Soft Science and Heavy Haul: Multivariate Statistics and Systemic Drivers

Dr. Dave van der Meulen
Specialist: Planning & Technology, Spoornet
dave@railcorpstrat.com

Summary: This paper introduces the application of multivariate statistical procedures to the heavy haul environment. It develops practical application of cluster- and factor analysis to understanding derailments of heavy trains, leading to examples of four archetype derailments and their systemic drivers. It concludes that soft science can offer valuable understanding in complex situations.

1. EXTRACTING UNDERSTANDING

1.1. Introduction

The author has attempted to extract understanding from complex situations, in a railway with both heavy haul and general freight components, to identify systemic drivers. This paper describes some of the successes, and some of the difficulties encountered. It offers empirical guidance based on that experience, because applying a technique in a new field is fraught with methodological obstacles, a selection of which are described below.

1.2. Organizational boundaries

Organizational boundaries partition businesses into constructs, such as departments, that originate within human discretion. Internally, they focus management attention, hold people accountable, facilitate financial control, specialize by discipline, and so on. Externally, they align with customers, suppliers, and other stakeholders. Organizations also reorganize as they adapt and develop. As the Information Age has supplanted the Industrial Age, organizational elements have decreased in size and increased in diversity, while the pace of change has accelerated, to maintain alignment with a dynamic environment.

However, technical problems in a turbulent organization respect no such boundaries. Train handling that results in coupler failure on particular days of the week, interaction between bearing failure, braking characteristics, and topography, and even incidents that seem related to tides, are some multidisciplinary examples. Any mapping from such problems to organizational structure can be little more than coincidental. The author therefore hypothesized that many systemic drivers remain latent below organizational structures, impervious to structural design or redesign. The work reported here addressed the challenge to render them discernible.

1.3. Core- and generic technologies

The author posits that a core technology is relatively impervious to organizational turbulence, because the output of a productive organization depends on the supporting assets and processes. Asset-specific organizations, of which railways are the archetype, necessarily rest on a core technology. The author argues that many systemic problems are encapsulated in the exploitation of that technology. Thus the manufacture of potato crisps, say, requires a different technology, and creates different problems, compared to, say, a railway. Such core technologies are distinct from generic technologies, which apply to diverse business organizations. Generic technologies come and go as their life cycles progress. They are therefore relatively exposed to the cut-and-thrust of a competitive environment. Information technology is the present archetype of this kind — enterprise management software is as applicable to the manufacture of potato crisp as it is to railways. Although information technology may facilitate productive processes directly, e.g. in ECP braking, or indirectly, e.g. in condition monitoring, the author has also observed a vain expectation that information technology will by default also offer insight into systemic drivers. That can however only happen when applications are implemented with that objective in mind.

1.4. Operational data

The Information Age has potential to facilitate understanding many issues that carried over from the Industrial Age. However, the author has observed that operational data still tend to be processed into statistics that are predicated on Industrial Age structures, e.g.

departments, such as human resources, finance, rolling stock, etc., *locations*, such as regions, corridors, routes, etc., or *components*, such as bearings, motors, wheels, etc. Such descriptive- or summary statistics may obscure or lose the underlying drivers within the larger system, and instead focus attention on incremental improvement within organizational structures. While they may support analysis of departmental performance, they are also amenable to massaging to demonstrate compliance with departmental objectives, while obfuscating contribution (or lack thereof) to overall objectives. They cannot explain why, despite apparent improvement within departments, an organization as a whole may decline, or fail to optimize.

1.5. Organizational behaviour

One negative consequence of an obsession with management is holding people accountable for that over which they have inadequate control. Management's toolkit resides in understanding the relations among the variables in their area of responsibility. However, if data structured according to organizational boundaries obscure systemic understanding, the author has observed that managers tend to cover their behinds: Furthermore, when pushed for results, their natural inclination is to impute causality where causality does not necessarily exist.

2. STATISTICAL ANALYSIS

2.1. Application to technical problems

Technical problems in railways include many complex relations. Heavy haul railways, in particular, have apparently simple business- and technology models, but they stress equipment to the limit. Furthermore, large-scale system optimization does not necessarily occur at active constraints, so enlarging them may prove fruitless. Effective problem solving or optimization therefore needs grounded understanding of the relations. Observable phenomena frequently provide the only clues to what is driving a complex situation. The author has found that the most comprehensive way in which to explore them is first to identify and describe all relevant variables, and then to list all cases. This approach may yield a matrix of hundreds of variables and thousands of cases, or upwards of 10^5 data points that defy unaided human pattern recognition, and hence require application of statistical procedures.

2.2. Appropriate procedures

One can analyze such data by case and by variable. The intuitively obvious way is to compare cases with each other, and group those that are similar: Cluster analysis is the multivariate statistical procedure that does just that. The more obscure way is to examine how variables co-vary with one another: Factor analysis is the multivariate statistical procedure that constructs a reduced set of latent variables, rotated in multidimensional space to purify the construct underlying each. The procedures are introduced by way of example in Sections 5 and 6.

2.3. Their background

Cluster- and factor analysis are exploratory procedures that together extract maximum understanding from data that represent a complex system. They originated in the soft human- and social sciences, to analyze intangible constructs from large samples — antithetical to the hard science of engineering. However, much railway hardware finds application in complex systems that subsume human behaviour; hence, they cannot be adequately understood by applying only engineering science. Consider issues such as train driver behaviour when service exigencies conflict with social commitments, or wagon examiner behaviour when a mine hands over trains late. In the general freight and heavy haul environments, cluster- and factor analysis also become powerful investigative tools.

2.4. Omitted variables

The author has observed that managers or researchers may not be aware of all relevant variables, or, even if they do know about them, they may not be accessible for measurement. They may also be latent, e.g. the propensity of a train driver to pass signals at danger. One may observe that D follows A, and hence infer that A causes D. If, however, in truth A causes B, B causes C, C causes D, and B and C are unknown or immeasurable, the inference may be unreliable. Extraneous variables not accounted for in the research design may have influenced B and C. The author regards this as a high-risk area for managers, because their forceful nature inclines them to impute causality where even an association might be tentative.

2.5. Root- and contributory causes

It is often difficult to establish causality with any confidence. Frequently, all that one can confidently assert is that a particular category of occurrence was present or not. It is therefore useful to distinguish between causality and association. While it may be impossible to establish causality, because of omitted variables, it is usually possible to establish association. In practice, that may provide sufficient insight with which to address a problem. The risk-laden human tendency to assign root- and contributory causes at the time of data capture thus rests on ignorance and presumption.

2.6. Interdependency

Factor- and cluster analysis apply to settings where variables are interdependent, i.e. they are interrelated, without designating some dependent and others independent. This is typical of a heavy haul railway, which connects a mine with and a port in a cyclical operation. Purposive efforts to manage the system, and events that disturb it, are difficult to attribute solely to one entity or one part of the cycle. A delay on an empty train may influence scheduled maintenance at the mine, which in turn influences the loading of another train. They are thus interdependent.

3. MEASUREMENT

3.1. Variables in practice

A database field, or a spreadsheet column, typically represents a variable. The author has found it useful to present this notion as minimum learning for practical operating people who acquire data. They know where to get it, but need also to appreciate its value and fragility. One can also describe a variable as a symbol to which one assigns numerals or values [1]. The numerical value assigned depends on the variable's properties [2]. At minimum it may have only two values, reflecting the presence or absence of a property, e.g. *did a bearing overheat, or not*?

Causality is often neither accessible, nor necessarily sought. One needs to design data sets accordingly. To simplify, the investigator can reasonably designate variables known or believed to be relevant to the research question. It is wise to err on the generous side, because factor analysis loads redundant variables onto the same factor. By inspection, one can cull them until all contribute some understanding. For example, there is probably no need for both *train mass* and *number of wagons* for unit trains, because they vary in unison.

3.2. Levels of measurement

The author has found that many people believe that measurement needs a physical instrument, such as an accelerometer, caliper, thermometer, or whatever. There are, however, many other instruments, like those from the behavioural sciences, whose apparently crude measures nevertheless yield more insight than no measurement at all.

There are four levels of measurement, namely nominal, ordinal, interval, and ratio, with corresponding scales. Nominal measurement assigns a label, which cannot be ordered or added. Therefore it does not support all statistical procedures. Ordinal measurement requires that the objects of a set be rank-ordered on an operationally defined property. They are frequently treated as equal appearing, as in the *strongly agree — agree — do not know — disagree — strongly disagree* scale, without significant loss of rigour. Interval, or equal-interval, scales possess the characteristics of nominal and ordinal scales, and in addition, numerically equal distances on interval scales represent equal distances in the property being measured. They support mathematical manipulation in statistical procedures. In addition to possessing the characteristics of nominal, ordinal, and interval scales, a ratio scale has a natural zero that has empirical meaning [1], e.g. train speed, where zero means standstill. Ratio measurements minimize error variance, but one should consider using lower levels of measurement whenever ratio is not available. In cluster analysis, simple yes/no binary measurements support insightful clustering.

3.3. What cases are

The author has experienced difficulty in leading people to grasp what cases are, and thereafter to define them. They may not be intuitively obvious in a complex setting. A case is a single example spanning whatever set of variables is being measured. The bond that unites variables in a systemic set is that each case has a measurement for each variable. For example, a case may be a particular incident in a set of derailed trains, or a particular train driver in a study of job-related stress in a train driver population. Cases ought normally to emanate from management-driven, value adding, activities such as transactions, consignments, or in railway operations, trains. This is why many commercial information technology applications provide abundant valuable data: Processes that support value chains drive them, and cases are implicit in such data.

3.4. What cases are not

A day cannot be a case, because days relate to Earth's rotation, and not to human effort. However, cases do accumulate over time in many settings, thereby also promoting longitudinal- rather than snapshot analysis. They may also appear at once, such as a set of train drivers that have passed signals at danger. However, an incident is not necessarily a case. The author's work has led to appreciating that the act of numbering incidents sequentially at capture in a data base summarily severs them from each other, thereby precluding the possibility of discovering relations among them based on the outcome of human effort. All that they typically then share is a date, rather than their rich relation to systemic drivers of the railway on which the incidents occurred.

3.5. Cases in railways

In a railway, trains as cases make statistical sense, because they represent the prime tangible outcome of efforts to deliver service. As such, they bond all essential and legitimate value chain activities, plus all deviations there from, such as loads released late, crews taken ill, delays during preparation, locomotive failures underway, tipplers broken down, and the myriad other disturbances known to railway people, into a single case. Trains as cases make particular sense in a heavy haul system, because the operation is relatively easy to describe, and most extraneous disturbances are known. In practice, the author has found that a case may be a particular train in a set of trains from origin P to destination Q, or a category of incident, representing a sub-set of all trains on a particular route, e.g. those that derailed.

3.6. Avoiding bias

Not all trains suffer incidents — many indeed perform as intended. Unless there is a specific objective regarding a subset of all trains, such as studying derailments, one should include all trains, both good and bad, to achieve unbiased understanding of the systemic drivers. The author has found this perspective to be useful when a system approaches or exceeds its design capacity.

3.7. Stillborn statistics

Absent an appreciation of how fragile good data can be, the author has found many data to be captured in a diary, particularly incident statistics, such as break-in-twos, hot bearings, etc. The mindset seems to have originated in the handwritten diaries or logs of antiquity, when it was arguably the only way to do so, and continued into electronic databases without question. Such data are thus related by day and not by purpose. There exists a notion that data mining can extract understanding from poorly structured data or databases. However, it is not possible to mine relations that do not exist, or that have been destroyed in the process of capturing. The author uses the term stillborn to describe such statistics, because although conceived with a purpose, they cannot deliver.

3.8. Human issues

Human attributes are thought to contribute to incidents where judgment is questioned, such as signals passed at danger, exceeding speed limits, etc. Personal data is often complex, frequently unavailable, privacy protected, or subject to interpretation by registered professionals. There is therefore a significant risk that one might omit relevant variables, and so obscure understanding. Sometimes a surrogate measure may come to the rescue. For example, train driver stress is difficult to define operationally, more difficult to measure, and frequently not admissible in a unionized environment. However, in an electronically-controlled pneumatic braking plus distributed power pilot scheme, the author has measured contributors to stress as a surrogate for measuring stress.

First reaction to an incident typically implicates the closest person, frequently the train driver. Most incident records therefore name the person(s) involved. This may be one reason why train drivers are heavily unionized. However, the author has found that relevant personal data, such as domestic- and medical circumstances, often are conspicuously absent. Similarly, data relating to training and certification, and hours of duty and rest, which may be stored in off-line repositories, are hard to come by. Names are not manipulable by multivariate statistical procedures: Their anonymity ought thus not to stir contention when applied to associate human attributes with incidents. Nevertheless, personal attributes seem available only to the persistent.

4. CREATING A DATA SET

4.1. Workshops with stakeholders

The author has found it beneficial to workshop a study with line management and senior knowledgeable people, to secure buy in, and scope the project to deliver the expected understanding. It also taps the vast repository of knowledge, albeit unstructured and incoherent, that is available in such people.

4.2. Alphanumeric data

Except for classificatory procedures such as cross-breaks, which can only extract a small portion of the latent understanding in a data set, powerful statistical procedures such as factor- and cluster analysis, cannot process alphanumeric data. Columns or fields must therefore contain numeric measurements only. While some data can be no other than alphanumeric, e.g. Area 27 or Driver Smith, the author has found that many other apparently alphanumeric data can be converted to numeric data. A column or field that contains nominal variables, e.g. textual descriptions, such as *points run through*, or *cattle on track*, needs instead to spread to a column for each category with a binary yes/no, or 1/0, response. This may substantially increase the number of fields or columns, but it does make understanding accessible.

4.3. Missing values

Cluster- and factor analysis need a complete data set, i.e. measurements for all cases and all variables. Discarding either cases or variables that have missing values can eliminate most variables and/or cases, rendering a data set almost useless. One therefore has to diligently fill all cells with values. As a last resort, one may fill missing values with rationally determined substitutes, to avoid losing an otherwise useful case or variable. For example, if dynamic braking effort at the point of derailment was not recorded for a particular case, the normal dynamic braking effort at that point may be a fair estimate. However, if a measurement is only available for a few cases, one should rather delete the variable.

Maintaining, or reconstructing, a database can be a major workload. Not all variables are relevant to all cases. For example, a pantograph hookup shares nothing with a signal failure, or a telemeter failure. Normal behaviour is to leave such fields blank. However, statistical procedures need a full set of cases for all variables. The author therefore structures variables, columns or questions so that the default answer is no, false, or zero. All such fields can then become optional, making statistical analysis possible by simply inserting zeroes in all empty cells.

5. APPLICATION TO DERAILMENTS

5.1. The setting

This section illustrates application of the procedures described to 200-wagon trains on the Ermelo-Richards Bay coal export operation. The study examined systemic drivers of derailments only, over a five-year period. Inclusion of non-derailed trains would not have added any value. Cluster analysis would simply have grouped non-derailed trains in their own cluster, and they were therefore excluded from the data set. The variables ultimately selected appear in the left hand column of Table 1, in the order they loaded onto factors: Their descriptions follow (all at moment of derailment):

Braking and Traction kN — the amount of braking or traction exerted by the locomotives.
Slack — stretched, neutral, or bunched.
Gradient (per mil) — self-explanatory.
Brake Pipe Reduction — the amount of braking applied deliberately by the train driver.
Inflection — order of vertical curvature (+3=inflection on upgrade, +2=crest, +1=upgrade, 0=level, -1= downgrade, -2=dip, -3=inflection on downgrade).
Loaded/Empty — wagon condition.
Gross Trailing Load — self-explanatory.
Red Signal/Flag — train stopped by external initiative.
1st Derailed Car = Jumbo — the other wagons are smaller.
Running Number — new wagons are numbered consecutively, thus an indicator of age.
1st Derailed Bearing Defect — first derailed wagon had a bearing failure.
1st Derailed Wheel Defect — first derailed wagon had a wheel failure.
Transition Present — first derailed wagon in entry transition, in curve, or in exit transition.
Curve Radius (m) — self-explanatory.
Stock Rail Broken — self-explanatory.
Date (Excel Number) — to mark the passage of time.
Rake 1 Slackless Present and Rake 2 Slackless Present — problems were experienced with derailment of wagons with slackless drawgear.
Other Derailed Quantity — number of wagons derailed in addition to the first one.
Other Derailed Gap Present — number of non-derailed wagons between derailed wagon groups.
Rand Cost — direct cost of damaged assets.
Rail Joint Broken — self-explanatory.
Speed (km/h) — self-explanatory.
kN Instability — tractive/braking effort unstable due to wheel slip/slide or locomotive fault.
1st Derailed Coupler Defect — coupler broke on first derailed wagon.
Day of Week — numbered 1 to 7.
Gradient Length — length of gradient on which train derailed.
Neutral Section Involved — self explanatory.
1st Derailed Position — position in train of first derailed wagon.
Line Voltage Unstable — self-explanatory.
Temporary Speed Restriction — self explanatory.
1st Derailed Spring Defect — first derailed wagon had a spring defect.
Points Involved — self-explanatory.
1st Derailed Brake Defect — first derailed wagon had a mechanical brake defect.
Hour of Day — numbered 1 to 24.
Month of Year — numbered 1 to 12.
Emergency Brake Applied — self-explanatory.

Note that the order of loading onto factors differs from the logical, organization-oriented, groups in which the variables were conceived, namely:

First 100 wagons description, and second 100 wagons description (two generations of wagon are in use),
Track description,
Track condition,
Overhead equipment description,
Train speed,
Traction condition,
Automatic brake operation,
Slack condition,
First wagon derailed,
Other wagons derailed, and
Cost.

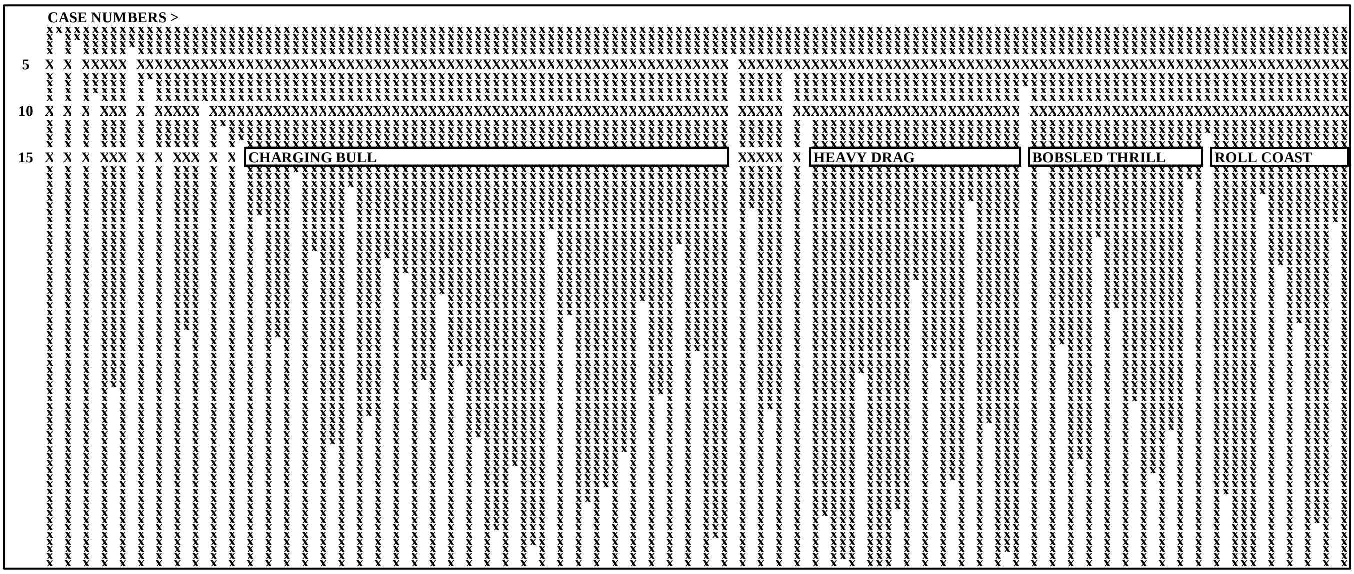
The change in order calls attention to the difference between the discretionary nature of organizational perspectives, and the primary nature of systemic drivers.

The set of relevant incidents limited the number of cases. One would probably have liked more data, but because derailments are managed down, the data available is inherently meager. It was therefore necessary to constrain the number of variables, to balance the number of cases.

5.2. Cluster analysis

In the hierarchical method, clustering begins by finding the closest pair of objects (cases or variables) according to a distance measure, and combines them to form a cluster. The algorithm continues one step at a time, joining pairs of objects, pairs of clusters, or an object with a cluster, until all the data are in one cluster [3]. At the top of the hierarchy, all cases are thus in the same cluster (because they reside in the same data set). At the bottom of the hierarchy, all cases are unique (to the extent that a finite number of variables cannot distinguish them sufficiently, some cases may appear identical). Applying the SPSS cluster analysis procedure yields the dendrogram in Figure 1. It is difficult to scale it to show minute detail. Simply, the vertical icicles represent cases that cluster together in ever increasing numbers. The underlying distinctions among clusters are found by inspection, after

Figure 1: Dendrogram



grouping the members of each cluster together in a spreadsheet and searching for similarities — a tedious task.

5.3. Interpretation of clusters

By virtue of their soft science origin, even mathematically rigorous results from cluster- and factor analysis require interpretation. The outputs from both procedures are still mere numbers. The number of clusters finally selected for interpretation thus depends on the judgment of the analyst. The trade-off is between a small number of relatively non-homogeneous clusters, such as empty- and loaded trains, and a larger number of relatively homogeneous clusters of which the distinctions among them are perhaps incomprehensible. The clusters themselves must have some meaningful relation to the observed phenomena, otherwise, the analysis will not engender confidence. The cuts that resulted in five, ten, and fifteen, clusters are marked in Figure 1. For the present analysis, the author selected fifteen clusters. These comprised first four major clusters, marked by name in Figure 1, and discussed below, and second, the remaining eleven clusters. The latter, identified by upper-case X's, are not significant because they represent isolated derailments e.g. due to a broken rail, whereas the major clusters are train-related. The first major cluster, empty trains that derailed at moderate speed (mean 44km/h) due to slack run in precipitated by a red flag- or signal, was named *Charging Bull*. The second major cluster, 20800-tonne trains that derailed near balancing speed (32km/h) on ascending gradients due to a broken coupler and subsequent intra-train collision, was named *Heavy Drag*. The third major cluster, 18500-tonne trains that derailed at relatively high speed (mean 47km/h) on descending inflections because of bearing failure on the last 100 wagons, was named *Bobsled Thrill*. The fourth major cluster, 20800-tonne trains that derailed on descending inflections at moderate speed (mean 41km/h) because of wheel failure on the first 100 wagons, was named *Roller Coaster*. During feedback to train drivers, they immediately recognized the clusters, and were amazed that unobtrusive observation and ex post facto analysis could yield such accurate insight.

5.4. Factor analysis

Factor analysis studies the correlations among a large number of interrelated quantitative variables by grouping them into a few factors. After grouping, the variables within each factor are more highly correlated with variables within the factor than with variables in other factors [3]. The factor analysis procedure yielded the left two sections of Table 1, namely a listing of the variables and a matrix of their loading on each factor. The author ultimately selected the 38 variables depicted, after a few trial analyses to ascertain which variables gave strong factor loadings, and which seemed to be redundant. Note that the variables belonging to each factor load high on that factor (shown boldface), and low on all other factors. This is a consequence of enhancing the analysis by varimax rotation, to maximize the loading within factors, and minimize the loading among them.

5.5. Interpretation of factors

By virtue of their soft science origin, factors need to be interpreted and named, to give meaning to the latent variables they represent, which in some way underlie the variables that load onto them. Being latent means that interpretation requires insight in the setting through experience and prior insight. The right-hand column in Table 1 shows the names given by the author. They were designed to inspire appreciation of the findings among people who needed to take corrective action. The following executive summary of the interpretations discusses only highlights, because one could devote an entire paper to the subject.

Factor 1, *Train-track Interaction*, speaks to the variables that relate how a driver handles a train over particular topography. The name should come as no surprise. However, contrary to conventional wisdom, the variable Speed does not load onto this factor, an insight that prevents wasting corrective effort on a misguided mission.

Factor 2, *Train-control Interaction*, supports the Charging Bull cluster. The opposite signs between Loaded/Empty and Gross Trailing Load on the one hand, and Red Signal/Flag on the other, indicates that a restrictive signal associates with empty trains and low tonnage. The coincident factor loading of Loaded/Empty and Gross Trailing Load suggests that one or other of them may be redundant. After all, an empty train is light and a loaded train is heavy.

Factor 3, *Wagon Defective*, supports the Bobsled Thrill and Roller Coaster clusters. Note that Wheel Defect and Bearing Defect, on the first derailed wagon, have opposite signs, supporting the mutual exclusivity of these two defects as an essential difference between the Bobsled Thrill and Roller Coaster clusters.

Factor 4, *Track Situation* relates to the presence of a transition and the sharpness of a curve, which from inspection of the clusters is revealed as an issue regarding

Table 1: Factor Loading Matrix

Variable	Factor Loading													Factor Name
kN Braking	**-0.8**	0.3	-0.1	0.0	0.1	0.0	0.0	-0.1	0.0	-0.1	-0.1	0.1	-0.3	1
kN Motoring	**0.8**	0.2	0.1	0.0	-0.1	0.0	0.0	0.1	-0.3	-0.2	0.0	-0.1	0.2	Train-
Slack	**0.8**	-0.1	0.1	0.0	0.1	0.0	0.1	0.1	0.0	-0.2	0.0	0.0	0.0	track
Gradient (per mil)	**0.7**	-0.4	0.1	0.1	0.1	0.0	0.0	-0.1	0.1	0.2	0.2	0.0	0.0	Interaction
Brake Pipe Reduction	**-0.6**	-0.4	0.0	0.2	0.1	-0.1	0.2	0.0	0.3	-0.1	0.1	-0.1	0.1	
Inflection	**0.6**	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.1	0.3	0.2	0.0	-0.2	
Loaded/Empty	-0.2	**0.9**	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.1	0.0	-0.1	0.0	2 Train-
Gross Trailing Load	-0.1	**0.9**	0.2	0.1	-0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0	0.0	control
Red Signal/Flag	-0.1	**-0.8**	0.1	0.3	0.1	0.1	0.2	0.1	0.1	0.0	0.0	0.0	-0.1	Interaction
1st Derailed Car = Jumbo	0.1	0.1	**0.9**	0.0	0.2	0.0	0.1	0.0	-0.1	-0.1	0.0	0.0	0.0	3
Running Number	0.1	0.1	**0.9**	0.0	0.2	0.0	0.2	0.0	-0.1	-0.1	0.0	0.0	0.0	Wagon
1st Derailed Bearing Defect	-0.4	0.3	**-0.6**	-0.2	0.1	-0.1	0.2	0.0	0.2	-0.1	0.1	-0.2	-0.1	Defective
1st Derailed Wheel Defect	-0.4	0.2	**0.5**	0.0	-0.1	-0.1	-0.2	-0.4	-0.2	0.2	0.1	0.2	-0.2	
Transition Present	0.0	0.0	0.0	**0.8**	0.0	-0.1	0.0	-0.1	-0.1	-0.1	0.0	0.0	-0.2	4
Curve Radius (m)	0.1	0.3	-0.1	**-0.8**	0.1	-0.1	-0.1	-0.1	0.1	0.0	0.0	-0.1	0.0	Track
Stock Rail Broken	0.0	0.3	0.0	**0.6**	0.1	0.0	-0.3	0.1	0.2	-0.1	-0.3	-0.1	0.3	Situation
Date (Excel Number)	0.0	0.1	0.0	0.1	**0.9**	0.0	0.0	-0.1	0.0	0.1	0.1	0.2	-0.1	5
Rake 1 Slackless Present	0.0	-0.1	0.2	-0.1	**0.7**	0.2	0.2	0.1	-0.1	0.1	-0.2	0.0	-0.2	Slackless
Rake 2 Slackless Present	0.1	-0.2	0.2	-0.2	**0.7**	0.0	0.0	0.2	0.0	-0.1	0.0	0.0	0.2	Bogeyman
Other Derailed Quantity	0.0	0.1	0.0	0.1	0.1	**0.9**	0.1	0.0	-0.1	-0.1	-0.1	0.2	0.2	6
Other Derailed Gap Present	0.2	-0.3	-0.1	-0.2	0.1	**0.7**	-0.1	0.0	-0.1	0.0	0.0	-0.2	0.0	Richter
Rand Cost	-0.2	0.5	0.2	0.0	0.1	**0.5**	0.1	-0.1	0.1	0.3	0.0	0.2	0.0	Scale
Rail Joint Broken	0.1	0.2	0.2	-0.1	-0.1	**0.4**	-0.1	-0.2	0.4	-0.1	0.2	-0.3	-0.2	
Speed (km/h)	0.1	0.0	-0.1	0.0	0.0	0.0	**-0.6**	0.0	0.1	0.1	-0.1	0.1	0.0	7
kN Instability	0.3	0.2	0.1	-0.1	0.2	0.0	**0.5**	0.0	-0.1	-0.1	-0.1	0.1	0.1	Ride
1st Derailed Coupler Defect	0.4	0.2	0.0	-0.1	-0.2	-0.1	**0.5**	0.0	-0.1	0.0	-0.3	0.1	-0.1	Quality
Day of Week	0.0	0.1	-0.1	0.0	0.1	0.0	-0.1	**0.7**	-0.1	0.2	0.1	0.0	-0.3	8
Gradient Length	0.2	-0.1	0.1	0.0	0.1	0.1	0.4	**0.6**	0.0	0.0	0.1	0.1	0.1	Homeward
Neutral Section Involved	-0.1	0.0	0.3	-0.1	0.0	-0.2	-0.4	**0.5**	0.1	-0.2	0.0	0.0	0.0	Hurry
1st Derailed Position	-0.1	-0.1	-0.3	0.0	0.0	-0.1	-0.1	0.0	**0.8**	0.0	0.2	0.0	0.0	9
Line Voltage Unstable	0.2	0.2	0.1	0.1	0.2	0.1	0.0	0.2	**-0.5**	-0.1	0.4	-0.3	0.1	Trigger
Temp Speed Restriction	0.2	0.2	-0.1	0.3	0.0	0.0	0.3	-0.1	**-0.4**	-0.2	0.3	0.1	0.0	
1st Derailed Spring Defect	0.1	0.1	-0.1	0.0	0.1	-0.1	-0.1	0.2	0.0	**0.8**	-0.1	0.0	0.1	10
Points Involved	-0.1	0.0	-0.1	-0.1	0.0	0.3	-0.1	-0.2	0.1	**0.6**	0.3	-0.2	-0.1	Defective
1st Derailed Brake Defect	0.0	0.0	-0.3	-0.3	0.3	0.1	-0.3	0.0	-0.1	**-0.3**	-0.2	0.2	-0.1	Equipment
Hour of Day	0.1	0.0	0.0	-0.1	-0.1	-0.1	0.0	0.2	0.1	0.0	**0.8**	0.1	0.0	11, 12, 13
Month of Year	0.0	0.0	0.1	0.1	0.2	0.1	-0.1	0.0	0.0	-0.2	0.1	**0.9**	0.0	Not Inter-
Emergency Brake Applied	0.1	0.0	-0.1	-0.2	0.0	0.2	0.0	-0.2	-0.1	0.1	0.1	0.0	**0.8**	pretable

empty trains. It points to a contribution from wheel unloading.

Factor 5, *Slackless Bogeyman*, relates to magnification of the size of a derailment due to the lack of energy absorption in slackless drawgears. Its association with Time relates to the phasing in of that drawgear, which was later eliminated from the wagon fleet.

Factor 6, *Richter Scale*, speaks to the size of a derailment. Noting the mean cost of derailments, Charging Bull at R230 000, Heavy Drag at R360 000, Bobsled Thrill at R3 370 000, and Roller Coaster at R7 000 000 (US$1 = R7.80 at time of writing), leaves no doubt where corrective action ought to focus.

Factor 7, *Ride Quality*, relates to the Heavy Drag cluster. Note that the variable Speed loads on this factor, but with a negative sign, indicating an association with low speed. Unstable tractive effort (kN Instability) and a Coupler Defect on the first derailed vehicle lead to coupler failure near the front of a train, with consequent derailment, on ascending gradients.

Factor 8, *Homeward Hurry*, is the first factor to suggest human responsibility. A high loading on Day of Week associates with a weekend, Gradient Length associates with longish inflections, and Neutral Section Involved indicates hasty train handling that can disturb train dynamics. The name is self-explanatory. Note that Inflection does not load on this factor, but on Train-track Interaction.

Factor 9, *Trigger*, suggests that Line Voltage Unstable and Temporary Speed Restriction tend to influence derailments towards the rear of a train. Note the opposite sign between the foregoing variables and 1st Derailed Position. This factor also suggests human responsibility, apparently a train driver challenged by an unfamiliar or abnormal situation. This factor, and the foregoing one, show the need for personal data, with which to probe what sort of person or circumstances associate with particular types of incident.

Factor 10, *Defective Equipment*, should come as no surprise. Such plain findings provide the researcher with a useful reality check on other interpretations that may have required deep thought.

Factors 11, 12 and 13, are not interpretable because they load only onto themselves, and are therefore not really factors at all. They are thus simple variables that represent no more than their names indicate. One could delete them from the factor analysis without loss of understanding.

6. APPLICATION TO SYSTEMIC DRIVERS

An examination of operational data over a twelve-month period, relating to planned and actual throughput, to loss-causing incidents, and to the performance of mine- and port logistics partners, is being applied on the Sishen-Saldanha system to assess its robustness against disturbances. This study, which requires identification of systemic drivers, exemplifies inclusion of non-incidents, because they are an integral element of system functioning. The notion of a train as case was conceived for this study, because each case needs to terminate somehow, and one may not duplicate a spreadsheet cell entry. The solution was to develop the notion of a terminal incident, which would normally be arrival, either at the halfway crew change or at final destination. However, a break-in-two or derailment, where the train crew and/or -composition may change, is also regarded as a terminal incident. The continuation of the trip is treated as a new train, or new case, for statistical purposes.

The grouped variables include, at high level, Case Identification (date, train number), Train History (train-, locomotive-, and crew attributes, and productive output), Operating History (dispatch particulars, time appointments, and external performance), Delays (due to occupations, alarms, power supply, signals, communications, and operations), Terminal Incident (situation, train handling, equipment, train parting, and/or derailment), and Cost. The variables total 156.

The study will span a full seasonal cycle. Upwards of 2000 cases will be recorded. To record them with minimum effort, the data set is structured such that only relevant variables need a value, but other cells default to zero. It has facilitated exploring incidents in a way that is non-threatening to organizational structures and unobtrusive to field personnel, but yields insight into the latent systemic drivers. It has also driven home the value of recording data into a central data base, because an effort to reconstruct historical data from dispersed sites required three months' work to recover one month's worth of data — clearly a losing battle. The study was ongoing at time of writing.

7. CONCLUSIONS

7.1. Successful outcomes

Factor- and cluster analysis have led to identification of four derailment archetypes, and understanding of the circumstances under which they occur. This understanding underpins a technology upgrade plan, of which a cable-based integrated ECP braking plus distributed power pilot scheme is a major first step. Appreciation that failed bearings and failed wheels occur in similar circumstances has suggested, first, that hot box detectors should be placed near inflections rather than at regular intervals, and, second, that heat due to braking may be the last straw that fails a bearing. The understanding so acquired has enabled management to apply remedial interventions directed at systemic outcomes, rather than departmental posturing.

7.2. Maximizing understanding

Cluster- and factor analysis offer useful tools for understanding heavy haul systemic drivers. Exposure to these procedures generates an appreciation of the value of diligent data accumulation for subsequent scientific analysis. The procedures assist managers both to avoid courses of action that cannot lead to a solution space, and to maximize the understanding of the complexity within spaces in which solutions might exist.

7.3. Data entropy

Given the sensitivity of data to capturing procedures, and the author's caution regarding stillborn statistics, one may ask whether the notion of data entropy has substance. The test might be whether corporate data consistently lead to new understanding, or whether they simply generate descriptive and summary statistics that do not lead to new insights.

7.4. Understanding complex systems

In a heavy haul environment, engineering science yields relatively rigorous understanding of numerate issues, but falls short of describing systemic complexity. Soft science and multivariate statistics yield understanding of systemic drivers despite complexity, but lack the ability to deliver numerate understanding. The author trusts that this paper illustrates the value that soft sciences can contribute in complementing engineering science.

8. ACKNOWLEDGEMENT

The author expresses his appreciation to the anonymous reviewer for the constructive criticism offered.

9. REFERENCES

[1] Kerlinger, F.N. 1986. Foundations of behavioral research. Fort Worth: Holt, Rinehart, & Winston.

[2] Emory, C.W. & Cooper, D.R. 1991. Business research methods. Homewood: Irwin.

[3] SPSS. 1999. Base 9.0 applications guide. Chicago: Author.