文章编号：1005-8451（2019）9-0049-05

动车组大数据处理及分析

严　皓
（中国铁路成都局集团有限公司　成都动车段，成都　610051）

摘　要：将动车组运用维护过程中产生的海量数据进行处理与分析，解决数据不能转换为可用结果的问题。依托大数据分析软件Tableau、按照大数据分析的流程进行处理及分析，实现了闸片磨耗率、磨耗差异、偏磨量等多角度展示，同时，结合地图与GPS坐标对故障分布进行了标记。分析结果表明，大数据处理技术可提高数据分析的效率及准确性。

关键词：数据清洗；线性回归；磨耗率
中图分类号：U266.2 : TP39　**文献标识码：**A

Large data processing and analysis of EMU

YAN Hao
(Chengdu EMU Depot, China Railway Chengdu Group Co. Ltd., Chengdu　610051, China)

Abstract: This article processed and analyzed the massive data generated during the EMU operation and maintenance, and solved the problem that data can not be converted into available results.Relying on the big data analysis software Tableau and according to the process of big data analysis, the brake wear rate, wear difference, partial wear and other aspects were displayed. At the same time, the fault distribution was marked with the map and GPS coordinates.The analysis results show that the big data processing technology can improve the efficiency and accuracy of data analysis.

Keywords: data cleaning; linear regression; wear rate

随着动车运用检修工作的不断进步，传统计划预防修模式引发的“过度修”等问题日渐暴露，以大数据为核心的“可靠性”状态修模式应运而生，张春[1]等人从理论的角度基于动车组等故障概率研究了状态维修建模体系，吕晓艳[2]等人结合生产运用中的大数据对运输组织及管理进行了优化，本文结合状态修理论以及现场大数据的应用，从动车运用工作中实际问题出发，研究相应部件磨耗、老化或故障规律，探索可靠性维修体系，确定合理的检修方式、检修类型、检修间隔周期等，从而争取以最少的维修资源消耗确定合理有效的检修方针。

1　动车组大数据分析

动车组大数据分析流程[3]概括如下。

1.1　需求分析评估

数据分析的本质是服务需求，如果缺乏专业可支撑的分析流程、不能以解决现场实际问题为导向，盲目地进行数据规律统计分析，会致分析工作不能有效转换为可用结果。

1.2　数据收集

动车组大数据收集必须立足于准确、完整反映分析问题的本身。尽可能从更规范的数据库、更完整的信息源中进行提取，围绕动车组方面，建议遵循“优先选取设备检测数据、过程数据大于结果数据”的原则；此外，同一问题的信息往往分散在多个数据源中，视情况需要，可按相关程度由重及轻分步收集。

1.3　数据清洗

数据清洗[4]是数据分析前期准备工作中非常关键的环节。如果不对错误数据、坏数据、缺失数据、不规范数据等进行处理，可能会致后期无法得到合理的结果甚至没有结果。

1.3.1　数据规范

将相同属性的字段按统一命名规则进行整理。例如，CRH1A-1021 与 CRH1021A，表示相同的意思，按统一命名规则将其规范表达为 CRH1A-1021。

1.3.2　异常值筛选

异常值筛选主要从两方面进行甄别。

收稿日期：2018-11-26
基金项目：中国铁路成都局集团有限公司科技研究开发计划项目（2016CX1639）
作者简介：严　皓，工程师。

（1）逻辑判断，即从专业基础知识角度将错误数据、坏数据等剔除或更正，例如，CRH1A-790 型闸片厚度 38 mm，因为闸片厚度 38 mm 超出出厂标准值，该数据不存在，因此需要将其剔除。

（2）统计判断，通过置信概率判断某些值严重超出误差范畴，去除不合理的数据，例如，拖车闸片磨耗周期为 1 000 ~ 2 000 km 或超过 200 万 km 等，为不合理数据，需要去除）。

1.4 缺失值补充

数据清洗过程中，会导致部分数据被整条删除或关键字段被清理，造成数据样本减少，过少时会对分析结果带来重大影响，需要采取缺失值处理。缺失值补充主要有两种方式。

（1）利用同类信息进行相似匹配（例如，两个数据库中均存在同一车组当日闸片更换信息，可进行信息互补）。

（2）利用数学方式进行补充，常见的有线性图形补充、平均值补充、回归方程补充等。

1.5 数据初步分析

（1）按照问题导向的原则确定关键字段，确定行列间的度量和维度，选取能反应问题特征的图形作为数据展示方式，如温升折线图、故障时间散点图、问题总量柱状图等。

（2）分析图形 / 数据的基本特征，如分散程度、中心位置、均值、标准差值等，得出统计数据的可参考度，统计最终结果和分布规律。

1.6 数据建模及相关性分析

当对数据本身特性有足够的掌握、且分析深入性和系统性程度较高时，可结合专业知识特性对某个部件 / 系统进行建模分析，选取相关的参数，确定合适的算法（如用于分类的 KNN 算法、线性回归法），描绘对应展示图形 / 曲线，分析部件 / 系统的运行平稳性，摸索各关键参数间的相关性。

2 数据预处理

在动车组大数据分析的基础上，结合 Tableau 大数据分析软件 [5]，以配属动车组磨耗件为分析对象，按照上述分析流程进行数据预处理。

2.1 分析需求

通过分析配属动车组磨耗件的磨耗规律，掌握动车组磨耗件平均更换周期，以经济性及安全性为评价指标，为一级、二级修中磨耗件检修限度的制定提供依据，为物料采购、合理化库存提供基础数据。

2.2 数据收集

按照“优先选取设备检测数据、过程数据大于结果数据”的原则，本次数据收集选取了轮对检测棚的 LVZ 闸片检测数据、LY 滑板检测数据、二维码磨耗件数据、动车组信息管理系统中故障管理数据、现场日生产信息等。

2.3 数据清洗

在数据清洗过程中，发现明显问题数据 1 641 条、占总故障数据比例约 20%，由于数据间存在先后的关联性，故障数据对整体的影响大于 20%，主要问题如下。

（1）数据逻辑错误。轴承位置与端位数明显不匹配，例如轴承数超过 4 个、不带轴盘制动的动车端位数超过 16 等。

（2）信息不完整。例如故障部件无车厢号、车厢号为全列、无具体端位数、车组走行公里值为 0 或为 null。

2.4 缺失值补充

针对清洗后出现信息不完整的情况，采取两种措施进行缺失值补充。

（1）用动车组故障信息中相同数据进行特征匹配，补充位置信息。

（2）通过 Tableau 的双表融合功能，用动车组交路信息进行走行公里补充。通过系列数据清洗、补充方式，最终从 8 305 条原始数据中整理出 8 062 条可用数据，约有 3% 的数据被清理，确保了大数据的完整性。

2.5 建立线性回归分析方程

此处考虑动车组在运行过程中，闸片磨耗过程电制动与空气制动功能均处于正常状态 [6] 且运行交路基本不变，闸片厚度与走行里程呈线性关系，借助 Tableau 工具，以时间 / 走行里程为维度，以车组、车号、闸片位数等为度量，建立线性回归分析方程，如图 1 所示。图中，每个夹钳左右两侧闸片分别以红、绿两色标识，其中，红色曲线代表左侧、绿色

曲线代表右侧，纵坐标为闸片厚度（单位 mm）、横坐标为走行公里（单位 km）。筛选同一夹钳左右两侧闸片磨耗趋势相近进行线性拟合，图中的方框显示了 CRH3A-3092 04 车 5 位左、右闸片的磨耗趋势，模型方程为：

闸片厚度 =–4.702 92e^{-5} × 走行公里 + 26.837 2 (1)

式（1）中，–4.702 92e^{-5} 为斜率，26.837 2 为截距。与此同时，自动计算出 R 平方值（测定系数）和 P 值（显著性检验水平），如图 2 所示。P 值是衡量置信水平的关键[7]，通常，P 值≤ 0.05 时，表明回归方程的趋势可信度较高。

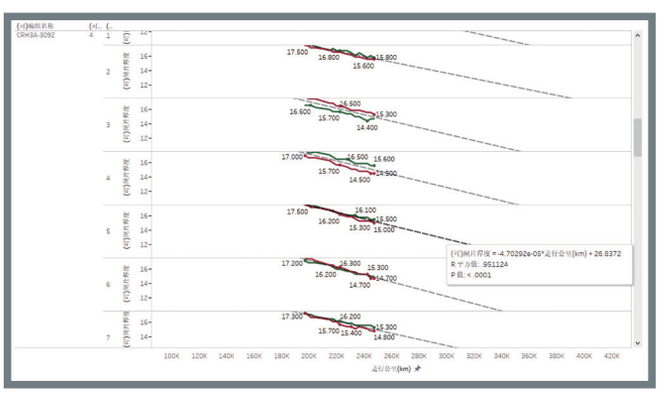

图1　闸片厚度-走行公里线性回归分析

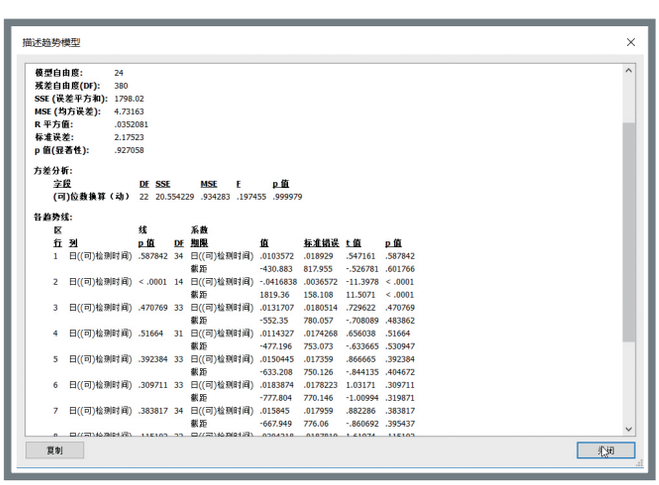

图2　趋势模型分析

从线性回归方程的斜率可得到闸片的磨耗速率，再从测定系数、显著性检验水平筛选出其中置信度较高的数值，对上述值进行二次统计分析可得到闸片磨耗的各种参数值。

3　闸片磨耗率及更换作业分析

由于各种车型闸片磨耗的分析过程类似，此处，从闸片磨耗率[8]、闸片更换作业的角度出发，分别对 CRH3A 型动车组的拖车、CRH380A 型动车组进行分析。

3.1　CRH3A型动车组拖车闸片磨耗分析

3.1.1　拖车总体磨耗率

如图 3 所示，选择 0.05 mm 作为数据桶间隔，以柱状图对拖车磨耗率进行整体展示。图 3 中，各柱状图构成了较典型的正态分布特征，可见 CRH3A 型动车组闸片磨耗率集中在 0.35 ~ 0.5 mm/ 万 km。

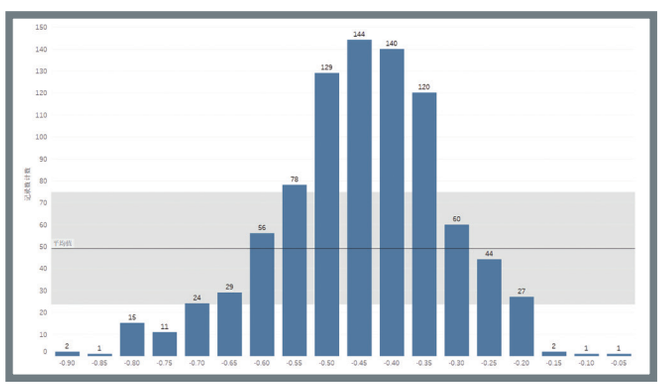

图3　CRH3A型动车组拖车闸片磨耗率（mm/万km）

3.1.2　线路磨耗差异分析

为进一步调查动车组在不同线路磨耗的差异，筛选出在成（都）绵（阳）乐（山）线运行的 CRH3A-5256 和 CRH3A-3107 型动车组，将上述 2 组车磨耗率总体分析结果在图 4 中用颜色标记（红色、黄色），图 4 中显示，磨耗率主要集中在 0.5 ~ 0.6 mm/ 万 km，远大于 0.35 ~ 0.5 mm/ 万 km。分析磨耗率上升因素，主要与成绵乐沿线车站较多、频繁启停有关。

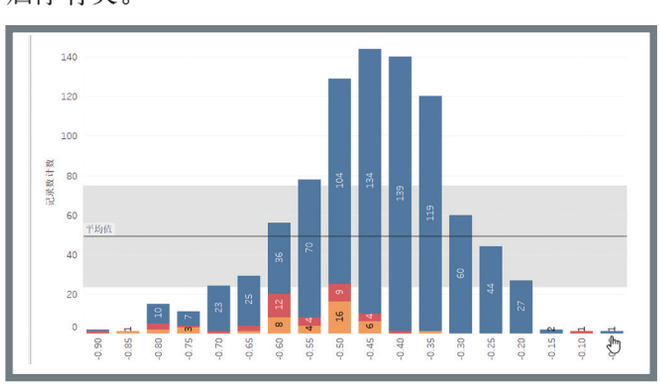

图4　拖车在成绵乐线磨耗率差异分析

3.1.3　偏磨分析

图 5 显示了拖车各端位偏磨情况，蓝色代表各

偏磨数据（同一闸片左侧与右侧闸片差值）的分布情况、黑色虚线代表偏磨量均值，均值位于 0 值左侧、代表该端位左侧闸片比右侧闸片厚度低。图 5 中，均值排名前 5 的数据都有一个统一的规律，全列以 01 车司机室为前端方向，所有闸片靠近轴端的一侧磨耗均高于远离轴端侧；同时，转向架对角线位置存在偏磨量较高的情况。

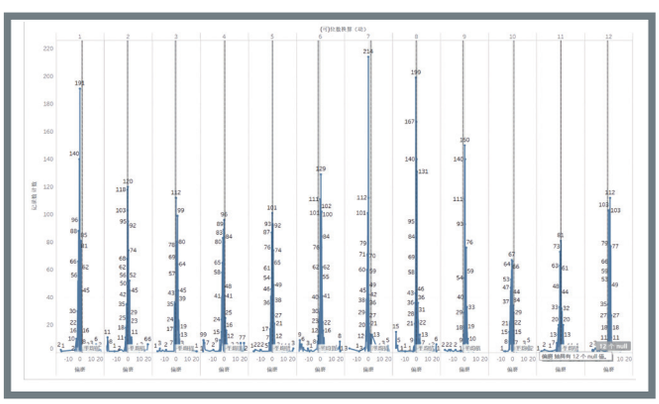

图5　拖车闸片各端位偏磨分析

3.2　闸片更换作业分析

图6展示了历年CRH380A型动车组闸片更换作业班组的分布情况，其中，二级修班组共完成了闸片更换总数的99.17%，一级修班组更换闸片记录为14次。说明现有的一级、二级修修程基本把闸片更换作业转移至二级修白班进行，避免了夜班大量更换闸片作业会出现错装、漏装或安装不到位的风险。

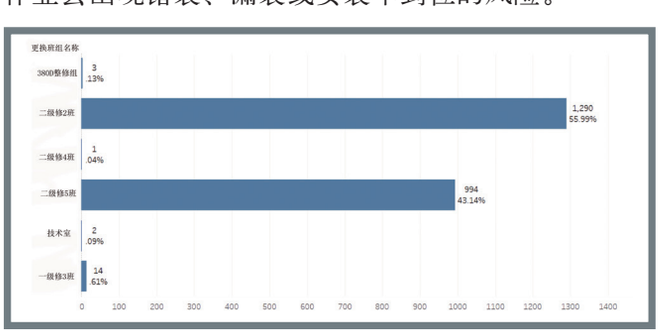

图6　CRH380A型动车组闸片更换作业班组分布

图 7 主要统计了闸片每季度更换总量、每年相同月份闸片更换量的同比上升、下降百分比。从总量来看，数据显示每年 3 季度、1 季度分别为闸片更换的高、低时期，需要提前做好闸片物料配备、人员安排等工作。同比上升、下降百分比方面，若后期配属车组数基本不变，可通过变化趋势大致掌握该车型的总体运行状况（如走行里程的长短、开行交路的不同等）。

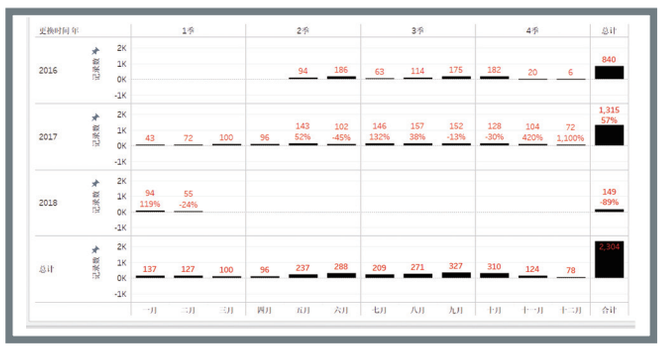

图7　CRH380A型动车组闸片更换量分布

4　大数据的地图分析功能拓展

大数据分析除了对数字本身进行统计归纳外，还可结合地图属性对地理位置相关的故障进行分析[9]，目前，可列举的如弓网硬点、踏面硌伤、网侧谐波、车辆晃动、过分相故障等均与具体位置有紧密联系，此处，以 CRH1A 型动车组曾经批量出现的网侧谐波滤波器过电流故障为例，结合 GPS 数据，展示地图分析功能。

图 8 显示了 2015 年 12 月 20 日单个动车组在同一天时间内先后多次出现网侧谐波滤波器过电流的情况，通过对车载数据中 GPS1、GPS2、GPS3 数据进行解析，得出每次故障的经纬度坐标，并在地图中以黄点标记，如图 9 所示。

图 9 中，3 起网侧过电流故障范围均锁定在遂宁附近，维度 30.55°，经度 105.53°，可见该故障具有一定规律特征，存在弓网供电及车组受电之间耦合作用不良的情况，通过故障区间的锁定，方便进一步排查车辆系统与供电系统的不稳定性因素。

5　结束语

本文相关大数据的分析结论已得到现场验证和应用，随着研究工作与生产作业的不断结合，就闸片检修工作而言，其主要问题在于如何合理制定检修限度、均衡任务，以及合理安排人员[10]。若检修标准制定过高、闸片浪费量巨大，但标准制定过低、作业量难以控制，可能会出现时而大量闸片更换、时而无闸片更换的情况，作业风险高，需要投入大量人力进行闭环卡控。下一步要研究的工作是：在本文闸片数据分析的基础上，进一步深化闸片状态检

修方案，以闸片状态修的方式取代预防修，达到闸片厚度利用率最大化、日 / 时检修任务均衡化的目的。

日期	车号	故障代码	故障描述	GPS 1	GPS 2	GPS 3
2015/12/20 16:17				287	27147	1818
2015/12/20 16:30				799	27150	4910
2015/12/20 16:43				287	27400	6918
2015/12/20 16:58				1567	27439	1815
2015/12/20 17:06				2847	27394	8474
2015/12/20 17:15				2847	27448	13338
2015/12/20 17:26				2335	27418	5657
2015/12/20 17:37				31	27405	12302
2015/12/20 17:42				31	27422	8
2015/12/20 17:46	5	3604	网侧谐波滤波器	287	27398	5634
2015/12/20 18:05				799	27192	8226
2015/12/20 18:21				31	27177	8473
2015/12/20 18:36				13598	27164	14350
2015/12/20 18:51				12318	27171	12804
2015/12/20 19:07				11038	26933	5937
2015/12/20 19:36				8734	26892	8464
2015/12/20 19:48				9758	26898	7169
2015/12/20 20:08				11550	26630	13341
2015/12/20 20:46				11038	26629	2830
2015/12/20 20:58				11294	26650	7713
2015/12/20 21:04				10782	26624	5420
2015/12/20 21:21				8734	26886	10769
2015/12/20 21:29	5	3604	网侧谐波滤波器	8222	26920	4127
2015/12/20 21:29	5	3604	网侧谐波滤波器	8222	26920	4127
2015/12/20 21:29	5	3604	网侧谐波滤波器	8222	26920	4127
2015/12/20 21:35				8222	26924	4896
2015/12/20 22:11				12318	27172	12548
2015/12/20 22:30				14622	27173	13589
2015/12/20 22:53				543	27158	14133
2015/12/20 23:09				287	27392	1552
2015/12/20 23:30				2847	27448	13338

图8　网侧谐波滤波器过电流故障数据

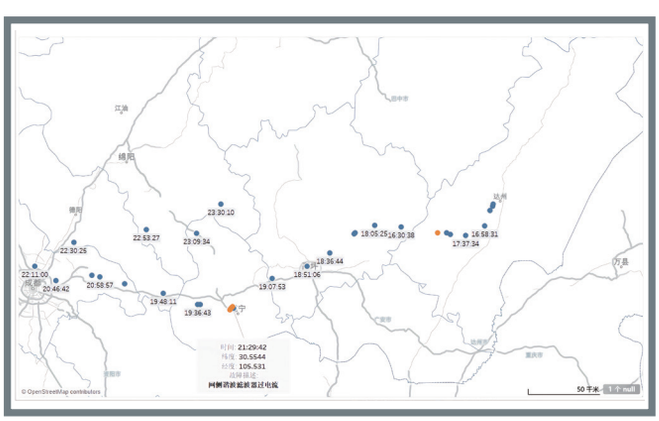

图9　故障坐标结合地图分析

参考文献：

[1] 张　春，李鹏程 . 基于状态维修的动车组关键部件寿命预测 [J]. 铁路计算机应用，2015，24（7）：1-4.

[2] 吕晓艳，刘彦麟，单杏花，等 . 基于大数据平台的铁路旅客群体分析应用研究 [J]. 铁路计算机应用，2016，25（9）：28-30.

[3] 陈建成，庞新生，李　川 . 统计学统计数据分析理论与方法 [M]. 北京：中国林业出版社，2013.

[4] 延婉梅 . 动车组大数据清洗关键技术研究与实现 [D]. 北京：北京交通大学，2015.

[5] Ashutosh Nandeshwar.Tableau Data Visualization Cookbook[M]. Birmingham Bretain:PacktPublishing,2013:1-133.

[6] 刘永科 . 地铁制动闸片异常磨耗原因分析及解决措施 [J]. 山东工业技术，2014（20）：48-49.

[7] 申玉伟，曹晓祎 . 我国用电量的多元回归分析 [J]. 电子技术与软件工程，2018（24）：225.

[8] 朱伟鹏 . 深圳地铁 11 号线受电弓碳滑板磨耗率研究 [J]. 铁道机车车辆，2018，38（4）：121-126.

[9] 姜春莹，何春光 . 基于电子地图的电力抢修智能调度系统的研究 [J]. 中国高新区，2018（12）：26.

[10] 丰雪霏 . 动车组运用检修的修程修制优化与实践 [D]. 北京：中国铁道科学研究院，2018.

责任编辑　王　浩