



(12) 发明专利申请

(10) 申请公布号 CN 103440164 A

(43) 申请公布日 2013. 12. 11

(21) 申请号 201310262859. 5

(22) 申请日 2013. 03. 28

(30) 优先权数据

1226/CHE/2012 2012. 03. 29 IN

(71) 申请人 穆西格马交易方案私人有限公司

地址 印度班加罗尔

(72) 发明人 K·克里斯南 D·德欣格拉

(74) 专利代理机构 中国专利代理(香港)有限公司

司 72001

代理人 徐小会 王忠忠

(51) Int. Cl.

G06F 9/46(2006. 01)

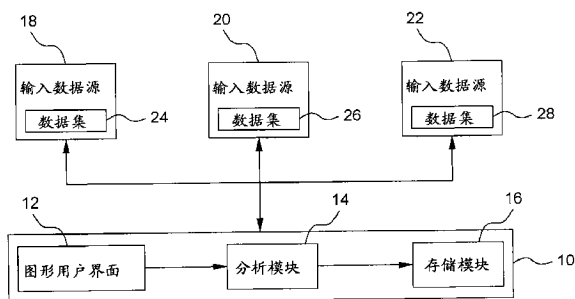
权利要求书2页 说明书7页 附图11页

(54) 发明名称

数据解决方案系统

(57) 摘要

本发明是数据解决方案系统。一种用于分析多个数据集以确定针对特定问题的一个或多个解决方案的系统被提供。该系统包括配置为从多个源接收多个数据集并且使用数据处理模块来分析所述多个数据集的分析模块,所述数据处理模块被配置为将所述多个数据集转换为分析数据集。该系统还包括配置为确定存在于所述分析数据集内的多个相关性的探索性分析模块,其中所述多个相关性被用于确定所述一个或多个解决方案。该系统还包括耦合于所述分析模块并且配置为使一个或多个用户能与所述分析模块进行交互的图形用户界面以及配置为存储所述多个数据集和所述分析数据集的存储模块。



1. 一种用于分析多个数据集以确定针对一个或多个问题的一个或多个解决方案的系统,所述系统包括:

配置为从多个源接收多个数据集并且使用下列模块来分析所述多个数据集的分析模块;

配置为将所述多个数据集转换为分析数据集的数据处理模块;

配置为确定存在于所述分析数据集内的多个相关性的探索性分析模块;其中所述多个相关性被用于确定所述一个或多个解决方案;

耦合于所述分析模块并且配置为使用户能与所述分析模块进行交互的图形用户界面;以及

配置为存储所述多个数据集和所述分析数据集的存储模块。

2. 如权利要求 1 所述的系统,其中所述分析模块还包括数据建模模块,其被配置为生成表示由所述探索性分析模块生成的一个或多个解决方案的一个或多个模型。

3. 如权利要求 2 所述的系统,其中基于所述分析数据集的平均值、方差以及协方差生成所述一个或多个模型。

4. 如权利要求 1 所述的系统,其中所述分析模块还包括报告模块,其被配置为使用户能在单个位置访问由所述探索性分析模块和所述数据建模模块生成的多个报告。

5. 如权利要求 1 所述的系统,其中所述分析模块还包括质量分析模块,其耦合于所述数据处理模块并且配置为评估所述分析数据集的质量。

6. 如权利要求 1 所述的系统,其中所述探索性分析模块被配置为对所述分析数据集应用一元分析,其中所述一元分析包括根据一个或多个统计属性表示所述分析数据集。

7. 如权利要求 1 所述的系统,其中所述探索性分析模块被配置为对所述分析数据集应用多元分析;其中二元分析包括确定相对于一个或多个统计属性的变化。

8. 如权利要求 7 所述的系统,其中所述探索性分析模块还被配置为生成所述分析数据集的可视化表示。

9. 如权利要求 1 所述的系统,其中所述分析模块还包括分割模块,其被配置为基于属性将所述分析数据集分组,其中所述属性由用户选择。

10. 如权利要求 8 所述的系统,其中由所述分析模块使用的多个边界参数由用户定义。

11. 一种计算机实现的系统,其包括:

一个或多个处理器;

一个或多个非暂时性的计算机可读存储介质,其包含配置为使所述一个或多个处理器执行操作的指令,所述操作包括:

从多个源接收多个数据集;

调节所述多个数据集以生成分析数据集;

对所述分析数据集执行探索性数据分析以确定存在于所述分析数据集内的多个相关性;

基于所述探索性数据分析的结果生成多个模型;其中每个模型提供一个或多个解决方案以达到由用户确定的预定义目标。

12. 如权利要求 11 所述的系统,其还包括评估所述分析数据集的质量。

13. 如权利要求 11 所述的系统,其还包括为探索性数据分析和数据建模生成多个报告。

14. 如权利要求 13 所述的系统,其还包括存储所述多个报告以使用户能从单个位置访问所述多个报告。

15. 如权利要求 11 所述的系统,其还包括基于由用户选择的属性将所述分析数据集分组。

数据解决方案系统

技术领域

[0001] 本发明涉及数据解决方案系统 (data solution system) 及技术。更具体地, 本发明涉及分析从多个源接收的若干数据集以为特定问题提供一个或多个最佳解决方案。

背景技术

[0002] 近来, 随着分析行业日益成熟并且竞争越来越激烈, 越来越需要证明分析花费的投资回报率 (ROI) 是合理的并且证明其商业价值。使分析保持商业的速度是至关重要的, 尤其是随着基于分析进行决策的商业问题的范围和数量呈指数增加。在当今的迅速发展的全球商业环境中, 对有力的分析解决方案的需要比以前大。

[0003] 然而, 对于现有解决方案的一些重要挑战是在组织内推动最佳实践以及确保团队之间的协作和相互学习的困难。还需要从编码和执行到商业解释释放并且重新规划资源的时间。此外, 提供在执行分析时朝最佳实践推动的工具是所期望的。

[0004] 因此, 需要可以构建实现复用性以及减少用于新雇员的斜升时间并且使来自当前的基础设施投资的价值最大化的平台的系统和方法。

发明内容

[0005] 简要地说, 根据本发明的一个实施例, 一种用于分析多个数据集以确定针对一个或多个问题的一个或多个解决方案的系统被提供。该系统包括配置为从多个源接收多个数据集以及使用数据处理模块来分析所述多个数据集的分析模块, 所述数据处理模块被配置为将所述多个数据集转换为分析数据集。该分析模块还包括配置为确定存在于所述分析数据集内的多个相关性的探索性分析模块; 其中所述多个相关性被用于确定所述一个或多个解决方案。该系统还包括耦合于所述分析模块并且配置为使一个或多个用户能与所述分析模块进行交互的图形用户界面以及配置为存储所述多个数据集和所述分析数据集的存储模块。

[0006] 在另一实施例中, 包含一个或多个处理器的计算机实现系统被提供, 其包括一个或多个非暂时性的计算机可读存储介质。该系统包括配置为使所述一个或多个处理器执行操作的指令, 所述操作包括: 从多个源接收多个数据集、调节所述多个数据集以生成分析数据集以及对所述分析数据集执行探索性数据分析以确定存在于所述分析数据集内的多个相关性。该处理器还执行包括基于所述探索性数据分析的结果生成多个模型的操作, 其中每个模型提供一个或多个解决方案以达到由用户定义的目标。

附图说明

[0007] 在参照附图阅读以下详细说明时将会更好地理解本发明的这些及其他特征、方面和优点, 在附图中, 相似的标记贯穿附图表示相似的部件, 其中:

[0008] 图 1 是根据本发明的各个方面实现的数据分析系统的实施例的框图;

[0009] 图 2 是根据本发明的各个方面实现的分析模块的实施例的框图;

[0010] 图 3 是示出根据本发明的各个方面的、来自不同源的各种数据集根据其被处理的一种方法的流程图；

[0011] 图 4 是根据本发明的各个方面实现的通用计算机的框图；以及

[0012] 图 5 至图 12 示出根据本发明的各个方面实现的图形用户界面的示例截屏。

具体实施方式

[0013] 在以下详细说明中对附图进行了参考，所述附图形成该详细说明的一部分。在附图中，类似的符号典型地标识类似的组件，除非上下文另有指示。在详细说明、附图以及权利要求中所描述的示意性实施例不是旨在限制。可以使用其他实施例，并且做出其他改变，而不背离在本文中所呈现的内容的精神或范围。将容易地理解的是，如在本文中一般地描述的以及在附图中所示出的那样，可以各种各样不同的配置布置、替换、组合、分开以及设计本公开内容的各个方面，所有这些都都在本文中明确被考虑。

[0014] 示例实施例通常针对用于分析从若干源接收的多个数据集以确定针对一个或多个问题的解决方案的数据解决方案系统。如在本文中所使用的那样，所接收的数据集可以指从各种社交媒体接收的数据集，数据集与产品的销售、围绕针对特定产品的市场营销活动所收集的市场营销数据等有关。

[0015] 图 1 是配置为从各种输入数据源接收多个数据集的数据解决方案系统的实施例的框图。数据解决方案系统 10 被配置为分析从各种源接收的数据集以为执行分析提供有引导的、交互式的白盒环境。将在下文中进一步详细地描述数据解决方案系统 10 的每个框。

[0016] 数据解决方案系统 10 被配置为与各种输入数据源 18、20 和 22 连接并且分别访问数据集 24、26 和 28。数据集的示例包括来自社交媒体、销售数字、市场营销渠道等的数据集。在一个实施例中，用户可以选择将从其获得数据集的输入数据源。如在本文中所使用的那样，术语“用户”可以指自然人以及作为“用户”进行操作的其他实体两者。示例包括公司、组织、企业，团队或者其他人群。还可以注意到的是，用户可以指受训练以对经由不同渠道接收的数据集执行数据分析的数据分析师。

[0017] 数据解决方案系统 10 包括图形用户界面 12，其被配置为使一个或多个用户能向分析模块 14 提供输入。在一个实施例中，图形用户界面包括使用户能选择所关心的选项的丰富菜单。

[0018] 分析模块 14 被配置为分析所接收的数据集以基于针对由用户定义的问题的详细统计分析生成最佳解决方案。这样的问题的示例可以包括确定产品销售的关键驱动力或者确定影响消费者的关键因素等。一般而言，分析模块 14 被配置为捕捉分析诀窍 (know-how) 并且以使执行过程有引导并且高效的方式设计工作流程。这又使用户能增加花费在形成深入了解 (insight) 上的时间。分析模块 14 还被配置为生成对分析数据集所执行的分析的可视化表示。

[0019] 存储模块 16 被配置为存储所述多个数据集和所述分析数据集。此外，存储模块 16 被配置为存储由分析模块生成的可视化表示。分析模块包括若干模块，将在下文中进一步详细地描述每个模块。

[0020] 图 2 是根据本发明的各个方面实现的分析模块的实施例的框图。如上所述，分析

模块 14 被配置为分析若干数据集并且生成使用户能确定针对由用户定义的目标的一个或多个解决方案的一个或多个数据模型。分析模块 14 包括实现若干统计过程以生成在做出关键商业决策时对用户有利的输出的多个模块。可以注意到的是,在下文中所描述的模块可以用户认为对于解决问题或达到目的所必需的任何顺序组合。将在下文中进一步地详细描述分析模块 14 的每个框。

[0021] 数据处理模块 30 被配置为将从多个源接收的多个数据集组合成分析数据集。分析数据集采用对于分析模块合适的格式。

[0022] 质量分析模块 32 被配置为确定分析数据集的属性。例如,唯一值设置、数据归档 (data profiling)、缺失或异常值处理及数据变换是质量分析模块所执行的功能中的一些。质量分析模块被配置为生成内容报告并且由此允许得出数据集中的所有变量的基本特征。

[0023] 探索性数据分析 (EDA) 模块 34 被配置为确定存在于分析数据集内的多个相关性。在一个实施例中,该多个相关性被用于确定所述一个或多个解决方案。EDA 模块 34 允许数据集操作、变量处理、数据汇总、数据探索以及数据处理。

[0024] 数据集操作允许在分析期间的任何阶段增加和输出数据集。该模块还允许跨数据集的变量进行数据分析。EDA 中的变量处理包括基于变量中的不同值 (distinct value) 重命名变量以及将变量归为数值、串 (string) 及手动分类。此外,其还包括新变量创建,所述新变量创建包括分类指示符、事件指示符、分仓 (binning)、广告库存变量 (ad stock variable)、滞后 / 超前变换、移动平均数等。

[0025] EDA 模块 34 的其他能力包括对于分析数据集的可视化表示的数据汇总、对变量中的唯一值的计数以及对于广泛的选项的统计汇总。延续地,数据探索也是 EDA 的关键支持能力中的一个。它支持可视化 (图表) 以及包括频率分析等的自定义模块。EDA 把数据看作为一元、多元、缺失、异常值以及变换处理。

[0026] 在一个实施例中,EDA 模块在分析数据集上实现一元和二元分析。在一个示例实施例中,通过一元分析对分析数据集执行定量 (统计) 分析。采用单个变量的描述及其适用的分析单元的属性来进行分析。一元分析允许像位置测量、分散度测量、正态性测试、分布、百分位值及它们的组合那样的属性。在另一示例实施例中,探索性分析模块被配置为对分析数据集应用多元分析。二元分析包括确定相对于一个或多个统计属性的变化。

[0027] 分析模块 14 还包括被配置为生成表示由用户规定的问题的一个或多个解决方案的一个或多个模型的数据建模模块 36。在一个实施例中,建模模块 36 使用回归技术来提供深入分析。在一个实施例中,基于分析数据集的平均值,方差以及协方差生成模型。数据建模模块被配置为支持多元处理、新变量创建和二元分析以研究自变量跨应变量的分布。

[0028] 诸如逐步变量消除、基于相关性及因素分析的变量分割等的模型构建选项可以被使用并且可以被构建在有偏差的总体上。其允许简单的变量消除通过多次迭代反复进行并且得到最佳适配的模型。其包括用于变量消除的算术回归并且还包含基于提前的 (advanced) 影响统计的多元异常值诊断。

[0029] 分析模块 14 还提供模型评估和验证能力。它基于模型统计、变量统计输出图表和表格。它对不同情景具有样本内 (in-sample) 和样本外 (out-of-sample) 验证以得到准确度和稳定性。可进行自助抽样 (bootstrapping) 以跨迭代比较模型统计。为多个冠军模型

提供评分并且比较输出的模型评分也被支持。

[0030] 报告模块 38 提供从单个用户界面对由分析模块生成的所有报告的容易的访问。报告类型的示例包括跨多个分类变量的所有不同级的内容报告、频率报告、一元汇总报告、多元汇总报告等。此外,可以生成对于不同变量和选项的多个报告并且可以直接将这些报告输出为诸如 excel、pdf 等的格式。

[0031] 报告模块确保所有输出在一个地方被比对以供用户更好地形成深入了解。可以在报告框架中的一个位置查看不同的报告并且还可以计算结果比较。可以容易地跨报告比较结果。深入了解的形成是这种能力的另一特征。可以使用报告框架快速地形成深入了解并且使该深入了解容易地与商业逻辑相关联。

[0032] 图 3 是示出根据本发明的各个方面的、来自不同源的各种数据集根据其被处理的一种方法的流程图。如上所述,不同数据集指来自销售、市场营销、社交媒体数据集等的数据集。将在下文中进一步详细地描述用于分析社交媒体数据的过程 40。

[0033] 在步骤 42 中,从一个或多个输入数据源取回数据集。从若干源接收的数据集被分析以确定针对特定问题的解决方案。一般而言,输入数据集可以包括某种产品的关键字、产品名称、企业或组织的名称等。在一个实施例中,数据集包括文本串和数值数据。

[0034] 在步骤 44 中,所接收的数据集被调节以生成分析数据集。数据处理被执行以通过应用某些条件来创建新变量。还可以通过处理现有数据集来生成新的数据集。

[0035] 在一个示例实施例中,对数据集执行一元处理。一元处理涉及选择递增或递减操作以及变量需要改变的特定值。在另一示例实施例中,对数据集执行二元处理。通过选择对于两个或多个变量的操作并且将操作值指派给新变量来执行二元处理。

[0036] 在步骤 46 中,分析数据集的质量被评估。质量评估需要标识对于操作重要的维并且需要精确定义构成这些维的变量。被用于质量评估的示例因素是准确性、完整性、一致性和时间性。

[0037] 在步骤 48 中,分割模块基于属性将分析数据集分组,其中所述属性由用户使用图形用户界面来选择。

[0038] 在步骤 50 中,对分析数据集执行探索性数据分析。探索性数据分析确定存在于所述分析数据集内的多个相关性,所述多个相关性协助确定针对用户定义的问题的一个或多个解决方案。探索性数据分析允许诸如一元分析、二元分析、基本及高级可视化、交叉表分析、频率及性能分析、相关性以及时间序列的多重分析。

[0039] 在步骤 52 中,数据模型被生成以确定一个或多个解决方案。数据建模提供回归技术的深入分析并且包括预建模处理。在步骤 54 中,资料库允许对在数据处理、质量分析、探索性数据分析以及数据模型生成步骤期间生成的所有报告的访问。

[0040] 上述技术可以由图 1 和图 2 所描述的数据分析系统执行。上述技术可以被实施为设备、系统、方法和 / 或计算机程序产品。相应地,上述内容中的一些或全部可以用硬件和 / 或用软件 (包括固件、常驻软件、微代码、状态机、门阵列等) 来实施。此外,所述内容可以采用在计算机可用或计算机可读存储介质上的诸如分析工具的计算机程序产品的形式,所述计算机可用或计算机可读存储介质具有用该介质实施以供指令执行系统使用或与指令执行系统连接的计算机可用或计算机可读程序代码。在本说明书的上下文中,计算机可用或计算机可读介质可以是能够包含、存储、通信、传播或传送供指令执行系统、装置或设备

使用或与其连接的程序的任何介质。

[0041] 计算机可用或计算机可读介质可以例如是,但不限于电子、磁、光学、电磁、红外或半导体系统、装置、设备或传播介质。通过示例的方式而不是限制,计算机可读介质可以包括计算机存储介质和通信介质。

[0042] 当所述内容在计算机可执行指令的一般上下文中实施时,实施例可以包括由一个或多个系统、计算机或其他设备执行的程序模块。一般而言,程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等。典型地,可以如在各种实施例中有所期望的那样组合或分布程序模块的功能。

[0043] 图 4 是示出配置为对于从各种源取回的数据集生成针对特定问题的数据解决方案的计算机 100 的实施例的框图。计算机 100 被配置为执行用于执行图 3 所描述的步骤的数据解决方案工具的指令。在非常基本的配置 102 中,计算机 100 典型地包括一个或多个处理器 104 以及系统存储器 106。存储总线 124 可被用于处理器 104 与系统存储器 106 之间的通信。

[0044] 取决于所期望的配置,处理器 104 可以是任何类型的,包括但不限于微处理器(μ P)、微控制器(μ C)、数字信号处理器(DSP)或它们的任何组合。处理器 104 可以包括一个或多个级别的高速缓存,诸如一级缓存 110 和二级缓存 112、处理器内核 114 以及寄存器 116。示例处理器内核 114 可以包括算术逻辑单元(ALU)、浮点单元(FPU)、数字信号处理内核(DSP 内核)或它们的任意组合。示例存储控制器 118 也可与处理器 104 一起被使用,或者在一些实现中,存储控制器 118 可以是处理器 104 的内部部件。

[0045] 取决于所期望的配置,系统存储器 106 可以是任何类型的,包括但不限于易失性存储器(诸如 RAM)、非易失性存储器(诸如 ROM、闪存等)或它们的任意组合。系统存储器 106 可以包括操作系统 120、一个或多个应用程序 122 以及程序数据 124。应用程序 122 包括被布置用于分析从不同源接收的多个数据集的数据解决方案工具 120。程序数据 126 可以包括社交媒体数据、市场营销数据、销售数据等。在一些实施例中,应用程序 122 可以被布置用于与程序数据 126 一起在操作系统 120 上进行操作,使得分散的设备与外部实体之间的交互被监控。这种所描述的基本配置 102 在图 4 中由内虚线内的那些组件示出。

[0046] 计算机 100 可以具有附加的特征或功能,以及具有附加的接口来促进基本配置 102 与任何所需要的设备和接口之间的通信。例如,总线/接口控制器 130 可被用于促进基本配置 102 与一个或多个数据存储设备 132 之间经由存储接口总线 138 的通信。数据存储设备 132 可以是可移动存储设备 134、非可移动存储设备 136 或它们的组合。可移动存储装置和非可移动存储装置的示例包括诸如软盘驱动和硬盘驱动(HDD)的磁盘设备、诸如压缩光盘(CD)驱动或数字多功能光盘(DVD)驱动的光盘驱动、固态驱动(SSD)以及磁带驱动等等。示例计算机存储介质可以包括以用于信息存储的任何方法或技术实现的易失性和非易失性、可移动和非可移动介质,所述方法或技术诸如为计算机可读指令、数据结构、程序模块或其他数据。

[0047] 系统存储器 106、可移动存储设备 134 以及非可移动存储设备 136 是计算机存储介质的示例。计算机存储介质包括但不限于 RAM、ROM、EEPROM、闪存或其他存储技术、CD-ROM、数字多功能光盘(DVD)或其他光存储装置、盒式磁带、磁带、磁盘存储装置或其他磁存储设备或者可以被用于存储所期望的信息并且可由计算机 100 访问的任何其他介质。任何这样

的计算机存储介质都可以是计算机 100 的部分。

[0048] 计算机 100 还可以包括用于促进从各种接口设备（例如输出设备 140、外围接口 148 以及通信设备 160）到基本配置 102 的、经由总线 / 接口控制器 130 的通信的接口总线 138。示例输出设备 142 包括图形处理单元 144 和音频处理单元 146，其可以被配置为经由一个或多个 A/V 端口 142 与诸如显示器或扬声器的各种外部设备通信。示例外围接口 148 包括串行接口控制器 150 或并行接口控制器 152，其可被配置为经由一个或多个 I/O 端口 148 与诸如输入设备（例如键盘、鼠标、笔、语音输入设备、触摸输入设备等）的外部设备或其他外围设备（例如打印机、扫描仪等）通信。示例通信设备 160 包括网络控制器 154，其可以被布置用于促进经由一个或多个通信端口 156 在网络通信链路上与一个或多个其他计算机 158 的通信。

[0049] 网络通信链路可以是通信介质的一个示例。通信介质可以典型地由计算机可读指令、数据结构、程序模块或诸如载波或其他传输机制的、经调制数据信号中的其他数据来实施，并且可以包括任何信息传递介质。“经调制数据信号”可以是具有其特征集中的一个或多个或者以将信息编码在信号中这样的方式改变的信号。通过示例的方式而不是限制，通信介质可以包括诸如有线网络或直接连线连接的有线介质，以及诸如声音、无线电频率 (RF)、微波，红外 (IR) 的无线介质及其他无线介质。如在本文中所使用的那样，术语计算机可读介质可以包括存储介质和通信介质两者。

[0050] 计算机 100 可以被实现为诸如手机、个人数据助理 (PDA)、个人媒体播放器设备、无线网页观看设备、个人耳机设备、专用设备或包括任何以上功能的混合设备的、外形因素 (form factor) 小的便携（或移动）电子设备的一部分。计算机 100 还可以被实现为包括膝上型计算机和非膝上型计算机配置两者的个人计算机。如上所述，数据分析工具及系统被配置为分析从社交媒体平台取回的社交媒体数据。数据解决方案工具及系统可以包括图形用户界面以促进用户提供输入数据以及选择由数据解决方案系统所提供的、所需要的操作。将在下文中参考图 5 至图 12 描述一些示例用户界面屏幕。

[0051] 图 5 是使诸如数据分析师的用户能对数据集执行数据处理操作以生成分析数据集的图形用户界面的截屏。数据处理模块使数据分析师能增加新变量或处理现有数据集，如屏幕 56 所示。数据分析师还可以为数据集选择共有的和专有的变量并且生成验证结果。数据分析师还可以从数据处理操作中生成相关的报告。

[0052] 图 6 是对分析数据集的数据质量分析的可视化表示的截屏。如可以清楚地看到的那样，在截图 58 中，质量分析支持通过一元汇总的定量（统计）分析。一元汇总每次允许多个变量的属性，像位置测量、分散度测量、正态性测试、分布、百分位值以及它们的组合。

[0053] 图 7 是对分析数据集的探索性数据分析的可视化表示的截屏。截屏 60 示出以不同制图类型的形式表示的分析数据集的一元分析，诸如概率图、盒型图、自相关图、直方图、平均百分位图和标准偏差图。

[0054] 类似地，图 8 的截屏 62 示出对于给定数据集的不同变量的频率及性能分析。图形用户界面允许数据分析师选择诸如频率、频率百分比、差别计数 (distinct count)、平均值等的各种参数以曲线图或表格格式来可视化。图 9 示出在多次迭代中对数据集的时间序列分析。截屏 64 描绘了对于单次迭代的时间序列图。

[0055] 图 10 是允许生成表示探索性数据分析的分析结果的一个或多个模型的数据建模

的截屏。如可以看到的那样,在图 11 的截屏 66 和 68 中,数据建模允许模型定义、模型构建、模型的模型诊断及可视化历史在诸如线性回归、逻辑回归、VARMAX、ARIMAX 等的类别下的可能性。在模型构建期间,基于分析数据集的平均值,方差以及协方差生成一个或多个模型。图 12 是各种报告和图表或曲线图的截屏 70,诸如内容报告、平均销售报告的所述报告和图表或曲线图可基于由数据分析师在各个阶段完成的数据分析生成。

[0056] 本领域的技术人员将理解,一般而言,在本文中并且尤其是在所附权利要求(例如所附权利要求的主体)中所使用的术语通常意在作为“开放式”术语(例如术语“包括”应被解释为“包括但不限于”,术语“具有”应被解释为“至少具有”,术语“包括”应被解释为“包括但不限于”等)。本领域的技术人员还将理解的是如果想要所引入的权利要求列举(recitation)有特定数量,则这样的意图将在权利要求中明确地叙述,并且在没有这样的列举的情况下则不存在这样的意图。

[0057] 例如,为了帮助理解,以下所附的权利要求可包含对介绍性短语“至少一个”和“一个或多个”的使用来引入权利要求列举。然而,即使在同一权利要求包括介绍性短语“一个或多个”或“至少一个”和诸如“一”或“一个”的不定冠词的情况下(例如“一”和/或“一个”应被解释为表示“至少一个”或“一个或多个”),这样的短语的使用不应被理解成暗示由不定冠词“一”或“一个”对权利要求列举的引入将包含这样引入的权利要求列举的任何特定权利要求限制于仅包含一个这样的列举的实施例;对于定冠词被用于引入权利要求列举的使用同样如此。另外,即使所引入的权利要求列举的特定数量明确地被叙述,本领域的技术人员将认识到这样的列举应被理解成表示至少所叙述的数量(例如在没有其他修饰语的情况下对“两个列举”的单纯叙述表示至少两个列举或者两个或更多个列举)。

[0058] 尽管在本文中仅示出和描述了若干实施例的某些特征,本领域的技术人员将想到许多修改和变化。因此,应理解的是所附权利要求旨在覆盖落在本发明的实质精神之内的所有这样的修改和变化。

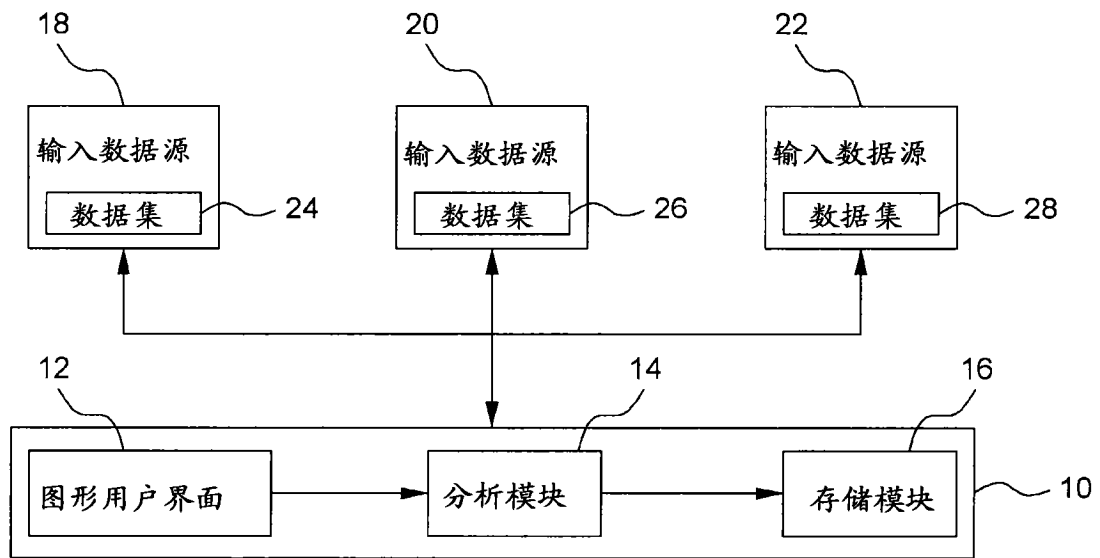


图 1

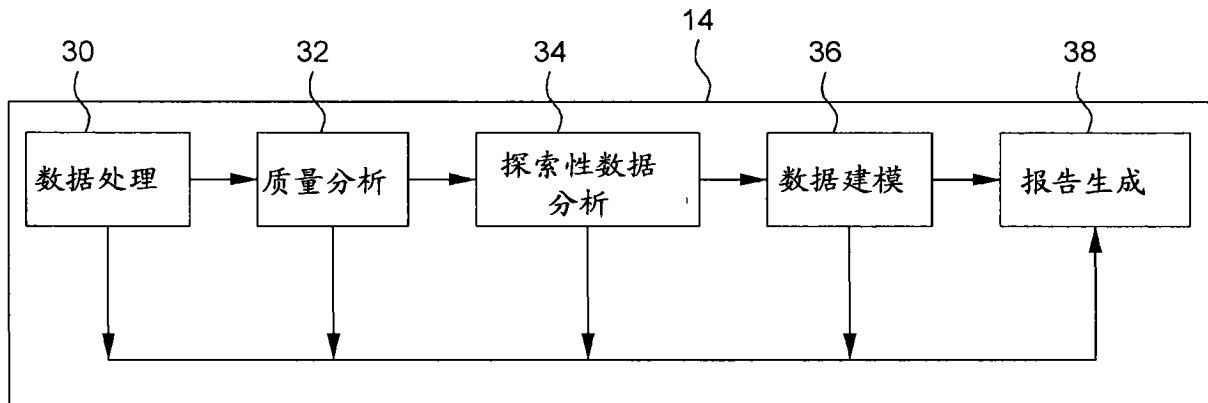


图 2

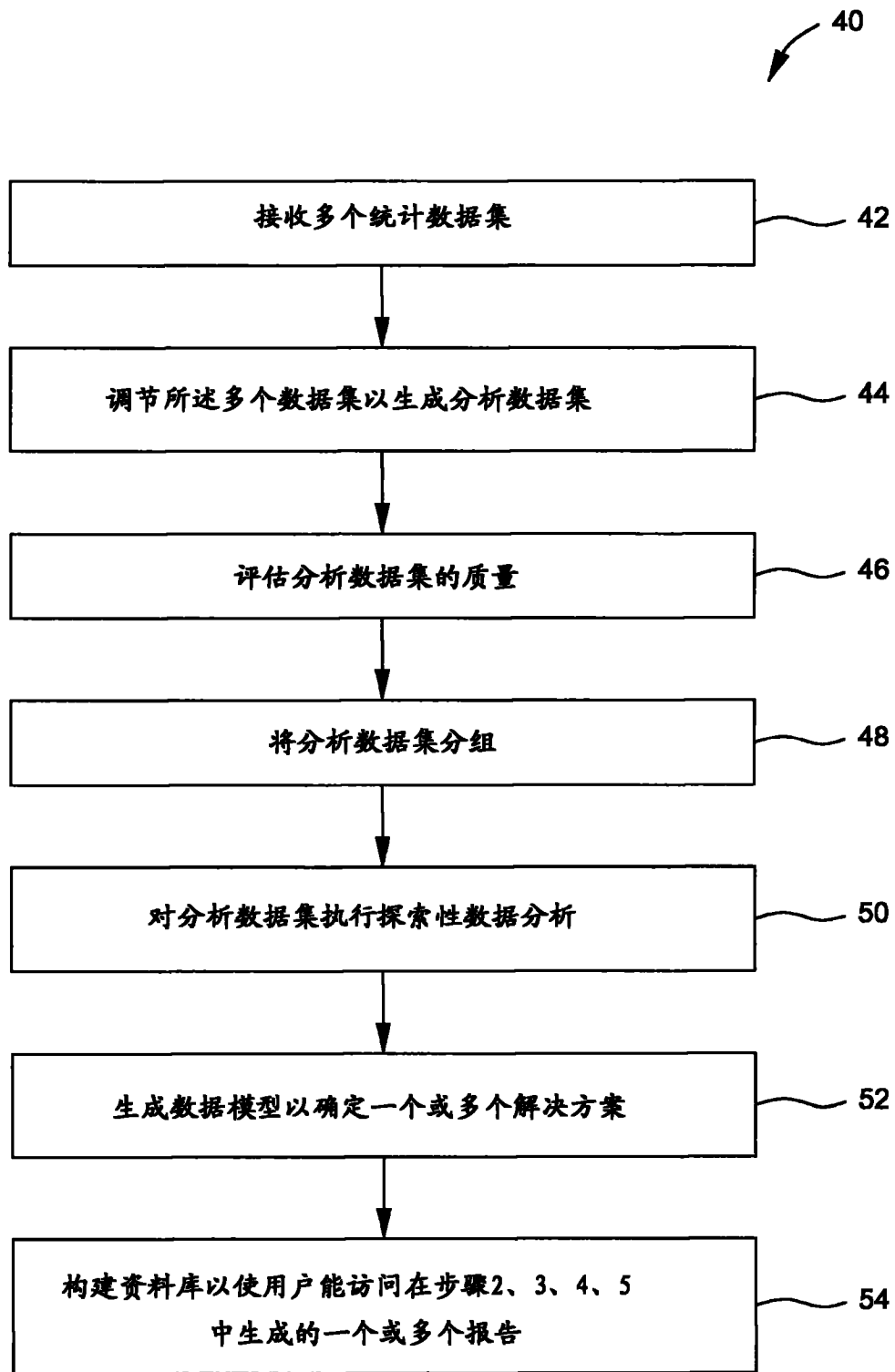


图 3

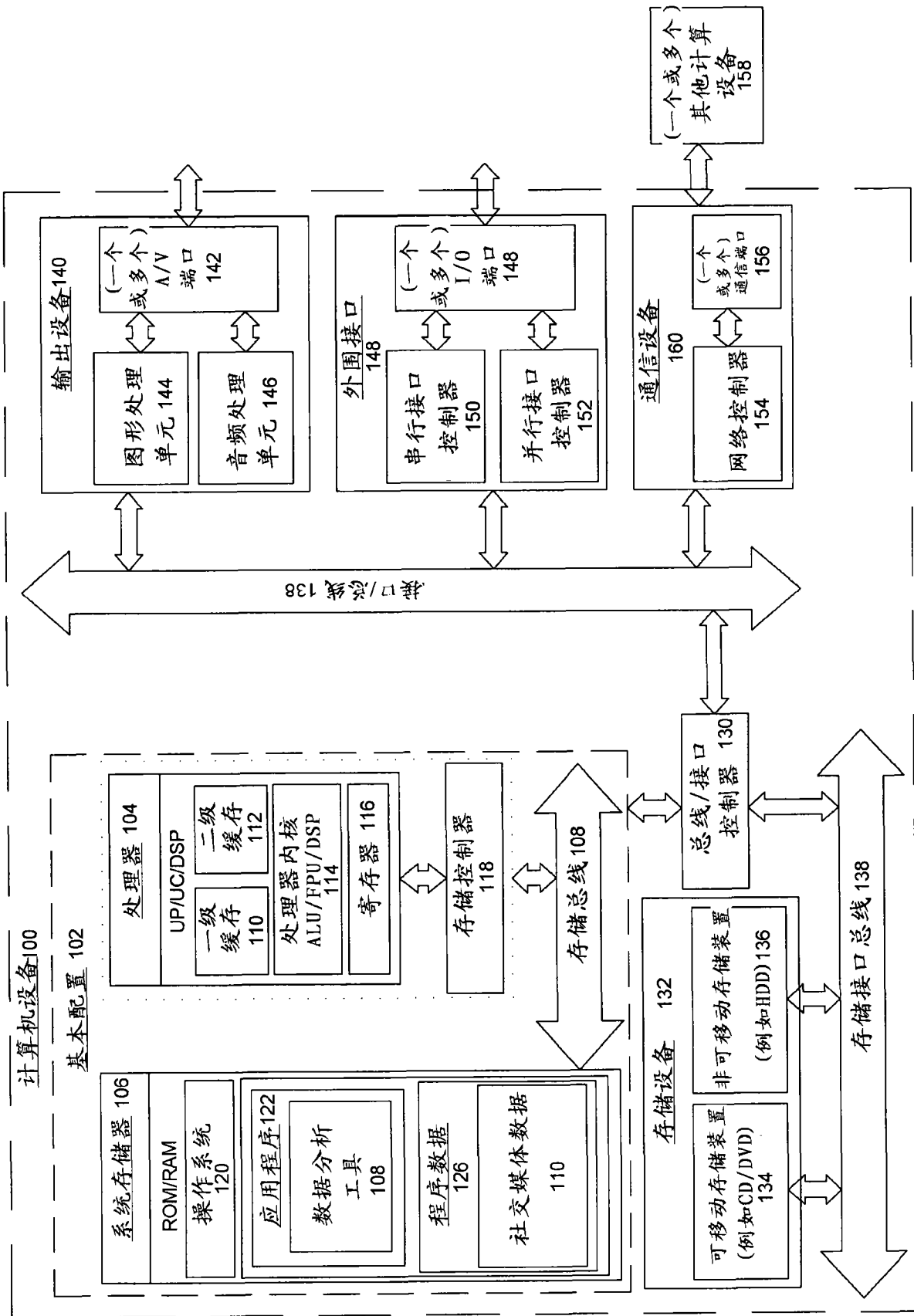


图 4

56

muPDNA
项目设置
数据处理
数据质量分析
EDA
建模
分割
报告框架

新变量创建
数据集处理

操作
报告

操作
附加

数据集列表

Base Da..	Child Da..	数据集	名称
<input checked="" type="radio"/>	<input type="checkbox"/>	<input type="checkbox"/>	DATA WOR
<input type="radio"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	DATA WOR

数据集视图 选择

专有变量
 DATAWORKING_LIN..
 DATAWORKING_NE..
 date_11
 date_date9
 date_datetime16
 date_ddmmyy6
 date_monyy5
 date_yymmdd8
 miles_acv

共有变量

<input checked="" type="checkbox"/>	ACV
<input checked="" type="checkbox"/>	Chiller_flag
<input checked="" type="checkbox"/>	Date
<input checked="" type="checkbox"/>	HHe_55_64
<input checked="" type="checkbox"/>	HHe_Income_75K..
<input checked="" type="checkbox"/>	Hispanic_HHe_Ind..
<input checked="" type="checkbox"/>	P164_Demand_In..
<input checked="" type="checkbox"/>	P28_Demand_Ind..
<input checked="" type="checkbox"/>	Store_Format
<input checked="" type="checkbox"/>	Total_Selling_Area

验证结果

变量	类型	类型	分派	是否	选择
ACV	数值	否	否	是	<input checked="" type="checkbox"/>
Chiller_f	数值	否	否	是	<input checked="" type="checkbox"/>
Date	数值	否	否	是	<input checked="" type="checkbox"/>
HHe_55	数值	否	否	是	<input checked="" type="checkbox"/>
HHe_Ind	数值	否	否	是	<input checked="" type="checkbox"/>
Hispanic	数值	否	否	是	<input checked="" type="checkbox"/>
P164_D	数值	否	否	是	<input checked="" type="checkbox"/>

历史

报告名称	基本数据集删除

新数据集名称:

图 5

58

主页	MPDNA	应用设置	数据操作	数据质量分析	EDA	线性回归	逻辑回归	因素分析	报告框架
▶ QA报告生成		QA报告		变量的唯一值					

数据集: 数据工作

数值变量	定义
ACV	竞争者指数
Chiller_flag	用于总体的标志变量
HHS_55_64	西班牙裔群体指数
HHS_Income_75K..	未添加定义
Hispanic_HHS_Ind..	未添加定义
P164_Demand_In..	需求指数
P26_Demand_Ind..	未添加定义
Total_Selling_Area	总销售区
black_hispanic	黑人西班牙裔群体
channel_1	channel indicator for first c..
channel_2	channel indicator for second...
channel_3	channel indicator for third ...

一元汇总的变量	变量
ACV	
Hispanic_HHS_Index	
P26_Demand_Ind..	
Total_Selling_Area	
sales	

一元汇总的选项	位置测量
<input checked="" type="checkbox"/> 平均值	<input checked="" type="checkbox"/> 平均值
<input checked="" type="checkbox"/> 中间值	<input checked="" type="checkbox"/> 中间值
<input checked="" type="checkbox"/> 模式	<input checked="" type="checkbox"/> 模式
<input type="checkbox"/> 中间平均值	<input type="checkbox"/> 中间平均值
<input type="checkbox"/> 截尾平均值	<input type="checkbox"/> 截尾平均值
<input type="checkbox"/> 缩尾平均值	<input type="checkbox"/> 缩尾平均值
<input type="button" value="全选"/> <input type="button" value="全不选"/>	
分散度测量	分散度测量
正态性测试	正态性测试
分布	分布
<input checked="" type="checkbox"/> 百分位值	增加: <input type="text" value="0"/>
<input type="checkbox"/> 相关性矩阵	<input type="checkbox"/> 相关性矩阵

附加报告生成选项	
<input checked="" type="checkbox"/> 生成按面板的一元汇总	
分散度测量	分散度测量
Chiller_flag	Chiller_flag
Store_format	
Channel_1	
Channel_2	
Channel_3	
通用报告	
<input checked="" type="checkbox"/> 选择生成内容报告 <input type="button" value="生成报告"/>	
查看所生成的报告	查看所生成的报告
报告名称	状态
rep1	已完成
rep2	已完成
rep3	已完成
	数据集名称
	DATAWORKING
	DATAWORKING
	DATAWORKING
<input type="button" value="查看QC报告"/>	

60

主页 MPDNA 应用设置 数据操作 数据质量分析 EDA 线性回归 逻辑回归 因素分析 报告框架

一元分析 二元分析 可视化 变量处理 交叉表 特性 频率 相关性

选择 报告

报告 rep1
编辑报告 删除报告

信息: 跨面板布局 | DATAWORKING

变量选择

Total_Selling_Area

南部

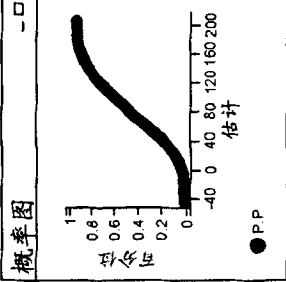
北部

Sales

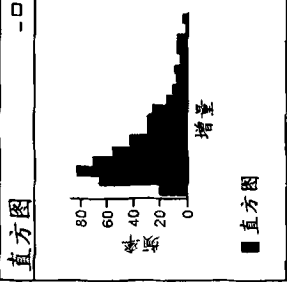
北部

南部

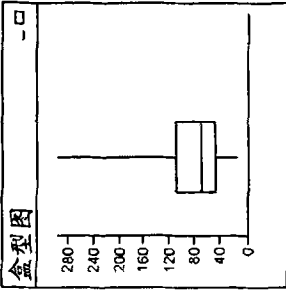
统计	估计
平均值	85.4717992
标准偏差	52.16561
number_of_missing	0
observations_used	463
最大值	296.740741
中间值	70.6204878
最小值	10.29
范围	286.450741
lqr	61.5456296
模式	74.2275
lb	-136.98289
ub	293.836518
no_of_zeros	463
noofoutliers	463



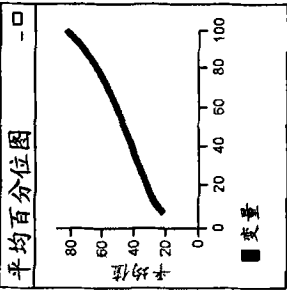
概率图



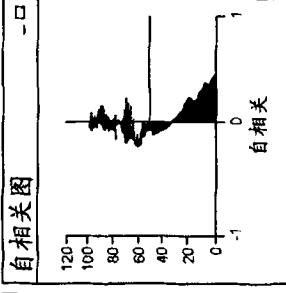
直方图



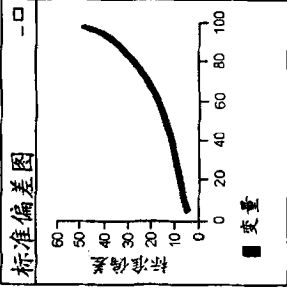
盒型图



平均百分位图



自相关图



标准偏差图

图 7

62

主页	MPDNA	应用设置	数据操作	数据质量分析	EDA	线性回归	逻辑回归	因素分析	报告框架		
▶ 一元分析		二元分析		高级可视化		交叉表		频率及特性分析		相关性	时间序列
报告生成											
报告查看											

报告: 测试

数据集: DATAWORKING_LINEAR | 频率选项: 唯一值 总计 | 特性选项: 平均值 中间值 范围 标准偏差 | 分析类型: 面板变量 | 选项: 导出到Excel

布局

变量

- continuous
- channel_3
- channel_2
- channel_1
- channel_4

频率
 频率百分比
 累积频率
 累积百分比
 平均值
 标准偏差
 直方图
 计数差别
 累积频率百分比
 百分比差别

图 8

64

数据集: DATAWORKING LINEAR

单一模式 多变量选择

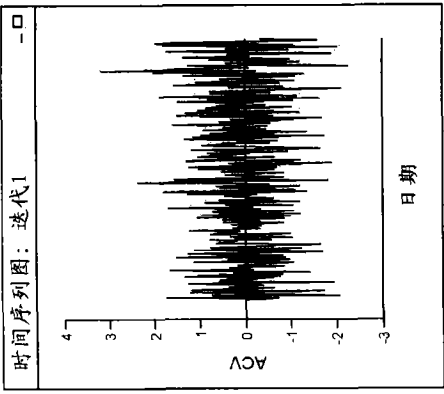
基本迭代: 0

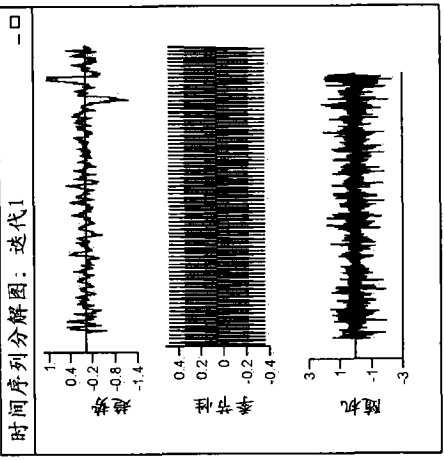
全选 / 全不选
 时间序列图
 时间序列分解图
 自相关图
 部分自相关图
 单位检验根
 白噪声测试

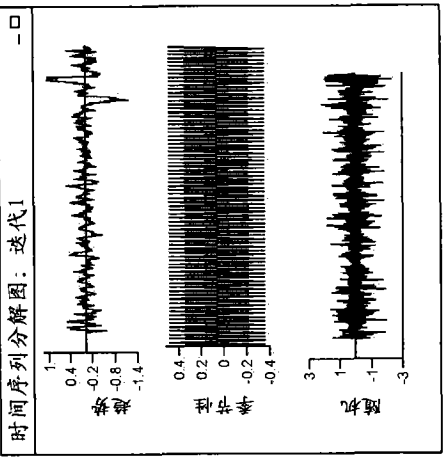
变换
 对数变换

按季节性调整数据
 创建季节性指数

求差: 算术 0 1
 标准 顺序 0 1
 季节性顺序 0 1

时间序列图: 迭代1


相关图: 迭代1


测试图: 迭代1


信息
 Time Series: ACV
 Transform...: true
 Seasonali...: false
 Order of di...: 3
 Order of s...: 5

66

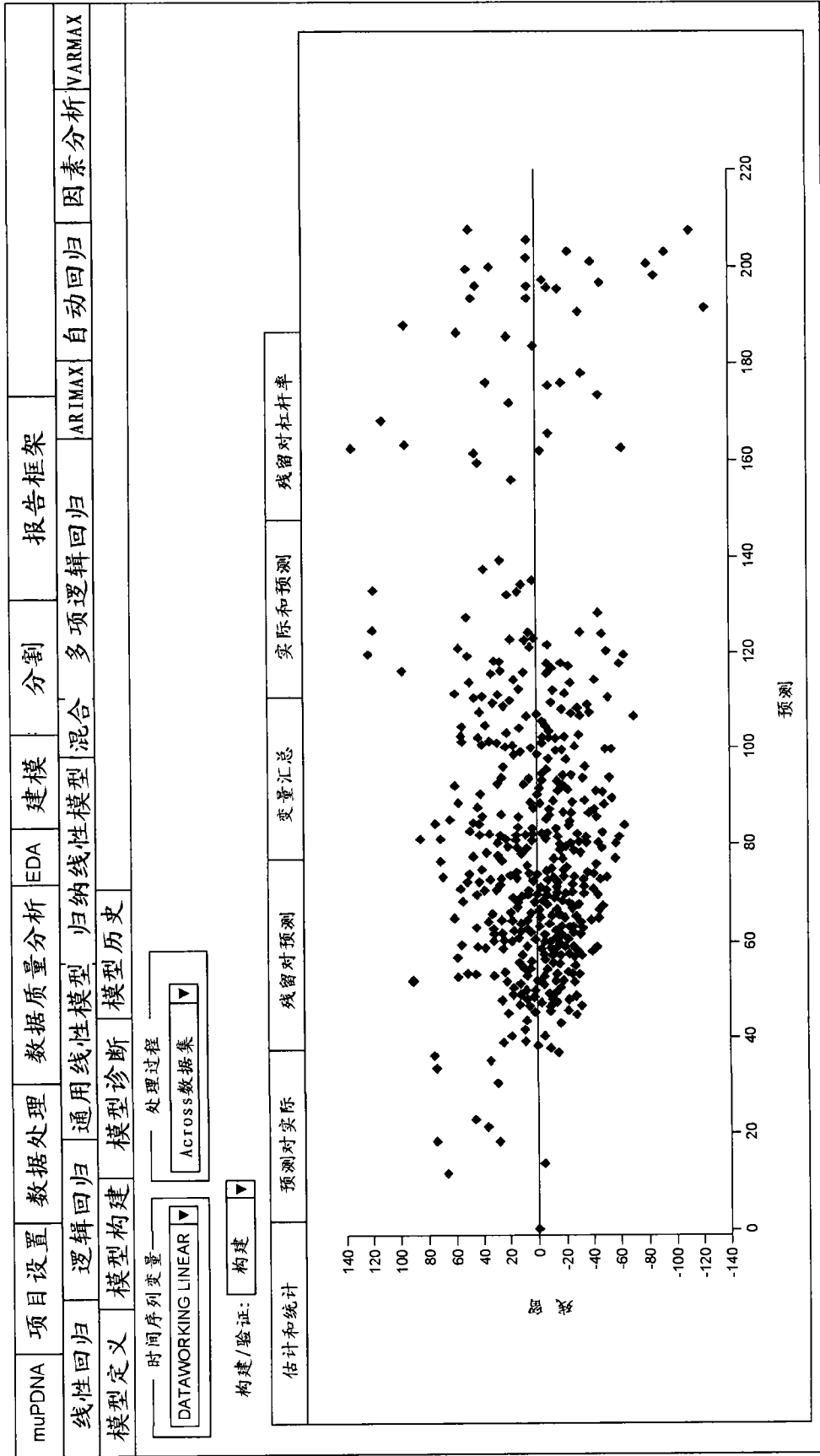


图 10

68

muPDNA

项目设置

数据处理

数据质量分析

EDA

建模

分割

报告框架

线性回归

逻辑回归

通用线性模型

回归线性模型

混合

多项逻辑回归

ARIMAX

自动回归

因素分析

VARMAX

模型定义

模型建立

模型诊断

模型历史

时间序列变量
DATAWORKING LINEAR

处理过程
ACROSS数据集

应变量
Sales, Total_Sel...

迭代
1

响应变量:
sales

选择输出
 构建
 验证

预测对实际
一元模型诊断

残留对实际
残留分析

部分自动回归

模型构建

模型历史

模型统计	
统计	Total S... Total S...
AIC	26.00000 26.00000
AICC	26.00000 26.00000
Covariance of Innovation Total	2409474. 205476 9.
Covariance of Innovation Seri...	205476. 1716.282.
FPEC	3.80000. 3.80000.
HQC	26.00000 26.00000
MAPE	21.65933 54.14059
最大值	233000. 296.74074
平均值	52118.5. 84.15001.
最小值	18000.0. 10.29000.
Number of Nonmissing Obser...	523.0000. 523.0000.
Number of Observations	523.0000. 523.0000.
Number of Pairwise Missing	0.00000 0.00000
SBC	26.00000 26.00000
标准偏差	34871.2. 50.41435.
模型的类型	VARX(3_2) VARX(3_2)
估计方法	Least S... Least Sq...

图 11

70

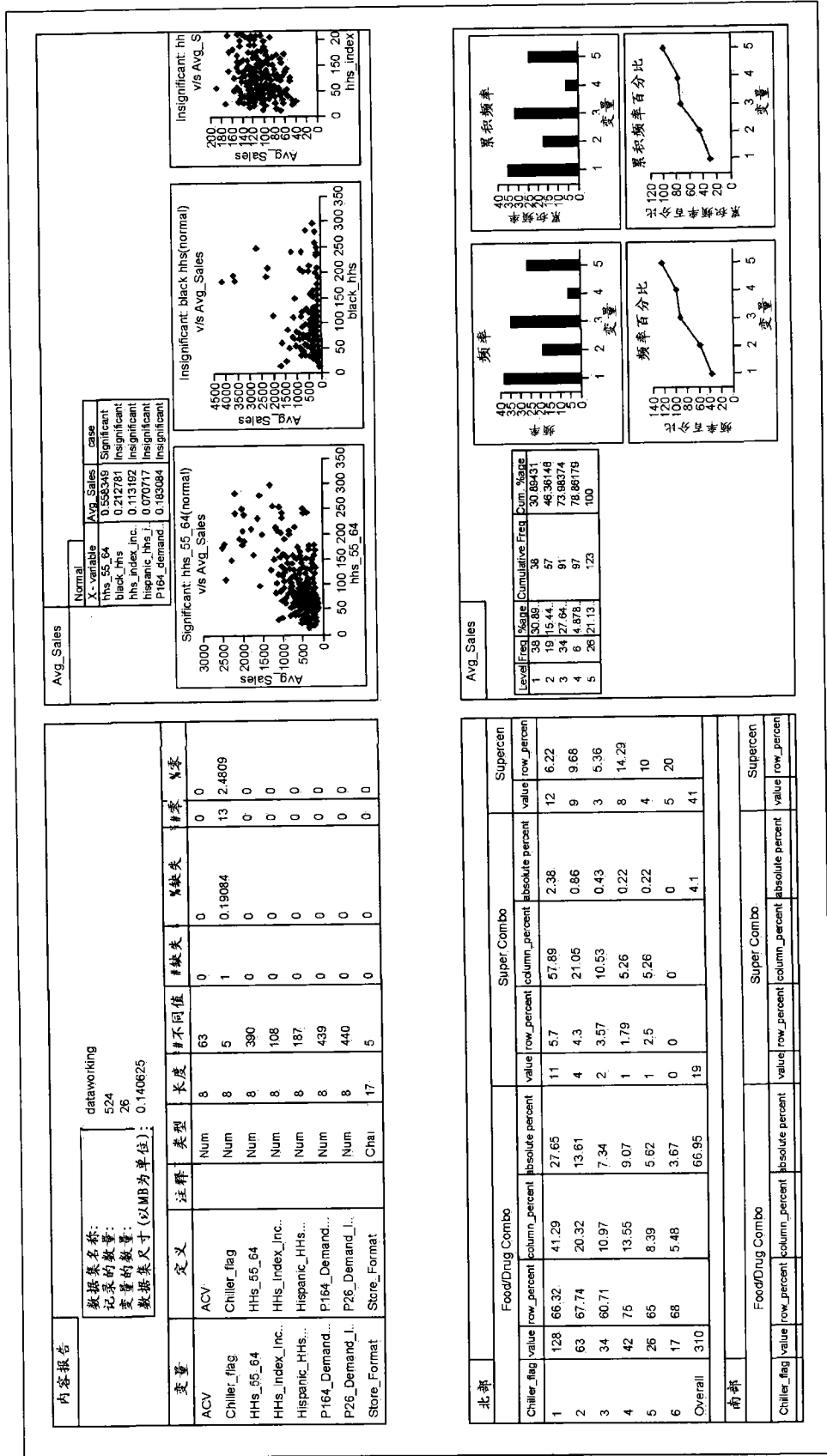


图 12