· 技术 / TECHNOLOGY ·

基于知识规则的 Excel 数据质量校验工具

苏贤明，沈志宏，刘宁

中国科学院计算机网络信息中心 科学数据中心，北京 100190

摘　要：　在分析现有数据质量校验方法与校验工具的基础上，借鉴科研领域的数据质量校验经验和规则引擎的相关技术，实现了基于知识规则的 Excel 数据质量校验工具，进而解决科研观测数据中异常记录判别、异常原因标识、数据可视化分析等关键技术问题。中国生态系统研究网络综合中心以及土壤分中心的应用表明，在不影响原有数据填报流程的前提下，该工具能很好地代替数据质量校验人员的手工查错工作，有效地提高数据质量校验的效率及准确性。

关键词：　知识规则；Excel；数据质量校验

Excel Data Quality Validation Tool Based on Knowledge Rules

Su Xianming, Shen Zhihong, Liu Ning

Scientific Data Center, *Computer Network Information Center*, *Chinese Academy of Sciences*, *Beijing 100190*

Abstract: Reviewing the existing methods and tools for data quality validation, this paper presents the development of an Excel data quality validation tool based on the customized knowledge rules database, learned from the experiences of data quality validation in scientific research. A number of key technical issues were solved in the research and observational data such as the discrimination of exception record, the identity of the reason for the exception, data visualization analysis and so on. The applications in Institute of Geographical Sciences and Natural Resources Research and Nanjing Institute of Soil, Chinese Academy of Sciences, showed that the tool could take the place of manual troubleshooting work and improve the efficiency and accuracy greatly in the data quality validation under the premise that the existing data reporting process was not affected.

Keywords: Knowledge rule; Excel; Data quality validation

基金项目：中国科学院计算机网络信息中心青年基金项目 (CNIC_QN_09007)

1 引言

随着数字化技术与网络的发展，科学研究日益成为数据密集型的工作。未来的科学技术创新将越来越倚重于科学数据，以及通过数据挖掘、集成、分析与可视化工具将其转换为信息和知识的能力。另一方面，科学数据资源的物理分布广泛、结构各异、关系复杂以及数据质量校验手段薄弱等特点，导致科学数据质量参差不齐。科学数据质量是科学研究的生命，如果科学数据质量无法保证，那么基于该数据所产生的科研结果将失去其价值。科研数据源中可能包括噪声数据、重复数据、不一致数据、错误数据和异常数据等，根据“进去是垃圾，出来也是垃圾 (garbage in, garbage out)”的规律，保障科研数据质量便成为数据集成与分析乃至科学研究的重要方面。

很多科研数据尤其是观测数据，由于观测人员的使用习惯以及需要适当的数据预处理等原因，无法实行在线入库，一般借助于 Excel 等办公软件来记录观测结果。对于数据质量的控制，大多依靠专家经验，这样的操作难免存在着遗漏，且发现问题常常滞后很长时间，还存在着人员变动带来的数据校验不一致性。因此，设计并实现 Excel 数据质量校验工具对于保证 Excel 数据源的科研观测数据质量具有重要意义。

本文在阐述数据质量问题定义及层次的基础上，通过分析现有数据质量校验方法与校验工具，借鉴科研领域的数据质量校验经验，研发基于可定制规则库的 Excel 数据质量校验工具，进而解决科研观测数据中异常记录判别、异常原因标识、数据可视化分析等关键技术问题。中国生态系统研究网络综合中心以及土壤分中心的应用表明，在不影响原有数据填报流程的前提下，该工具能很好地代替数据质量校验人员的手工查错工作，极大地提高了数据质量校验的效率及准确性。

2 数据质量校验方法与工具

2.1 数据质量问题定义

数据质量被定义为数据的一致性 (Consistency)、正确性 (Correctness)、完整性 (Completeness) 和可靠性 (Reliability) 在信息系统中得到满足的程度。数据质量衡量指标分为两类：数据质量指示器和数据质量参数。前者是客观的信息，比如数据的收集时间、来源等，而后者是主观性的，比如数据来源的可信度 (Credibility)、数据的及时性 (Timeliness) 等[1]。按数据质量问题所处的层次可分为模式 (Schema Level) 层数据质量问题和实体层 (Instance Level) 数据质量问题。数据质量问题的分类如表 1 所示[2]。

2.2 数据质量校验方法与工具

随着科学数据资源的急剧膨胀以及海量数据处理等相关技术的发展，数据质量校验方法与技术逐渐成为国内外学者的研究热点。房强[3]通过设计面向半结构化数据的数据质量控制模型，实现了半结构化数据的质量检测、问题数据处理以及数据质量评估功能，提出了该模型对半结构化类型数据的抽象方法，很好的解决了半结构化数据的异构问题。支丽凤[4]针对数据移植过程中的数据质量问题，建立了一套全面

表 1 数据质量问题分类

问题分类	层次	原因	典型的表现形式
单数据源问题	模式层	缺少完整性约束，糟糕的模式设计	唯一性约束 引用约束
	实例层	数据记录的错误	拼写错误 相似重复记录 互相矛盾的字段
多数据源问题	模式层	异构的数据模型和模式设计	命名冲突 结构冲突
	实例层	冗余、互相矛盾或者不一致的数据	不一致的汇总 不一致的时间选择

的数据质量控制体系，从而保证了信息系统在数据移植过程中的数据质量。殷俊[1]提出了基于统一工作流的 ETL 模型进行数据质量控制的方法，基于这种模型，每个数据表的 ETL 流程都按照 ETL 的特性统一分为 3 个标准步骤，即数据抽取/变换、数据转换和数据加载，每个步骤需要记录完整的处理中间状态及完善的日志信息。周世健等[5]认为动态数据的质量控制是要求其动态数据处理过程中能够自动检测出系统模型中的偏差, 消除这些偏差对估计结果的影响，并以 Kalman 滤波作为动态数据处理方法, 讨论了含有偏差的动态数据处理模型, 基于偏差探测的假设, 构造了偏差探测的统计检验量及基本公式, 并对偏差识别的过程及统计检验作了分析, 最后得到了偏差估计和校正的理论公式。韩成贵等[6]在深入分析人工智能领域内专家系统中的规则和推理、传统数据库中数据约束的缺陷等基础上，实现了基于知识库的数据校验。肖明[7]通过规则库来扩充数据库应用系统的校验功能，使系统能够对规则进行动态读取与分析，并根据分析结果对输入数据进行动态校验。

在地理数据质量校验领域，郭仁安[8]提出了简单对比检查的 GIS 属性数据质量控制模型，并根据概率论与数理统计理论，借助缺陷率这一指标，采用一般随机抽样和分层抽样统计模型计算缺陷率来对属性数据进行质量控制和研究。袁琳[9]通过设计面向对象约束条件的数据特征模型来对不同格式的原始数据进行质量检查和质量加强。这个面向对象的数据质量检查模型，可以通过对要素关系、行为和有效规则的定义来实现对现实世界系统更好地表达和描述，然后利用 ESRI 公司的 Geodatabase 这种已有的能够支持特征关系、行为和规则定义的高级特征模型来演示这种面向对象思想的数据检查模型是如何对不同格式的原始地理数据质量进行控制检查和约束加强。沈陈华[10]则在江苏省部分县市 1:1 万土地利用现状调查数据建库基础上, 研究了建库过程中影响数据质量的主要因素和因子，给出了影响数据质量的主要因素与因子的权重，并以模糊数学原理和方法为基础, 建立了数据质量综合评价模型, 同时利用 SPC 方法初步研究了数据建库过程中数据质量的控制。

数据质量校验工具的发展对于促进数据质量校验的研究具有重要意义。如针对专有数据格式的 MD5 值校验工具、Cmis30 数据校验工具、CRC 数据校验计算工具，针对 Web 表单数据的 Java 开源数据校验框架 Validator[11]，开源和商用的数据清洗工具 DataCleaner[12]、WinPure Clean & Match[13]、MatchIT Data Cleansing Software Suite[14]等。此外，Xin Wang 等[15]提出一种基于本体的数据清洗框架，可以在知识层而不仅仅是数据层发现数据质量问题。叶舟等[16]描述了基于规则引擎的数据清洗方法，其实验结果表明利用规则引擎比硬编码更加高效与灵活。

综上所述，模式层的转换集成成为国内外学者的研究重点，而对于实例层次上的数据质量问题，因其领域相关性，所受到的关注并不多。尤其对于 Excel 数据源的实例层次数据质量校验，大部分的研究集中在利用 Excel 软件自身的函数[17]或者 VBA[18]等功能来完成数据质量校验功能，其功能很难满足科研人员的需求，且难以扩展。

3 基于知识规则的 Excel 数据质量校验工具

规则引擎技术的发展使定制形色各异的校验规则成为可能。本文通过借鉴规则引擎的相关思想，基于 Eclipse RCP 相关技术，实现了基于知识规则的 Excel 数据质量校验工具。

3.1 需求分析

基于知识规则的 Excel 数据质量校验工具包括管理员与数据质量校验人员两类角色。

数据质量校验人员的需求如下：

(1) 数据转换：根据管理员配置好的转换规则，将数据库中的表转换为 Excel 实体表，以进行后续数据校验；

(2) 数据校验：利用管理员配置的校验规则对指定 Excel 数据实体进行数据校验；

(3) 可视化分析：对已校验数据进行多指标曲线统计、多空间曲线统计、多指标曲线统计、多时间点柱状统计等，以发现离群点；

管理员具备数据质量校验人员的所有需求，同时管理员还存在如下额外需求：

(1) 配置数据源；

(2) 数据实体的校验规则配置与管理，包括规则的导入、导出、复制、粘贴、增、删、改、查等；

(3) 对系统所有的用户进行管理，包括增、删、改、查等功能；

(4) 配置用户对转换规则和校验规则的可见性；

(5) 对所有的元数据进行管理，包括增、删、改、查等功能。

3.2 系统构成

根据上述需求，本文将基于知识规则的 Excel 数据质量校验工具设计为规则配置与数据质量校验两大模块。其中，数据质量校验模块包括数据转换、数据校验以及统计分析三个部分。数据质量校验人员登录后，校验系统根据用户的操作权限，提取相应的数据字典信息，封装成元模型对象。界面菜单的内容以及转换的参数等都通过元模型来生成。

(1) 规则配置：管理员配置转换规则与数据校验规则，同时配置统计分析的时间与空间等相关参数。规则定制可以参考历年观测数据的均值、最大值、最小值、年变异、空间变异以及观测方法等统计值与相关知识。同时，可以通过扩展相关的类来实现规则的定制与扩展。

(2) 数据转换：如用户需要校验和统计分析的数据存储于数据库中，可以根据工具提供的数据转换功能将相应的数据转换为 Excel 格式。同时，当用户导入 Excel 时，还可以选择需要导入并进行校验的字段，以屏蔽冗余数据。

(3) 数据校验：根据管理员定义的数据校验规则，对目标数据表进行校验，并将错误记录输出。

(4) 统计分析：用户可以根据需求选择相应的分析种类，进而发现离群点。

(5) 数据输出：用户可根据校验结果，选择是否将结果输出。输出有两种方式：导出为 Excel 表和输出到数据库中长期保存。

3.3 知识规则管理

本文定义了知识规则及其参数配置的标准格式 (Rule Definition Format, RDF)，规则存放于 XML 格式的规则文件中，便于规则的分发、编辑与修改，同时具有跨域与跨平台的特性，便于扩展和移植。规则描述的语言基于开源的模板语言 Velocity。同时，为支持规则扩展，定义了开放的、可扩展的规则工具箱 (Rule Toolbox)，用工具箱中提供的工具操作可轻松组装成一条数据校验规则。

基于知识规则的 Excel 数据质量校验工具根据领域专家的相关知识，内置了比较型、格式型、查找型等三种类型的数据校验规则。管理员还可根据自己业务上的需要自定义相应的规则，进行数据校验。

知识规则主要包括以下几项：

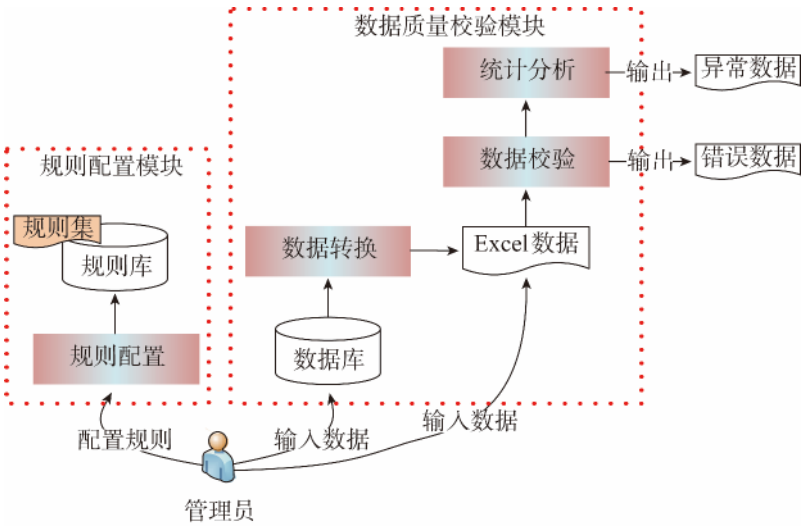

图 1 基于知识规则的 **Excel** 数据质量校验工具的系统结构

(1) title：规则名称，简要表示规则的含义；

(2) code：规则在一个规则文件中的唯一标识符，为以后进行规则编辑时预留；

(3) description：规则描述信息，可用几十个字描述规则的含义；

(4) suggestion：规则建议，用来表示数据不符合所指定规则时，用户该如何操作；

(5) expression：规则表达式，由一条 Vecocity 语句组成，通过规则引擎对校验数据进行校验，返回结果为 true 或 false；

(6) params：规则中的参数，由一系列规则参数 param 组成，param 组成规则表达式中的某一项用以规则编辑，当规则校验时这些参数会由规则引擎根据校验对象变成具体的值，进行表达式的计算。 规则参数 param 主要由以下几项组成：

① title：参数名称，在显示规则和编辑规则时，用来表示该参数的含义；

② name：参数名称，用来编辑和在规则表达式中引用参数的时候用；

③ value：参数值，现有的参数只支持字面值，如 ABC，或 123；

④ type：参数类型，现在只有两种参数类型：Text 和 Number。

一条具体的参数定义如下：

```
<param>
    <title>样地代码字段名</title>
    <name>plot_code</name>
    <value>Plot_code</value>
    <type>TEXT</type>
</param>
```

具体的规则示例如下：

```
<rule>
    <title>样地代码校验</title>
    <code>PLOT_CODE_RULE</code>
    <description>库中找不到对应的样地代码值：
        $record.get($plot_code)</description>
    <suggestion>请检查</suggestion>
    <expression>$tool.dict($plot_info,$plot_code).
        contains($record.get($plot_code))
    </expression>
    <params>
        <param>
            <title>样地代码表名</title>
            <name>plot_info</name>
            <value>Argo_plot_info</value>
            <type>TEXT</type>
        </param>
        <param>
            <title>样地代码字段名</title>
            <name>plot_code</name>
            <value>Plot_code</value>
            <type>TEXT</type>
        </param>
    </params>
</rule>
```

规则表达式主要由四类符号组成：包括逻辑运算符、四则运算符、取值运算符和比较运算符。运算符的含义及操作如表 2 所示。

表 2　运算符含义及操作

类别	符号	含义
四则运算符	+ - * /	对数值进行加减乘除
逻辑运算符	&&	逻辑与，符号‘ & ’在XML文件中需要写成‘ & ’
	||	逻辑或
	!	逻辑非
比较运算符	>或gt	大于(‘ > ’在XML文件中需要写成‘ > ’)
	>=或ge	大于等于
	<或lt	小于(‘ < ’在XML文件中需要写成‘ < ’)
	<=或le	小于等于
取值运算符	$record.get(xxx)	获得字段xxx的值

3.4 数据转换

基于 Excel 文件对科研观测数据进行校验是本文的出发点，为更好的支持数据库中相关历史数据的校验，本文通过研究数据库表与 Excel 表的相关特性，利用数据框架定义出专业报表的结构，并基于此开发完成数据转换功能。数据转换规则亦采用 XML 文件进行配置，可灵活定制与扩展。

3.5 数据校验

本文在借鉴规则引擎相关思想的基础上，设计并开发了数据校验模块。模块有两类输入数据：知识规则和待校验数据，各个部分可分解为：

(1) 管理员通过规则配置模块对数据校验规则进行配置，规则集合由规则导入模块载入引擎的工作内存；

(2) 数据质量校验人员通过外部程序输入待校验数据，进入规则引擎；

(3) 执行引擎通过后置处理器反馈给外部程序。这样的外部程序包括校验结果的显示界面，以及其它的异常通知程序，如邮件等处理接口。

规则引擎部分的模块结构图如图 2 所示。

规则引擎的执行过程主要有以下几个步骤：

(1) 规则解析器将 XML 格式的规则文件，解析封装为一个个规则对象，规则加载器将这些对象加载至工作内存；

(2) 数据加载器解析用户载入的 Excel 文件，封装为数据对象列表，并置于工作内存；

(3) 执行引擎接受外部数据校验事件的触发，将所得规则作用于对应的数据对象列表，在执行引擎的过程中，需要使用到规则语言解释模块，以及对数据对象进行预处理的前置处理器模块和触发外部事件的后置处理器模块。

3.6 数据统计分析

数据统计分析模块以图示化的方式直观地表示出多个指标数据的时间变化趋势或者不同空间对象上单一指标数据的时间变化趋势，既可以进一步发现数据超界或数据突变问题，也可以了解数据变化的趋势。数据统计分析模块内嵌 Web 容器，通过 Web 界面使用 Flash 的图形插件展示统计分析数据。统计分析模块基于元数据的配置确定指标、空间信息和时间信息创建界面，在界面内置浏览器用以展示统计分析的结果。用户选择好相应的参数，提交统计请求后，系统自动封装这些参数创建相应的 Http 请求，提交至内置的 Web 容器，内置的 Web 容器调用后台的分析处理器对该请求进行处理并做出响应，将结果返回至展示界面。其执行流程如图 3 所示。

数据统计分析模块的功能展示分曲线图和柱状图两种类型，曲线图和柱状图各自分别存在两种情况的统计展示。

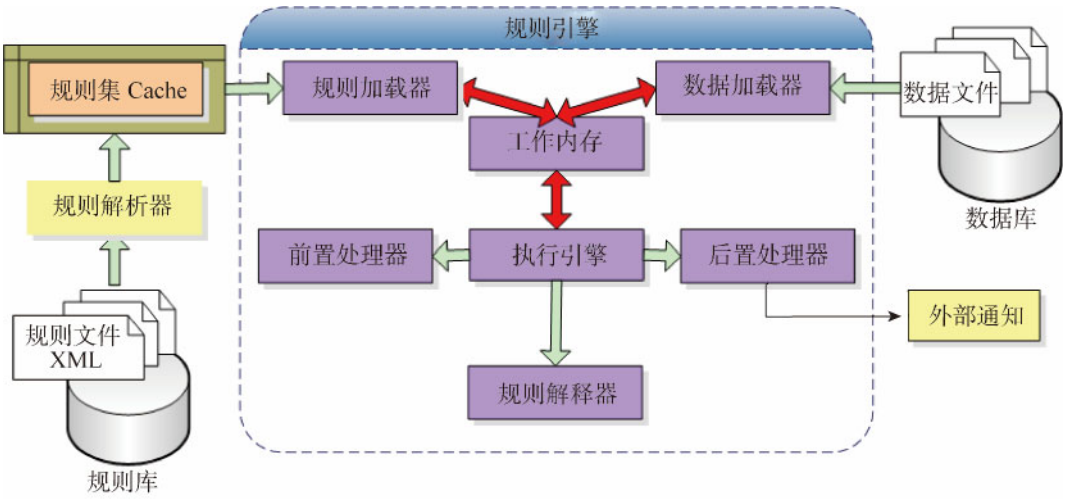

图 2 规则引擎模块结构图

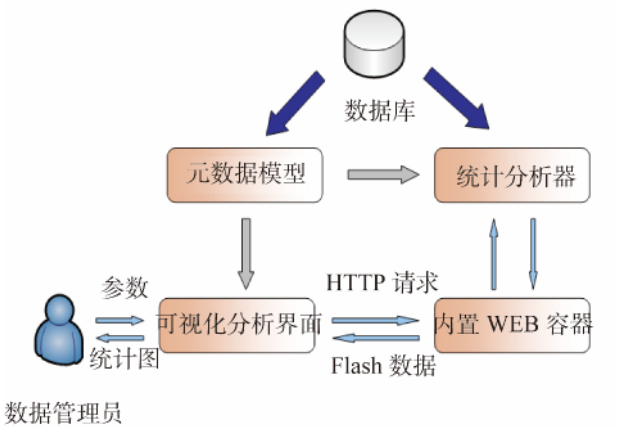

图 3　统计分析模块执行流程

(1) 多个空间点 (样地) 在同一时段内的某项指标的统计。用户可选择多个空间点 (样地)，选择某项指标，自定义一个时间段，则图表中展示在这个时间段内选中的多个空间点 (样地) 该项指标的值变化曲线，如图 4 所示。

(2) 单个空间点 (样地) 在同一时段内的某几项指标的统计。用户可选择某个空间点 (样地)，选择多项指标，自定义一个时间段，则图表中展示在这个时间段内选中的这个空间点 (样地) 这几项指标的值变化曲线，如图 5 所示。

(3) 多个空间点 (样地) 在某一时间点的某几项

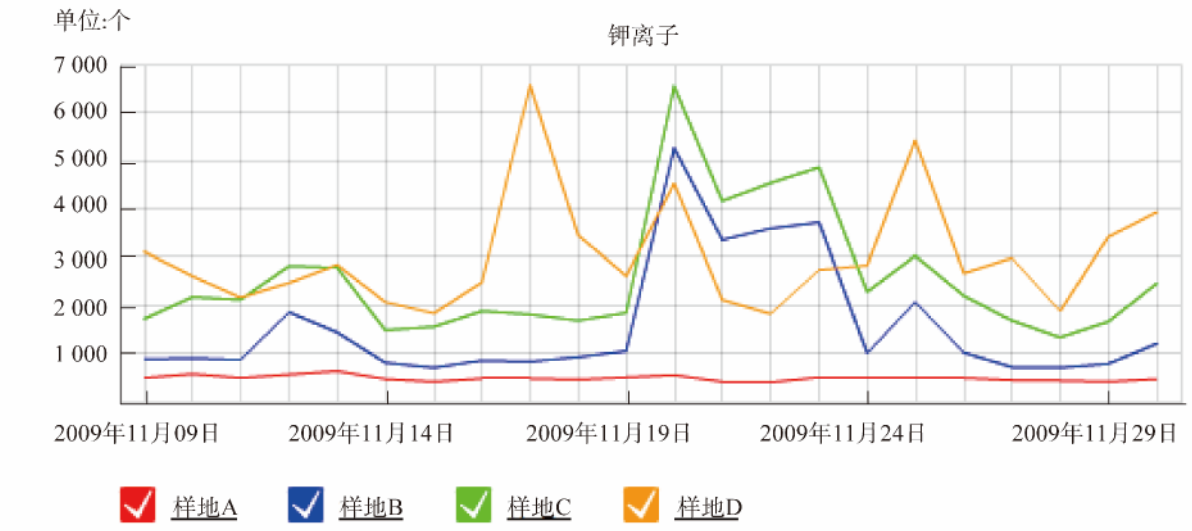

图 4　多空间曲线统计示例图

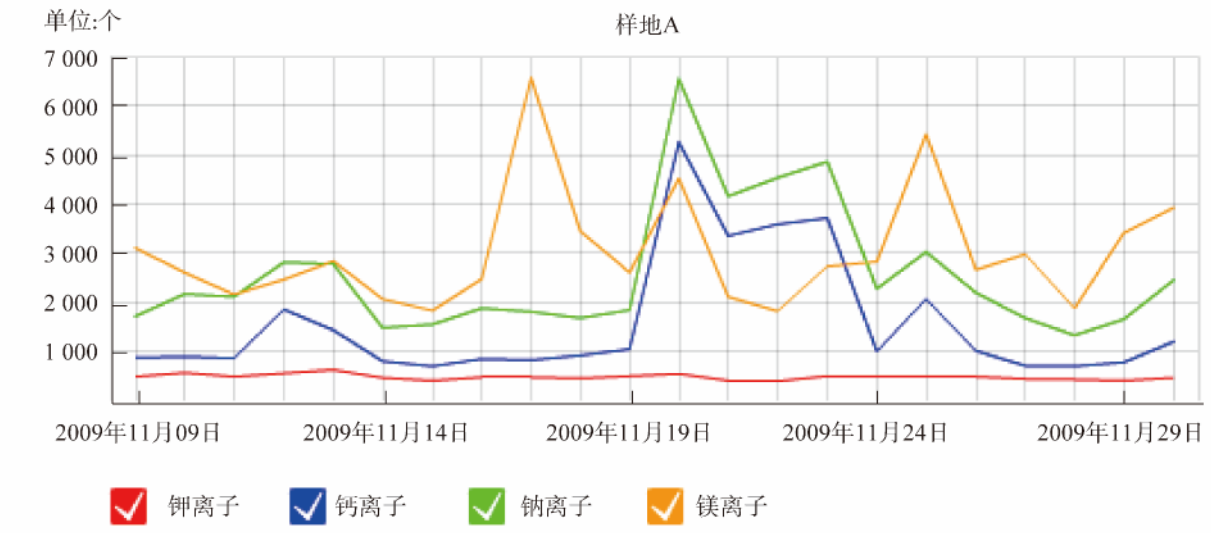

图 5　多指标曲线统计示例图

指标的统计。用户可选择多个空间点 (样地)，自定义一个时间点 (比如说 2009 年或者 2009 年 8 月或者 2009 年 8 月 8 日，在这种统计情况中一个空间点在一个时间点该项指标只有一个值)，则图表中展示在这个时间点的各个空间点 (样地) 这几项指标的值情况，如图 6 所示。

(4) 多个空间点 (样地) 的某项指标在多个时间点的统计。用户可选择多个空间点 (样地)，选择某项指标，自定义几个时间点，则图表中展示这几个空间点 (样地) 这项指标在不同时间点的值变化情况，如图 7 所示。

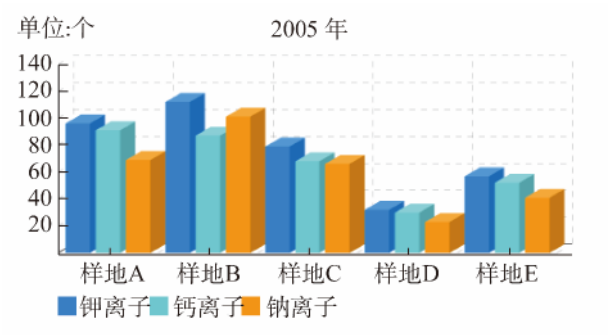

图 6　多指标柱状统计示例图

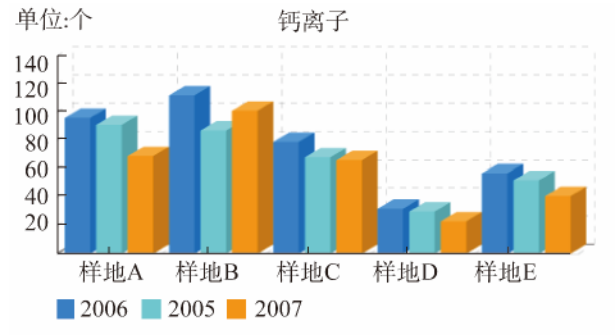

图 7　多时间柱状统计示意图

4　应用

基于知识规则的 Excel 数据质量校验系统目前已经应用到中国生态系统研究网络 (CERN) 综合中心以及土壤分中心，如图 8 和图 9 所示，用于对长期生态监测数据进行质量控制。综合中心利用该系统制定校验规则 300 多条，并依据这些规则对 CERN 的大气观测数据、土壤观测数据、水分观测数据、生物观测数据共计 200 多张数据表格、1 800 万个数据进行数据校验，辅助数据管理人员发现数据错误。土壤分中心利用该系统对 2009 年农田生态系统 18 个陆地站的土壤监测数据进行校验，涉及不同地区农田土壤交换量、农田土壤养分、农田土壤矿质全量、农田土壤微量元素和重金属元素、农田土壤硝态氮和氨态氮、农田土壤速效微量元素、农田土壤颗粒组成和容重的阈值校验、统计校验和关联知识校验等。同时，土壤分中心通过建立统一的背景参考库，设定相应的阈值标准，对观测数据的样地代码、方法一致性、标准样品方差等数据进行校验，保证了全局分析数据的一致性。

图 8　基于知识规则的 **Excel** 数据质量校验系统在地理所的应用

图 9　基于知识规则的 **Excel** 数据质量校验系统在南京土壤所的应用

应用表明，基于知识规则的 Excel 数据质量校验系统简便易用，实用性、通用性、扩充性强，具有较强的推广价值，能够很好的支撑综合中心、土壤分中心以及观测台站的数据质量校验工作，提高工作效率，具有很好的应用前景。

5 总结与展望

应用表明，在不影响原有数据填报流程的前提下，基于知识规则的 Excel 数据质量校验工具能很好地代替数据质量校验人员的手工查错工作，极大地提高数据质量校验的效率及准确性。本文所提出的数据质量校验模型虽然主要针对 Excel 数据源，但其灵活的知识规则定制方式，可以应用于任何关系型数据库管理系统 (DBMS) 中的任何关系表。对于其它领域的数据校验，仍有借鉴意义。

参考文献

[1] 殷俊. 基于ETL技术的电信业数据仓库质量控制模型的研究及应用. 上海: 上海交通大学, 2006.

[2] 郭志懋, 周傲英. 数据质量和数据清洗研究综述. 软件学报, 2002, 13(11): 2076–2082.

[3] 房强. 面向半结构化数据的数据质量控制系统的研究与实现. 沈阳: 东北大学, 2008.

[4] 支丽凤. 数据移植过程中的数据质量控制方法的研究. 上海: 同济大学, 2007.

[5] 周世健, 鲁铁定, 臧德彦. 动态数据自适应控制中的质量控制. 江西科学, 2006, 24(6): 414–416.

[6] 韩成贵. 基于知识库的数据校验. 北京: 北京工业大学, 2003.

[7] 肖明. 基于规则的数据校验在数据库应用系统中的实现. 计算机与信息技术, 2007, (21): 337–339.

[8] 郭仁安. GIS中属性数据质量控制的研究与探讨. 地理信息世界, 2001, (1): 19–22.

[9] 袁琳. 地理信息数据录入的数据质量控制研究. 青岛: 中国海洋大学, 2009.

[10] 沈陈华. 土地利用现状数据建库数据质量评价与控制研究. 遥感信息, 2006, (1):50–53.

[11] http://commons.apache.org/validator/.

[12] http://datacleaner.eobjects.org/.

[13] http://www.winpure.com/.

[14] http://www.helpit.com/.

[15] Xin Wang, Howard J. Hamilton, and Yashu Bither. An Ontology-Based Approach to Data Cleaning. Technical Report July, 2005.

[16] 叶舟, 王东. 基于规则引擎的数据清洗. 计算机工程, 2006, 32(23):52–54.

[17] 汪洋. 快速校验Excel数据的两种方法. 电脑知识与技术, 2009, 5(15):4069–4074.

[18] 亓靓, 宋传真, 唐俊. Excel VBA对环境统计软件数据审核功能的改进. 环境科学与管理, 2010, 35(11): 36–38.

收稿日期：2012 年 3 月 18 日

苏贤明：中国科学院计算机网络信息中心科学数据中心，工程师，硕士，研究方向为信息系统开发。
E-mail: *suxianming@cnic.cn*

沈志宏：中国科学院计算机网络信息中心科学数据中心，高级工程师，研究方向为科学数据集成与管理。
E-mail: *bluejoe@cnic.cn*

刘　宁：中国科学院计算机网络信息中心科学数据中心，工程师，硕士，研究方向为科学数据治理和数据质量管理。E-mail: *ln@cnic.cn*