Discovering hidden connections with big data

Analytics revealed hidden cost drivers – and some surprises – for a transportation maintenance agency

By Kim Patel, Sekar Sundararajan, Charalambos Marangos and Emory W. Zimmers Jr.

Experience working with service sector businesses has shown that current and historical operational data may contain significant undiscovered trends linked to opportunities for improvement. Often, standard reporting tools do not help visualize or identify these hidden big data relationships, preventing companies from improving business processes and outcomes.

This article outlines a strategic approach for applying data mining methodologies to maintenance operations at a transportation services agency, but it is transferable to other organizations. The article is a case study that shows how advanced software and emerging analytics techniques, including data modeling and data mining, can analyze and transform current and historical data into insights to guide decision-making in maintenance operations and related areas. It includes strategic development and how data mining tools were used to identify opportunities and pinpoint specific job characteristics or best practices most likely to improve performance and reduce cost.

The transportation organization in this case study performs approximately 150 different maintenance activities on thousands of miles of roads and highways annually. Management thought that, based on traditional reporting structures, anecdotal input and input from managerial and operational personnel, the program could reduce maintenance costs. The transportation services agency was under continuous management pressure to reduce overall maintenance spending.

The agency had collected multiple years of maintenance data and developed standard metrics and reports. However, the agency didn't use data mining tools to analyze the data further and discover unidentified cost drivers. The agency wanted an approach that would regularly analyze historical data, drilling down to pinpoint areas of improvement within maintenance operations. Using a case study example with representative data, the project used a six-step approach.

1. Understand business objectives

Visible management support is necessary when a program starts. As a first step, senior executives helped to define the business objectives of the project, providing the overall goal. Specifically, the objectives were to:

- Analyze performance trends to existing maintenance standards.
- Identify statistically valid correlations and controllable variables to improve production efficiencies.
- Recommend new production standards for consideration.
- Provide guidance to align data collection methods to enable better accuracy and decision-making.
- Pilot the use of data mining and analytics techniques and tools on agency maintenance data to determine if these methodologies can provide insights to improve performance.

2. Form cross-functional team

The next step was to form a team with a balanced combination of experienced agency personnel who had contextual business knowledge, employees with software systems experience, and people with specific skill sets in applying analytics techniques.

For example, the team for this project comprised managers from the maintenance organization with institutional knowledge; information technology staff from the agency itself were selected for their ability to extract data from the existing enterprise resource planning (ERP) or associated systems; and a student, faculty and mentor consulting team based in Lehigh University's Enterprise Systems Center's Advanced Analytics Laboratory, which had expertise in analytics.

3. Establish KPIs and set goals

The team then developed the key performance indicators, such as how well specific activities met the engineering resource utilization standards, which helped define the success of the project. In some cases, federally mandated guidelines were a factor in calculating a given resource utilization standard.

Historical agency knowledge of the organization's specific pain points helped focus on high-cost and high-utilization maintenance operations. The team identified pilot areas for examination based on preliminary data analysis, which identified costs by activity.

For example, in order to reduce the overall spending on 149 activities, the strategic approach focused on analyzing one activity, which agency management helped choose. Let us call it "Activity MP."

As is often the case, management thought that analyzing and improving operations for the activity with the highest expenditure would reduce costs the most. The agency set this particular maintenance activity's performance standard at a specific tonnage per man-hour. Activity MP costs an average of approximately $17.5 million a year with a set number of standard labor hours per year.

4. Create a master data file

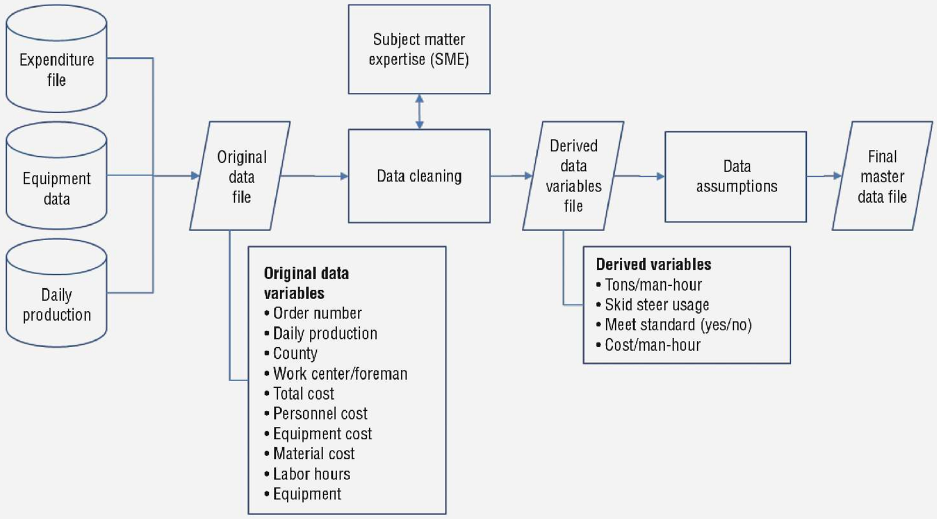
Next, the team had to identify data sets that contained the cost, labor and other recorded details of road maintenance activities, information such as location, name of foreman, equipment used, etc. The agency's IT environment did not allow for direct connection to the raw data, so access to reports and flat files was provided. Collecting the raw data involved some back and forth with the agency ERP system.

The Advanced Analytics Laboratory team figured out how to extract the relevant information from the ERP system to create a master file that served as the data repository for data mining analysis. As the project team discovered new data links, it had to access the ERP system repeatedly for appropriate information. In this case, the team collected five years of maintenance operations data for "Activity MP."

FIGURE 1

Master your data

The analytics team used this approach to compile a master data file.



The team chose SAS Enterprise Miner software for the data mining because it could use a flat data file (in this case a large, two-dimensional, sometimes sparse table) as the input to the analysis. Three sources of historical data from the ERP system were used to compile a flat file of this activity's total expenditure, the maintenance equipment used for the activity, the chosen activity's daily production and 25 additional data variables. In the end, the combined data file consisted of 28 variables. Figure 1 shows the approach used.

Each row of the table created is a record, such as a work order, and each column is an attribute, such as miles repaired or road type (rural, city). For example, vehicle routing information tables, maintenance staffing scheduling tables, activity breakdown tables, activity schedule, route classification tables and related factors had to be extracted from the ERP systems and combined to form the flat file.

Such a large table of data that spanned 365 days and five years across about six dozen maintenance territories with multiple data entry processes contained errors and data entry inconsistencies (e.g., zero production time with associated cost, outperforming standards in unexplainable ways, and other data entry adjustments to adapt to the ERP system's limitations on data entry). This data had to be "cleaned," that is, removed from consideration or adjusted and fixed accordingly.

Additionally, this step identified correlations among variables, such as the use of a specific tool and the cost associated with that tool. Agency team members who knew about specific maintenance activities were essential to identifying and explaining the data inconsistencies, such as performing an activity without using labor, and variable correlations.

5. Conduct data mining analysis

For the next step, the sample data was imported into the software to explore the data set and visualize trends. The software highlighted missing values, such as missing job orders associated with significant costs (methodology used to reconcile project costs), and inconsistencies, such as performing an activity without using people (again, an empirical cost reconciling methodology).

The results were presented to the agency. The team focused its analysis on a major repetitive activity across the entire organization to serve as a proof-of-concept of the analytics tool and approach. The preliminary analysis showed a significant change in productivity over the five-year period as measured by classical industrial engineering and financial metrics. The largest cost component was labor.

This initial finding peaked curiosity about what other insights could be gained from this flat file. The team asked and answered questions never envisioned before the introduction of the data mining tool. New questions focused on machine utilization, individual foreman performance, overall territory performance, the agencywide influence of strategic decisions like the use of mechanized assistive technology and the relationship of many of these to the traffic density (urban, suburban and rural). Additionally, the preliminary findings shed light on the increased amount of data that needed to be captured, as well as the potential for modifying some job descriptions and changing policies to permit that new level of data granularity.

This led to a critical phase: the need to understand the root cause of the productivity change and what job, activity or machine characteristics or attributes should be investigated and monitored to understand the productivity variation and share best practices across the organization.

Analytics identified the factors responsible for the change in labor cost. In response, variables were computed to look more closely at the data. The derived variables included the tons produced per man-hour, the cost per man-hour, whether the activity was performed using hand tools or mechanized tools, and whether the activity performance met the current resource utilization standard on a daily basis. Then, the data set was modified to distinguish the independent and dependent variables. The independent variables were the location, daily production, maintenance equipment used and workforce composition. The dependent variables were labor hours, labor cost, maintenance equipment cost and cost of the material used.

Dependent variables were dropped from further analysis, so only independent variables were used to establish the pattern for a service that meets or exceeds the standard. The data set was broken into two parts, one set for testing the hypothesis and the other to validate the results. The SAS decision tree tool sets identified the most influential variables and found correlations among variables.

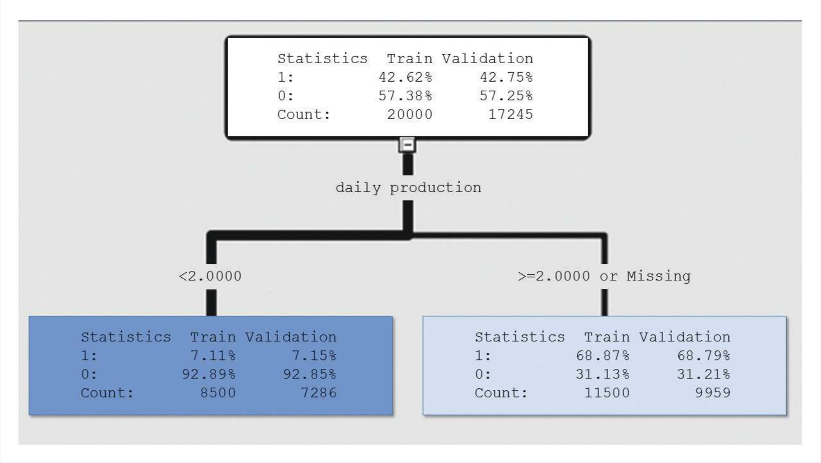
Four different analyses were performed on the data, namely, geographic location of maintenance activity, daily production volume, equipment type used and experience of employee:

1. Analysis (location): This is an interactive decision tree to analyze the independent variables. The analysis showed a scattered performance, with some locations underperforming significantly relating to the standard, while other locations outperformed the standard by a significant measure. This lack of consistency prevented conclusions. However, the lack of information found through the decision tree analysis highlighted the need to examine other variables, such as daily production volume, equipment type and employee experience.

FIGURE 2

Daily production decision tree

This data analysis shows the likelihood of maintenance personnel meeting performance standards when producing less than, greater than or equal to 2 tons per day. Note that "1" indicates the observation met the standard, while "0" indicates the observation did not meet standard.



```
Statistics  Train  Validation
1:          42.62%     42.75%
0:          57.38%     57.25%
Count:       20000      17245
```

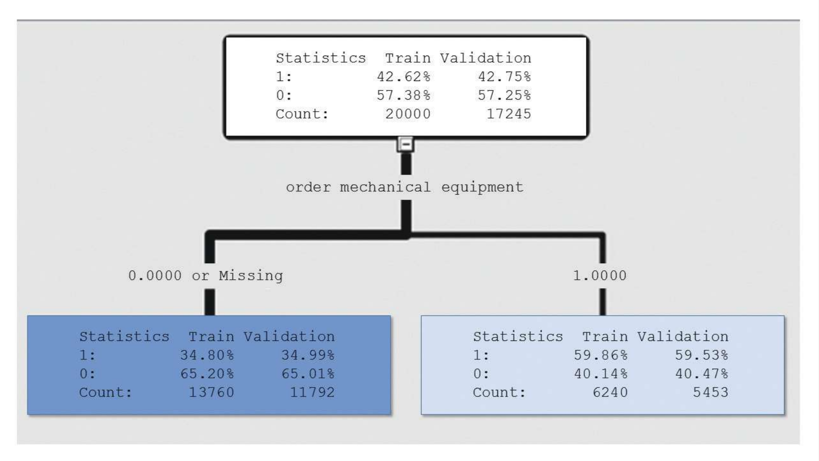
```
Statistics  Train  Validation
1:           7.11%      7.15%
0:          92.89%     92.85%
Count:        8500       7286
```

```
Statistics  Train  Validation
1:          68.87%     68.79%
0:          31.13%     31.21%
Count:       11500       9959
```

FIGURE 3

Manual vs. mechanical

This decision tree shows the likelihood of meeting performance standard using automated maintenance equipment (the right branch) or performing manually (the left branch). Again, "1" indicates the observation met the standard, while "0" indicates it did not.



```
Statistics  Train  Validation
1:          42.62%     42.75%
0:          57.38%     57.25%
Count:       20000      17245
```

```
Statistics  Train  Validation
1:          34.80%     34.99%
0:          65.20%     65.01%
Count:       13760      11792
```

```
Statistics  Train  Validation
1:          59.86%     59.53%
0:          40.14%     40.47%
Count:        6240       5453
```

2. Decision tree (daily production): This analysis determined the percentage of work orders that met the performance standard. Less than one-third met the performance standard on 50 percent or more of the days they performed the maintenance operation. On days when a lower tonnage was produced, there was a 92 percent probability of not meeting the production standard (Figure 2). This showed that the agency should consider scheduling the activity to increase the percentage of meeting the daily production standard and reduce labor hours.

3. Decision tree (manual tools vs. mechanically assisted equipment): This analysis showed that daily production met standard 34 percent of the time when performed manually and 59 percent when the activity was performed using automated maintenance equipment (Figure 3). As expected, mechanized tools were more efficient, but this analysis did not consider the investment for the tool or the increased expense of using a mechanized resource with limited availability. The team suggested that the agency study this area more, including the impact on overall cost, production efficiency and updating standard operating guidelines for activity MP.

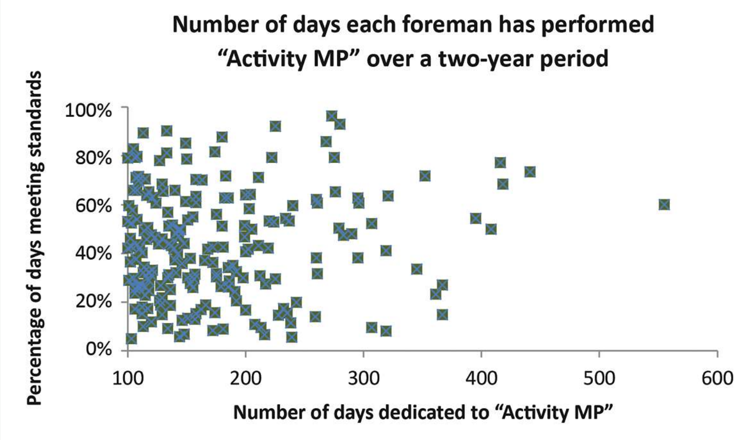
Decision tree (employee): This analysis revealed that employees who frequently performed the specific maintenance service, the people considered more experienced, did not have a higher likelihood of meeting the performance standard. The agency did not expect this finding and identified this as a possible training issue. The analysis proved, as shown in Figure 4, that the time-on-the-job vs. meeting-the-standard hypothesis was not supported.

This does not cast aspersions on the employee, as workers can excel when they have the proper training and support. Instead, these findings point out that some employees are better skilled at the maintenance service, and the agency can use their performance to determine consistent best practices, such as improving scheduling by grouping multiple activities together or adding on-the-job training.

FIGURE 4

Experience? Not that important

This graphic shows that the percentage of foremen who met standards did not correlate to the number of days they have performed the maintenance activity.



6. Summarize findings and define opportunity

Examining the variables using data mining analysis led to a number of findings. They include:

1. The geographic location (or the maintenance district) of the maintenance activity did not have a bearing on achieving the performance standard. The maintenance activities performed in locations within the same district also did not have a bearing on achieving the performance standard.
2. Days with less tonnage scheduled, or when a minimal amount of the maintenance activity was performed, had a high probability of missing the performance standard. Planning and scheduling to increase the number of maintenance activities to increase the daily tonnage performed could improve the rates of meeting the performance standard.
3. Employees who used mechanically assisted equipment met the performance standard more often than those who performed the maintenance manually. Increasing the use of mechanically assisted equipment could increase the agency's rate of meeting the performance standard, especially when scheduling multiple or grouped activities that use costly and limited equipment.
4. Experience was not a significant factor in achieving the performance standard.
5. The current process used in capturing maintenance activity data should include collecting data for each step involved in completing the maintenance activity. Some examples include travel time to maintenance shed and job location, weather conditions, time of day, and mapping of exact activity location, such as by using GPS.

The above findings represent a sample of the overall analysis of project-specific variables using data mining techniques. Unexpected trends can be found within data sets, leading to deeper analysis and new insights. The analysis showed that geographic location and employee experience did not contribute to whether the maintenance activity was likely to meet the performance standard.

However, using mechanically assisted equipment and careful scheduling of maintenance activities appeared to improve productivity. The analysis also discovered that when the original standard was developed, mechanical equipment of this type was not available as an option. In addition, the importance of the proximity of repair locations was not formally modeled. The interactions of these "Activity MP" specific variables were highlighted by the analytics techniques used in this project. We anticipate that these insights will be important in developing the agency's next generation of standards and operational best practices.

Key insights and critical success factors

The work on this project and similar ones has led to the following insights and critical success factors in big data analysis. These insights can guide organizations in exploring how to use data mining tools for their operational improvements.

Ascertain data granularity and availability. Using data mining and analytics modeling techniques, companies can look deeper into their big data in search of trends in performance. Like the transportation services agency case presented here, others can identify their largest cost drivers and examine potential operational changes to reduce spending significantly.

This case study found that this agency collected data for myriad variables, but the data was not granular and consistent enough. The agency discovered that granular and consistent data would help develop more insights through data mining and modeling.

Data availability may be an issue, but in general it is advisable to get started with what you have. Data mining is often a discovery process. It is like peeling layers of an onion. As one peels a layer, new questions to ask and test are found. But this should be treated as a victory or insight because it may lead to the capture of new data that are likely to provide greater opportunities to reduce costs in the future.

In this case study, the inconsistencies and missing values identified during the data integration, selection and cleaning process have prompted management to institute data entry training.

Take a balanced view of data mining. In general, people often end up with one of two viewpoints. Proponents of the technique and technology may at times oversell and overpromise. On the other end, opponents fear that it is a black box, and they do not trust the insights because they do not understand the approach's theory and mechanics.

While working with this transportation services agency it became clear that many held the view that data mining was a black box that provided answers. However, the project showed that data mining can provide information for decision-making with the data that is available if a business context savvy team can ask the right questions.

Teams using this approach should validate their results against actual scenarios in the field, deploy the findings and be willing to adjust the initial approach and plan as an improved view of reality emerges. Fortunately, the tools at hand are capable of testing against the new reality quickly.

Selecting the right team leadership is an important element in overall success. The team must be technically competent, patient and persistent. Because of the various views about data analysis, facilitating a convergence of disparate points of view usually will obtain the best value for the business.

Business insight needs to be coupled with technology insights. Often, people misunderstand, thinking that new technologies will provide intelligence without the user's input. But data mining tools can lead to insights not easily obtained only when a group of business experts coupled with technology specialists collaborate to brainstorm and ask the next "Why?" After initially distancing themselves, some agency personnel soon realized the power of the analytics tool. When they became intimately involved with the data, analysis and patterns, they realized how this information would enable them to leverage their contextual knowledge.

Use a proof of concept approach. Many of the findings may be counterintuitive and will be challenged by management and practitioners in the field. A proof of concept approach provides the opportunity to build a foundation, validate the model and obtain buy-in from key personnel who will be instrumental in implementing and institutionalizing a changed approach or process.

Data mining is an adaptable methodology

Applying data mining methodologies to current maintenance operations at this transportation services agency helped reveal significant opportunities for improvement. In the example described here, standard reporting tools did not sufficiently visualize or identify hidden relationships in the data.

This case study demonstrated the general approach and showed how the agency and project team collaborated to help develop new strategies. The team used data mining tools to identify gaps in performance and pinpoint specific job characteristics or best practices most likely to improve performance.

The selected case study described the use of advanced software and emerging analytics techniques, including data modeling and data mining, which were employed to analyze and transform current and historical data into insights to guide decision-making in maintenance operations and related areas. As a final note, other organizations should be able to adapt this general approach for their specific circumstances. ❖

Kim Patel is an analyst within the analytics division of ITG (Investment Technology Group), working with asset management firms to analyze trading data and identify opportunities for improvement in trading performance. She holds a master's degree in management science and engineering from Lehigh University.

Sekar Sundararajan is a senior manager at Kurt Salmon, a global management consulting company. He has worked across multiple industries, including retail, consumer goods, automotive, industrial goods and logistics. He has more than 20 years of industry experience in retail strategies, strategic operations, agile manufacturing, sourcing and supply chain.

Charalambos A. Marangos is a consultant for the Enterprise Systems Center's Advanced Analytics Laboratory and president and founder of Zephyros Inc., a consulting company that specializes in business and industry. Marangos has served as a developer and instructor for part of an innovative sophomore-level engineering graphics course that introduced students to SAS.

Emory Zimmers is a professor of industrial and systems engineering at Lehigh University and director of the Enterprise Systems Center, including its Advanced Analytics Laboratory. In this capacity, he works extensively with companies to identify critical new areas for research and provide experiential learning opportunities for students through industry projects. Zimmers supervises the senior capstone course as well as the engineering leadership minor. He has consulted for numerous companies and government agencies.