-05-31

此外,本文另一个目的是为了利用当前的信息可视化技术来实现数据质量的控制,近 5 年来,信息可视化问题日益受到人们关注. 信息可视化主要是指非空间数据的可视化,虽然它的出现是应大型数据,甚至是海量数据的存储、传输、检索的要求而迅猛发展起来的[3],但可视化技术也是为了了解数据之间的相互关系和适应数据管理的发展趋势,为此本文主要试图通过这种可视化技术来进行海量数据的质量控制.

1　GIS 数据库质量控制的多层次概念模型

众所周知,位置、时间和专题特征是表达现实世界空间变化的 3 个基本要素,而空间数据是记录这些变化信息的载体,数据质量则是描述和度量空间数据在表达这 3 个要素时,所能达到的准确性、一致性、完整性、合理性. 据研究,影响空间数据质量的因素很多,作用也不尽相同,而对 GIS 中影响数据质量的主要原因也有过较多的研究,如今形成了以下两种观点:其一是以何建邦为首的从事城市地理信息系统研究与开发的 GIS 科技人员给出的,这种观点是从 GIS 系统建设的角度出发,认为地理信息系统的数据质量问题,实际上是伴随着数据的采集、处理和应用过程而产生,并表现出来的. 根据这一认识,可以把地理信息系统的数据质量问题形成过程划分为数据的采集与保存、数据的录入与转换、数据的分析和处理 3 个阶段[4];其二是以李军为代表的观点,即从空间认知的角度给出产生数据质量的来源,他们认为数据质量问题是受人类认识世界的方式和手段限制而产生的,其形成的主要的误差有:空间实体和地学过程的自身不稳定性、空间实体和地学过程表达方式误差、空间数据处理中的误差、数据使用误差等[5]. 鉴于数据质量影响因素都是多阶段、多方面的,因而对数据质量的控制也相应地应该有阶段和分层次. 为此,本文在综合分析数据质量影响因素的基础上,提出了一种多层次的 GIS 数据库质量控制概念模型(见图 1).

从某种意义上讲,该模型与数据质量问题成因的第 1 种观点是一致的,这种模型就是从数据的采集到录入和处理分析逐步深入地进行质量控制. 从图 1 可以看出,对数据质量的控制分为如下 3 个层次:开始是从数据源的角度加以控制;随后是从 GIS 系统建设的流程给出基于数据库的质量控制,其可

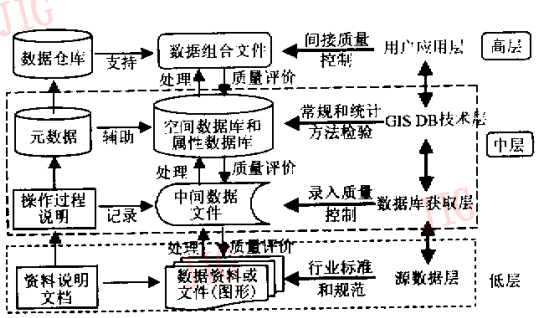


图 1　GIS 数据库质量控制的多层次概念模型

以看作是技术层的质量控制;最后是面向具体应用的质量控制,也是相对高层次的质量控制. 其中,源数据层的质量控制是初级和低层次的,但由于数据源本身反映了对该地理现象或者地理过程的认识程度,因此是最终质量评价的前提条件,这里对数据源的质量控制可以按照数据采集的途径和基础空间数据标准、规范以及行业标准和规范来进行,相应的质量控制报告是资料说明文档;随后是伴随 GIS 系统的技术流程所带来的质量控制,由于 GIS 数据包括图形数据和属性数据两部分,因此在数据获取阶段的质量控制也包括了这两部分,数据进库后又有常规 DB 的各种约束控制以及属性与空间的关联控制,在此称之为直接质量控制方法,与该阶段相对应质量控制报告有过程说明文档和元数据库,众所周知,从 20 世纪 90 年代初,便提出了元数据概念,在此可以看作是质量控制的辅助技术支持,因为元数据是记录数据的数据,随着元数据理论的成熟和技术的完善,直接通过元数据库就可以了解和分析数据的状况和精度情况,甚至误差处理的过程;最后,结合到具体的应用,通过间接的方法来寻找隐藏在数据中的错误. 近年来,因研究了很多的数据仓库技术,从而为该阶段的质量控制提供了技术支持. 上述 3 个阶段是动态、反复的过程,通过数据仓库和元数据库的相应技术可以作到自动响应数据改正.

2　海洋渔业数据库质量控制体系

2.1　海洋渔业数据库质量控制特点

由于相对于陆地系统,海洋的空间背景数据比较单一,变化也较少,而与空间数据相连的属性数据则变化较大,并且数量也较大,因此对海洋渔业数据库的质量控制不同于常规的陆地地理信息系统,其主要有以下特点:

(1) 以渔区为单位的匹配检验

在海洋渔业地理信息系统中,空间和属性关联的关键字段是渔区代码,由于属性数据不仅在空间上是以不同的渔区级别为统计单元而形成多重格网的统计数据,而且在时间上也是以不同的时间段来统计,同样产生了多重时间段的统计数据,因此,对渔业数据库质量的控制首先是进行以多种渔区为基准的匹配校验.

(2) 以属性数据的质量控制为主

正如前面所讲的,由于海洋渔业综合数据库的主要份量集中在隐含有空间特性的属性数据(包括统计数据和环境数据)中,因此以属性数据为主的综合数据库质量控制也是海洋渔业 GIS 质量控制的特点之一.

(3) 数据量大,自动控制

由于海洋渔业综合数据库中的属性数据是一长时间序列数据,且属于同种数据格式的数据量较大,因此采用自动质量控制方法,是面向渔业 GIS 的需要而应有的质量控制特点

2.2 海洋渔业数据库质量控制体系

针对海洋渔业数据库质量控制的特点,在深入理解质量控制多层次概念模型的基础上,给出了进行海洋渔业综合数据库质量控制的体系结构(见图 2). 由图 2 可见,该体系具有 3 层结构,其基础层是数据层,包括原始数据的整理和编辑,以及 GIS 最初阶段数据(具体的是海洋渔业的背景数据和各种属性数据)的采集和录入;其二是质量控制的方法层,包括如下直接、间接和智能 3 大类方法:①直接控制法是在数据录入时,面向数据库本身的初级控制,其中,采用 Arc/Info 的错误检查方法是用于控制空间数据,而双录入法和统计检验法则用于控制属性数据,②间接方法是在初级控制的基础上面向应用的高级控制,主要是针对属性数据,但在渔区,则是隐含的空间特征,③智能方法是面向整个系统更高层的质量控制方法,同样以渔区为线索,自动地根据数据层进行控制和修正;其三是经过方法层逐步深入地质量控制后,数据提交给用户层,从另一个角度看,也可以认为用户层是质量控制最初始的驱动层,因而可以选择质量控制的层次和深度,以及相应的方法.

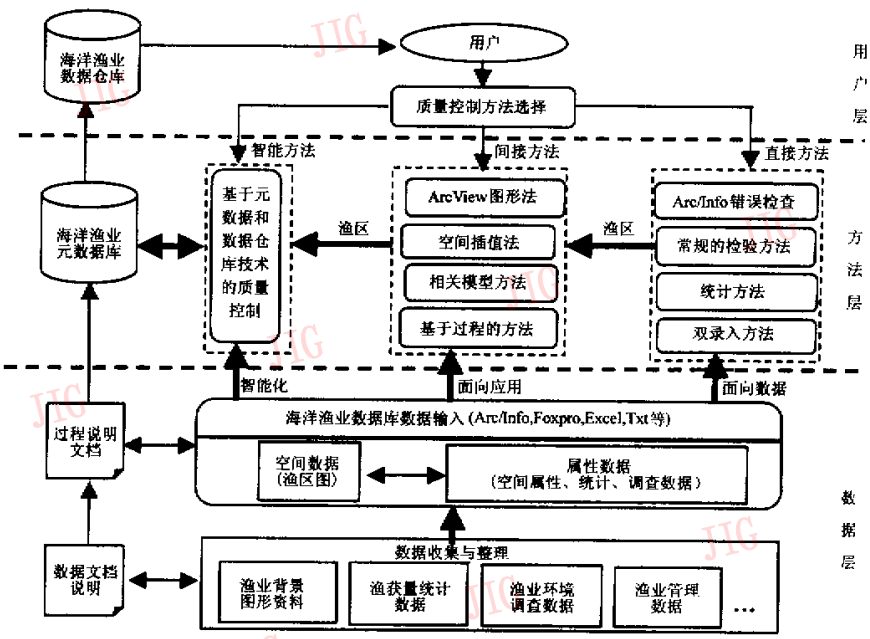


图 2 海洋渔业数据库质量控制体系结构

3　质量控制方法研究

按照海洋渔业数据库质量控制的体系结构,在具体进行数据库建设时,应分步骤地采用各种方法来全面地进行质量控制. 同时,海洋渔业数据的特殊性,也决定了数据控制方法的特殊性. 人们通过具体的实验,总结了一套比较实用的质量控制方法.

3.1　直接的质量控制方法

所谓直接的控制方法,在此也称作面向数据本身的质量控制方法,它是最简便和快速的方法. 从 GIS 的流程来看,它是处于数据的采集和入库阶段的质量控制方法. 针对海洋渔业数据多为隐含有空间特性的属性数据这一特点,直接控制方法主要是面向该类属性数据,该类方法有以下几种:

(1) 双录入方法

双录入法是应处理大批量的渔业捕捞统计数据的需要而采用的双人录入方法,也是从录入的方式上,控制数据质量的一种方法. 具体实施中,可以采用多种输入软件,只需要针对输入的软件,通过编制相应的错误检查小程序来自动寻找双录入中不一样的记录,即可进行核实和纠正. 从概率的角度看,采用该方法出现错误的概率从一开始就大为降低,而查找速度反而提高. 可见不失为一种好的方法.

(2) 满足标准和规范的质量控制

满足标准和规范的质量控制是经过双录入检查的数据入库后,需进行的首要质量控制. 一般来讲,不同行业有不同的标准和规范,渔业也不例外,反映到数据上,是类型和取值都有特定的规定. 比如,渔业的作业类型和渔船的编号都有一定的标准和规范.

(3) 阈值检验方法

所谓阈值检验法,可以看作是异常值检验方法. 大家知道,海洋异常数据的处理方法是海洋数据资料质量控制的重要研究课题[7],因为在海洋资料中(尤其是环境观测数据),会发现少数比正常数值大得多或小得多的异常数据,即通常所说的异常值. 按照性质其可以分为正确的异常值和含有过失误差的异常值两类. 阈值检验方法就是检验这种过失异常值的方法,它是通过海洋要素的正常取值范围来判断异常值. 该方法是继初级控制后,用于稍高层次质量控制的方法.

(4) 统计方法

由于采用阈值检验方法只能判别比较明显的异常值,而一些虽然满足取值范围,可是从数据序列的角度来看,虽属于异常的值,但并不能用阈值检验方法查出,因此一些基于概率统计的误差检验方法应运而生,如箱线图就是其中比较常用的方法. 该方法是利用最大值、最小值、上分位、下分位数和中位数这 5 个特征值来表征一组数列的统计量,即可以用箱线突出数列中部 50%的数据,而上界线一般采用上分位数加上四分位数间距的 1.5 倍,下界线则采用下分位数减四分位数间距的 1.5 倍来确定. 该方法认为落在上下界线之外的值为异常值,且认为正态分布的数列有 95%的数据落在上下界线范围内,因而它能详细地表示 95%之外的异常值[6].

3.2　间接的质量控制方法

经过直接的质量控制,单从数据的角度来看,数据库中的数据已经满足了规范,并且已经粗略地排除了异常值,但从数据与数据之间的相互关系来看,依然会存在误差. 这种情况就必须结合一定的应用或者根据一定的模型来进行深层次的错误检查. 换一个角度来看,也可以认为间接质量控制就是在数据的处理和分析时,用以发现数据的错误,并加以改正,其主要方法如下:

(1) 空间展布法

所谓空间展布法,这里是指把统计或观测值展现在空间上,并利用空间位置来进行纠错的方法. 在本系统中,海洋渔业环境数据库的数据主要是以点位信息作为隐含空间信息的观测数据,因此对其空间点位的错误检查可以采用空间展布法. 在本系统中,是利用 Arc/Info 和 ArcView 来把断面调查数据展现在海洋渔业背景图上,如果点位落在陆地区域,或者偏离某一条航线很远,或者相距很近的两个观测点的某个环境要素差距很大,则说明数据记录有问题. 该方法既可用于直观的点位查错,又可以进行隐含的观测要素查错. 图 3 是利用该方法进行的断面点位查错的例子.

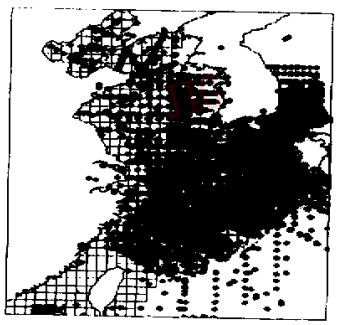


图 3　利用 ArcView 进行断面数据的点位查错

(2)空间插值法

该方法是更进一层的观测数据查错方法,由于单点数据即使利用上述方法展现在平面上,对于单要素的数值检验来说,依然不很直观,因此试想把表达连续现象的点位观测数据内插成面状数据,再采用三维图形方法表达,并进行隐含错误的检查就可能会直观些. 本系统中,对断面数据的温度和盐度要素采用克吕格插值法进行同一时间段内观测值的平面内插,并用相应的软件进行立体显示. 图 4 是利用该方法进行的断面观测数据查错的例子,如果从图中可明显看到起伏剧烈的区域,则该区域中必定包含异常的观测数据.

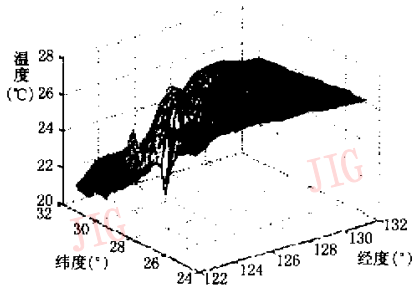


图 4 利用空间插值法进行的观测值查错

(3) 基于过程的质量控制方法

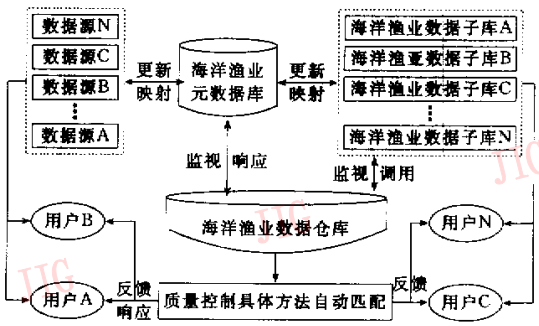
空间展布法和空间插值法对数据的质量控制是一种静态的质量控制方法,对于长时间序列的统计数据或观测数据,该方法的重复性较大,并且受到时间取段长短的限制,由于数据的异常与否与所取的时间长短有关,如,在短时间内如果两个相邻观测点的温度差值大于 2 C为异常,长时间内则不一定就是异常,因此针对本系统中长时间序列数据多的特点,采用了一种动态的质量控制方法,即基于时间过程的方法. 具体是在上述两种方法的基础上,采用时间纵向和横向分段的方法来对数据进行批处理和空间图形显示,再依靠数据随时间相对变化程度来探寻隐含在数据中的错误. 具体实施则采用 ArcView 和 Matlib 等商业软件的二次开发功能来完成数据时间维的动态表达,从而发现错误,具体程序略.

(4) 相关模型方法

从空间图形法到基于过程的方法虽然实现了质量控制从静态到动态的进步,但它们都只是对单要素的控制方法. 由于无论是陆地还是海洋系统,各个要素之间都是有关系的,因此可以利用关系密切的其他要素来辅助进行当前该要素的控制. 由于该方法需要建立多个要素之间的相关模型,因此需要一定的背景数据,这对于那些具有长期观测资料的用户非常有效. 针对海洋渔业领域,渔业的渔获量同海洋的环境要素很有关系,如何采用渔获量与海洋环境相关模型的方法来进行质量控制,是随着对海洋渔场形成机制的深入理解而逐步深入的. 目前在物理海洋领域采用这种方法的比较多,主要是利用温度、盐度、密度资料进行相互校验.

3.3 基于元数据的智能型质量控制方法

如前所述,在质量控制的多层次概念模型中已提到了基于元数据的智能化质量控制方法,该方法起因于美国海洋观测系统计划中提到的在数据管理中对数据质量的多层控制,由于不同的人对数据的使用不同,因此对数据质量提出的要求也不一样. 为实现多层控制,则必须考虑到多尺度的问题,最好的办法是结合元数据给出不同程度的控制[7]. 所谓元数据是关于数据的数据,其中,从数据源的说明到 GIS 各项操作都有记录,由于现有对空间元数据库的研究无论是在技术上,还是理论上,都有了很大的进步,如今结合数据仓库技术,相应元数据已可以随着原始数据的更改而进行自动更新,因此,对数据质量控制来讲,若采用元数据库中对应的数据精度说明,就可以实现智能化地选择具体质量控制方法,从而满足特定的用户使用,该方法即为基于元数据的智能化控制方法,具体的概念模型见图 5. 从图 5 可以看出,不同的用户使用不同精度的数据,而具有不同精度的数据源会自动地提供不同精度的应用. 但由于本方法智能化程度还受数据仓库和元数据技术的影响,因此在此仅给出方法模型.



4 结 论

本文在总结分析现有数据库质量控制理论的基

础上,首先给出了质量控制的多层次概念模型;然后针对海洋渔业数据质量控制的特点,给出了海洋渔业数据库质量控制的理论体系和具体的切实有效的实现方法;最后,结合现有的元数据和数据仓库技术,从概念上提出了一种智能化的质量控制方法,但该方法尚有待于具体的应用和实践.

参 考 文 献

1 Specialist meeting on data quality[EB/OL]. http://www.shef.ac.uk,1995

2 国家空间数据基础设施[EB/OL]. http://www.nsdi.gov.cn/comment.html, 2000.

3 唐泽圣. 可视化及虚拟现实技术的新发展[EB/OL]. http://www.e-works.net.cn/JCJS/id29.htm,1999.

4 阎正. 城市地理信息系统标准化指南[M]. 北京:科学出版社,1998:158～162.

5 李军. 地球空间数据集成基础研究及其应用[D]. 北京:中国科学院地理研究所,1998.

6 陈上及编著. 海洋数据处理分析方法及其应用[M]. 北京:海洋出版社,1991.

7 Susan Martin R, Toward a U. S. Plan for an Integrated, Sustained Ocean Observing System [EB/OL]. http://www-ocean.tamu.edu/goos/publications/sw/section3.html,1999

杜云艳　1973 年生,2001 年获中国科学院地理科学与资源研究所博士学位. 一直从事遥感和地理信息系统,尤其是海洋渔业地理信息系统的研究,已发表论文数篇.

周成虎　研究员,博士生导师,现任中国科学院资源与环境信息系统国家重点实验室主任. 主要研究领域为洪水灾害的遥感监测、基于知识的遥感影象理解、基于元胞自动机的地理时空模拟等. 出版专著 10 册,发表论文 60 余篇.

邵全琴　1962 年生,研究员,1987 年获南京大学硕士学位. 主要研究领域为海岸带滩涂资源与环境调查、洪水灾害遥感监测、城市地理信息系统、海洋渔业地理信息系统等,已发表论文 20 余篇,编写文集 5 本.

苏奋振　1972 年生,2001 年获中国科学院地理科学与资源研究所博士学位. 一直从事海洋遥感、海洋地理信息系统,数据挖掘等方面的研究.