

Data mass standard of digital human body

BI Si-wen, JING Dong-sheng

(State Key Laboratory of Remote Sensing Science, Application Institute of Remote Sensing,
Chinese Academy of Sciences, Beijing 100101, China)

[Abstract] The meanings of data mass of digital human body was given in this paper. The data mass standard of digital human body, the data mass contents of digital human body, the data gather mass analysis and data error transfer of digital human body were introduced, including concept of data mass, the influence elements to data mass, digitalization error of data mass and data mass model of digital human body and so on. Furthermore the evaluation to data mass of digital human body was introduced. This paper provided a theoretical gist for studying digital human body standardization system.

[Key words] Digital human body; Data mass; Standard; Error; Evaluation

数字人体数据质量标准

毕思文, 景东升

(中国科学院遥感应用研究所遥感科学国家重点实验室, 北京 100101)

[摘要] 本文研究了数字人体数据质量含义、数字人体数据质量标准、数字人体数据质量内容、数字人体数据集质量分析和数字人体数据误差传递, 包括数据质量概念、影响数字人体数据质量的因素、数据质量的数字化误差和数据质量模型等, 并进行数字人体数据质量评价, 为研究数字人体标准体系提出理论依据。

[关键词] 数字人体; 数据质量; 标准; 误差; 评价

[中图分类号] TP391.9 **[文献标识码]** A **[文章编号]** 1003-3289(2004)05-0781-05

1 数字人体的数据质量含义

1.1 数字人体数据质量的概念 数字人体信息是对现实物质人体的抽象和表达。由于现实世界的复杂性和模糊性, 以及人类认识和表达能力的局限性, 这种抽象和表达总是不能完全达到真实值, 而只能在一定程度上接近真实值。而且, 真实值往往是不可知的或不可测的, 因此误差总是存在着。

一般来讲, 质量(mass)指产品满足指标、状况和要求能力的特征总和。数字人体数据质量的好坏不是一个绝对的概念, 目前比较公认的定义是指数据对特定用途的分析和操作的适用程度。

1.2 与数字人体数据质量有关的基本概念 在论及数据质量的好坏时, 人们常常使用误差或不确定性的概念, 数据质量问题在很大程度上可以看作数据误差问题。而描述误差最常用的概念是准确度和精密度。主要包括: 误差, 数据的准确度, 数据的精密度和不确定性。

1.3 数字人体数据质量的测度 有许多数位可用来测度数

据质量, 最重要的有: 精度、精确性、可靠性。这些测度数位往往具有不同的含义; 精度是指结果和真实值相似程度, 通常是用标准差或其平方根给出。精确性是指在量测方法中小数位数字, 或称为有效数字。精确性和精度是不一样的, 大的有效位数并不表示测量位是精确的。可靠性给出有关数据的可控制性和错误(粗差)信息。用可靠性表示最小可查明的错误的数位。在正态分布量测值的情况下, 可靠性的数值可被确定。

2 数字人体的数据质量内容^[1,3]

2.1 数字人体数据质量描述 数字人体数据质量由数据质量元素来描述。数据质量元素分为两类: 定量元素和非定量元素。定量元素描述数据集质量的定量成份, 非定量元素描述数据集质量的非定量成份。

2.2 数字人体数据质量元素和子元素

(1) 数字人体数据质量定量元素: 数字人体数据质量定量元素包括六个方面内容: 完整性、逻辑一致性、位置精度、时间精度、专题(属性)精度、用户定义(数据质量元素或数据质量子元素), 主要内容有: 数据情况说明, 指对数据说明的全面性和准确性; 位置精度, 或称定位精度; 属性精度, 指实体的属性值与其真值相符的程度; 逻辑一致性, 指数据关系上的可靠性; 数据完整性, 指地理数据在范围、内容及结构等方面覆盖所有要求方面的完整程度; 时间精度, 主要指数据的现势性。

(2) 数字人体数据质量定量子元素: 数字人体数据质量子

[作者简介] 毕思文(1956—), 男, 北京大学和清华大学双博士后, 研究员, 北京医药信息学会“数字人体-人体系统数字学”专业委员会主任委员。研究方向: 数字人体-人体系统数字学。

E-mail: bisw@irsa. irsa. ac. cn

[收稿日期] 2004-04-07

元素指数据质量元素的组成部分,用于描述数据质量元素的某一方面。数据质量量子元素含义见表 1。

表 1 数字人体数据质量元素及其含义		
数据质量元素	数据质量量子元素	含义
完整性	多余	数据集中有多余的数据
	缺少	数据集中缺少应有的数据
逻辑一致性	概念一致性	符合统一概念模式规则
	值域一致性	同在界定的值域范围内
	格式一致性	数据存储与数据集物理结构、规定格式的一致性程度
	拓扑一致性	数据集逻辑特征和拓扑关系的正确性
位置精度	绝对精度	数据集坐标值与可接受的值或真值之间的接近程度
	相对精度	数据集中要素相关位置与各自对应的、可接受的相关位置或真值之间的接近程度
	格网数据位置精度	格网数据起始单元位置的值与可接受的值或真值之间的接近程度,分辨率大小
时间精度	时间的量测精度	数据集使用时间参照系统的正确性
	时间的一致性	时间序列的一致性
	时间的有效性	数据在时间上的有效性
专题(属性)精度	分类正确性	要素或属性相对于分类标准的一致性程度
	非定量属性的正确性	非定量属性描述的正确性
	定量属性的正确性	定量属性的精度

(3)数字人体数据质量非定量元素:数字人体数据质量的非定量元素用于描述数据集非定量的质量内容:目的、用途、数据志。目的描述生产数据集的原因和主要目的。用途描述数据集对于数据生产者 and 数据用户等的应用范围。数据志描述数据集的历史沿革,即数据集从获取、编辑到现状完整生命周期的有关描述。数据志包括两个独立的部分:数据源信息和数据处理步骤、重要处理事件(转换、维护)信息。

2.3 影响数字人体数据质量的因素 数字人体的数据质量问题,实际上是伴随着数据的采集、处理与应用过程而产生并表现出来的。根据这一过程,可以把数字人体的数据质量问题划分为三个阶段:

第一个阶段是数字人体数据的采集和保存;第二个阶段是数字人体数据库的建立,包括数字化、数据录入和必要的数

据转换;第三个阶段则是数字人体中数据的操作、分析和处理。每一个阶段都包含前一个阶段所带来的原有误差,并增加了本阶段所引入的新的误差因素。因而,数据质量的影响因素可用数据获取和应用过程的这三个阶段为线索来考查。每一个生产或应用过程,都通过影响数据质量六个方面的全部或一部分而影响到数据质量,主要包括:数字人体数据源影响数据质量的因素;数字人体数据库建立中对数据质量产生影响的因素;数字人体分析和处理过程产生的数据质量问题。

3 数字人体的数据集质量分析^[2]

3.1 数字人体数据质量本身的模糊性 数字人体数据具有多重性、复杂性、不精确性及不确定性,也即数据的模糊性,它给地理数据数学模型的拟构和分析带来许多困难,主要表现为:分布位置的模糊性,地理现象描述产生的模糊性,地质数据的特点及其类型。

3.2 数字人体的图像数据质量

(1)测量数据的质量问题:测量数据主要指使用人体系统高新技术和其他一些测量方法直接量测所得到的测量对象的空间位置信息。这部分数据的质量问题,主要是空间数据的位置误差。测量方面的误差通常考虑的是系统误差、操作误差和偶然误差。

①系统误差的产生与一个确定的系统有关,由于环境因素(如温度、湿度和气压等)、仪器结构与性能以及操作人员技能等方面的因素影响而产生。系统误差不能通过重复观测加以检查或消除,只能用数学模型模拟和估计。

②操作误差是操作人员在使用设备、读数或记录观测值时因粗心或操作不当而产生的。应采用各种方法检查和消除操作误差。操作误差一般可通过简单的几何关系或代数检查验证其一致性,或通过重复观测检查并消除操作误差。

③偶然误差是一种随机性的误差,由一些不可见和不可控制的因素引入。这种误差具有一定的特征,如正负误差出现频率相同、大误差少、小误差多等。偶然误差可采用随机模型进行估计和处理。

(2)图像数据的质量问题:图像数据是指由现有图像数字化产生的数据。图像数据质量问题中,不仅含有图像固有的误差,还包括图纸变形、图形数字化等误差。

(3)应用 RS 采集数字人体数据:遥感(remote sensing, RS)数据的质量问题,一部分来自遥感观测过程,一部分来自遥感图像处理和解译过程。遥感观测过程中,本身存在精确度和准确度的限制。这一过程产生的误差主要表现为分辨率、几何畸变和辐射误差,这些误差将影响遥感数据的位置和属性精度。

遥感图像处理和解释过程中,主要产生空间位置和属性方面的误差。这是由图像处理中的影像或图像校正和匹配以及遥感解译判读和分类引入的,其中包括混合像元的解译判读所带来的属性误差。遥感技术可以为数字人体的数据更新提供可靠、迅速的数据源。但在以遥感信息作数字人体数据源时有几方面问题需要解决:数据结构问题,多源遥感影像的融合问题,遥感信息与非遥感信息的复合。

(4)现有图像资料的数字化:现有图像数字化的精度受数

字化仪本身的精度、数字化方式、操作者经验及心理状态等多种因素影响。多数数字化仪的标称分辨率为 0.001 或 0.025 mm,一般能满足数字人体的要求。

3.3 数字人体的专题数据质量和时间数据质量

属性数据是数字人体的重要数据源,一般由调查统计方法得到,其中存在的数据质量问题主要包括调查随机误差和统计误差,这种误差通常为属性误差。

常规测量数据的采集中,通过各种电子数据记录器,数据终端能以数字形式自动记录野外测量数据,并转到计算机中实现数字人体的数据采集。

时间是信息的重要属性。随着数据库与信息技术的深入和发展,时态信息在信息系统扮演着日益重要的角色。不确定性存在于所有系统中。在时间数据库中,时间的不确定性是指“不知道什么时候”。这种不确定性来源于四个方面:

①微粒过小:在大多数情况下,数据库时间计时单位与计算事件发生的时间尺寸不吻合。例如,数据库在的计时单位为秒,而实际事件是以天来计算的。

②计时的不精确性:即使数据库的计时单位和实际事件发生的时间相一致,但大多数计时设备是不精确的。

③预测的不精确性:绝大多数系统的预测时间是不精确的。

④事件时间的不确定性:有时实际事件发生的时间是不确定的,如一个人的出生日期等。

时间的不确定性导致了时间数据库模型的复杂性,在目前的应用系统中,不确定时间模型仍很少见。

4 数字人体的数据误差传递^[2]

4.1 数字人体数据质量的数字化误差 对于数字人体,其数据库的建立可能通过两种途径:由现有的纸介质的图形、图像和文档数据进行数字化和数据录入而建仓;或由现有的图形、图像和文档数据直接或通过一定转换而建立数据库。因此,这部分数据质量问题除了包括数字人体数据源的固有误差外,还包含数字化和数据转换所引入的误差。在这个过程中产生的数据质量问题,是数字人体数据质量问题的最重要的部分,也是数字人体数据质量控制的重点。数字化方式主要有手工数字化和扫描数字化两种:

(1)数据预处理:在进行图像数字化时,必须采用适当的比例尺因子进行修正。如果从不同的图像上采集信息,应了解图像的投影方式是否一致,比例尺是否匹配,以估计由此可能产生的误差。

(2)跟踪数字化:跟踪数字化一般有点方式和流方式两种工作方式,在实际生产中使用较多的是点方式。用流方式进行数字化所产生的误差要比点方式大得多。影响跟踪数字化数据质量的因素主要是:数字化要素对象,数字化操作人员,数字化仪和数字化操作。

(3)扫描数字化:扫描数字化采用高精度扫描仪将图形、图像等扫描并形成栅格数据文件,再利用扫描矢量化软件对栅格数据文件进行处理,将之转换为矢量图形数据。矢量化过程有两种方式,即交互式 and 全自动。影响扫描数字化数据

质量的因素包括原图质量(如清晰度),扫描(仪)精度,扫描分辨率、配准精度,校正精度等。

4.2 数字人体数据质量的数据转换误差 数字人体中的数据转换可能有数据结构变换、数据格式转换、数据计算变换等。

(1)数据结构的转换,主要包括栅格向矢量格式转换和矢量向栅格格式转换。栅格向矢量的转换其实就是矢量化。矢量数据转换为栅格数据,主要产生属性误差和拓扑匹配误差,包括像元属性值错误和边界重复加粗问题。

(2)数据格式转换,主要是指数据在不同文件格式之间的转换。在转换过程中,由于各系统内部数据结构不同和功能差异,往往会造成信息的损失,包括数据精度上的损失。

(3)数据计算变换,指通过各种计算方法对数据进行的处理,包括数据坐标变换、比例变换、投影变换等。变换过程中,由于算法模型本身的局限而引入误差。

4.3 数字人体数据质量的计算误差

(1)数据库多元数据综合的数据质量问题:数字人体的数据库建立后,其中已经包含了数据源和数据库建仓所引入的误差。数据库中的多源数据,经过系统的各种分析、处理后,可以形成新的数据和最后产品。在这个过程中还会产生新的数据质量问题,这些问题包括:计算误差、叠加分析引起的数据质量问题以及数字人体中的误差传播问题。

(2)拓扑叠加分析引起的数据质量问题:叠加分析是地理信息系统中很常用的一种分析方法。通过同一地区不同内容的多幅地图的叠加组合,产生新的图形和属性信息。在这个过程中,往往产生拓扑匹配、位置和属性方面的数据质量问题。

叠加时多边形的边界可能不完全重合,产生若干无意义多边形。对这些无意义多边形进行处理的结果往往会改变边界线的位置。叠加后形成的新的多边形,其属性值的确定也可能存在属性组合带来的误差。

(3)数字人体中的误差传播问题:数字人体中,由于从数据来源、数据库建立到数据的操作和使用都引入了各种误差因素,因此系统应用分析的最终结果中也包含了这些误差因素的影响。误差传播的研究目的就是研究初始过程和中间过程中引入的误差因素对于最后结果的影响,并模拟误差的变化。

误差传播的应用研究包括确定误差指标、建立误差传播函数,并通过对所生成的地理数据的空间、专题和时间精度的评定来分析误差指标和误差传播函数的实用性。每个误差传播函数的确定主要取决于三个因素:欲研究的特定误差的度量指标、所选择的地理信息系统数据变换函数以及对空间数据误差特性和误差传播机理所作的一组假设。

目前,在误差传播问题方面的应用研究已有不少进展,提出了一些误差传播模式和误差传播函数,如分层误差传播模式、再选择函数、相交函数等。由于对地理信息系统误差传播机理的认识还不够深入,误差传播的很多方面都还处于研究和试验阶段。但是,对于数字人体的建设者和应用者来说,了

解数据的各类误差均会以某种方式在系统中传播并将对系统的最后应用结果的质量产生影响,对理解数字人体数据和数据产品的可靠性将是十分有益的。

5 数字人体的数据质量评价^[4,5]

5.1 数字人体的数据质量模型

(1)直接数据质量模型:从用户角度出发,直接评价质量模型包括可用性、可共享性和可加工性三个方面。可用性就是指空间数据在精度、现势性、逻辑一致性、要素完备性等技术指标方面能否满足用户应用的要求;可共享性反映的是数据存储格式的通用性和开放性;可加工性则表示要素分类编码的合理性,即分类编码标准化的程度或反映现实世界真实性的程度。直接评价质量模型由用户层和技术指标层两个层次来表达(见图 1)。

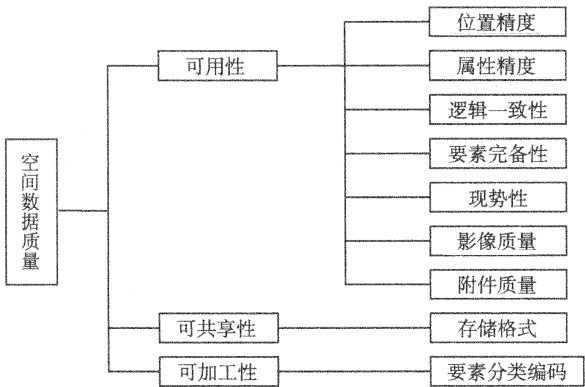


图 1 直接评价质量模型

(2)间接数据质量模型:间接评价依据是一些间接信息,没有直接面对最终的数据机,其质量模型的选取只能以间接信息来确定,而这些间接信息也只能按照数据集形成的数据流中的误差传播的途径来选择。间接质量模型通常有数据源质量、数据生产质量和数据加工处理质量三个层面。其中数据源质量也可进一步细分为源图形(像)数据质量、属性数据质量、文档数据质量;数据生产质量包括数字化质量和数据转换质量;数据生成、处理质量则分为计算误差、拓扑分析质量和图层叠置质量(见图 2)。

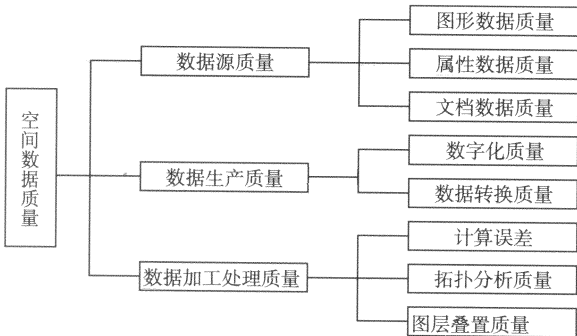


图 2 间接评价质量模型

(3)直接评价质量模型与间接评价模型的统一:两个质量模型的最底层不是一个统一的质量元素层。直接质量模型表现用户的可用性和可加工性的技术指标,对于数据集来说,已经反映在数据集的原数据中,无须评价,当用户需要时,可直接从元数据库中查询这些信息。因此,通过间接质量模型二级质量元素层的分析,其各项元素与之间质量模型的技术指标层存在着紧密的关联关系,如之间评价模型中的位置精度即是由间接质量模型中的原图形数据质量、数字化质量、数据转换质量、计算误差、拓扑分析质量、图层叠置质量的积累(见图 3)。

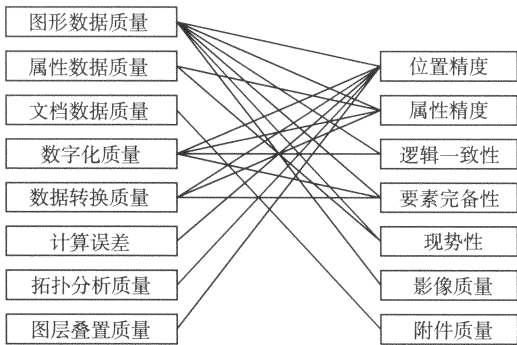


图 3 直接质量模型与间接质量模型的关系

5.2 数字人体的数据质量评价方法 数据质量评价过程指用于应用和报告评价方法的操作。

5.2.1 直接评价法

(1)数据质量定量元素度量方法:数据质量评价是对定量元素和非定量元素各方面作出衡量的过程,评价内容涉及定量元素及其子元素和非定量元素。由于数字人体所研究实体的复杂性和特性,以及人们对客观世界认识的不足和局限性、人类自身条件的限制、技术水平和仪器设备的限制,对数据质量的评价指标不能象传统测量数据那样只是精度或误差一项,而需要采用多种指标。一般可以用以下七种指标:数据溯源、精度(包括位置精度和属性精度)、逻辑一致性、完整性、可靠性、现势性和安全等级等。因此,在数字人体数据的存储、处理、叠加、综合时要考虑数据的这些特性。

(2)质量评价实施方法:①计算机程序自动检测方法:编写计算软件自动检测数据集中可以用逻辑判断对错的内容,同时也检测可能出现错误的内容,以缩小人工检查范围。当采用计算程序自动检测时,用于检测的自动检测程序的名称、算法和其他参考信息应当在数据质量报告中说明,以使评价结果可信。②人工检测法:借助于计算机或其他工具辅助或直接用人眼检查数据集错误。③随机抽样检查方法:对数据集的不同要素,在不同数据生产、集成阶段采用不同比例进行随机抽样检查。一般生产阶段自查或作业组检查比率为 100%,集成数据抽样比率为 10%,重要要素可以增大抽样比率。在抽样样本确定后,可以采用计算机或人工的方法检查。随机抽样检查方法应确定两个基本内容:①抽样样本:从被检

测的对象数据集总体中,随机抽取一部分为样本,通过对样本的检测判断数据集总体质量。②空间自动相关:数据生产者在设计抽样方案时,要进行抽样调查,以便识别有哪种空间自动相关及其特征。不同的自动相关应该采用不同的抽样方法。一般采用简单的随机抽样法,在复杂情况下,需要在有足够地理空间数据抽样检测经验的统计专家参加设计抽样方法。当应用基于抽样的评定方法时,评定过程必须作为质量报告的一部分予以简要说明。

5.2.2 间接评价法 间接评价法即基于相关知识或信息进行演绎推理来确定数据集的数据质量等级或符合程度。用于演绎推理的知识或信息包括用途、数据历史记录、数据源的质量、数据生产方法和系统的信息以及误差传递模型等。

用于演绎推理方法和知识或信息应当在质量评价报告中阐述,以提高演绎推理的可信度。

5.2.3 综合评价法 数据集、要素或属性的总体质量可以由综合它们的各组成部分的质量评定结果来确定。在综合过程中,需要考虑这些组成部分的出错率。

5.3 数字人体的数据质量评判 模糊综合评判是采用模糊数学方法,权衡各种影响所研究对象的因素,对所研究对象给出一个总概括式的优劣评价。对于空间数据而言,其评价步骤如下:

(1)根据对系统中空间数据质量的分析,得出其空间数据的质量特性包括:数据情况说明、属性数据精度、位置数据精度、数字化精度、逻辑一致性、数据完整性、数据安全性等 7 项作为评价因素集。

(2)建立评价集,对于系统中空间数据的质量,根据《测绘产品质量评定标准》的规定,设评价集为: $v=\{v_1(\text{优}),v_2(\text{良}),v_3(\text{合格}),v_4(\text{不合格})\}$ 。

结合数字测绘产品质量评定标准得出各质量特性的得分值,并由此值确定各评判数据质量特性的隶属程度。

(3)对各指标进行一级综合评价,确定评价因素的权重集 $A=\{0.15,0.20,0.25,0.15,0.10,0.10,0.10,0.05\}$,A 的确定可由专家评分再取均值,然后归一化而获得。在研究复杂问题时,需要考虑的因素很多。而且这些因素往往还分为不同的层次,如对反映空间数据质量的数据质量特性,可将其划

分为一线质量特性,二线质量特性等不同层次。为在较小范围内便于比较少数因素的相对重要性,便于确定各级评判中的权向量,常先把所有因素按某些属性分成几类,在每一类范围内进行综合评判,然后再根据各类评判结果进行高一级的综合评判,以至更多级综合评判。

(4)二级综合评判模型:在一级评判之前进行二级评判,以 B 作为其单因素评价,又可构成二级评价矩阵。

(5)模糊识别:得出系统的最高级别综合评判后,可根据最大隶属度原则进行模相识别或按择近原则归类,确定系统数据质量等级。

[参考文献]

[1] Bi SW, Wang X, Luo SQ, et al. Research and application in fundamental theory and framework of digital human body (human body system digital science) [J]. Chin J Med Imaging Technol, 2003,19(1):4-7.
毕思文,汪湘,罗述谦,等.数字人体(人体系统数字学)基础理论框架研究与应用[J].中国医学影像技术,2003,19(1):4-7.

[2] Bi SW. Metadata of space data of digital human body [J]. Chin J Med Imaging Technol, 2003,19(Suppl):140-145.
毕思文.数字人体空间数据的 Metadata [J].中国医学影像技术,2003,19(增刊):140-145.

[3] Bi SW. Pandect on digital human body-human body digital science [J]. Chin J Med Imaging Technol, 2003,19(Suppl):1-8.
毕思文.数字人体-人体系统数字学总论 [J].中国医学影像技术,2003,19(204):1-8.

[4] Bi SW. Study on digital human body standardization system [J]. Chin J Med Imaging Technol, 2004,20(2):286-289.
毕思文.数字人体标准化体系研究 [J].中国医学影像技术,2004,20(2):286-289.

[5] Bi SW. Study of currency standard and speciality standard system framework of digital human body [J]. Chin J Med Imaging Technol, 2004,20(3):460-464.
毕思文.数字人体通用标准与专业标准体系框架研究 [J].中国医学影像技术,2004,20(3):460-464.