CORRELATION AND CAUSATION: A POTENTIAL PITFALL FOR EFFICIENT ASSET MANAGEMENT

J.H. Heyns[1*] and P.J. Vlok[2]
[1,2]Department of Industrial Engineering
University of Stellenbosch, South Africa
[1]hannesheyns@gmail.com
[2]pjvlok@sun.ac.za

ABSTRACT

The successful coordination of activities and practices within a system rely on the organisation's ability to make informed decisions. Decisions must be made quickly and effectively, while ensuring efficient Physical Asset Management (PAM). Access to processed data, in the form of reliable information, on how sub systems interact greatly simplifies decision-making. Many organisations mistake correlation for causation when analysing this data. Such a mistake carries great consequences for organisations, since important decisions might unknowingly be based on self-invented problems, while the true problem is left unresolved. It is crucial to understand the difference between correlation and causation when practising root cause analysis within a PAM environment. Although root cause analysis is presumed a highly specialised field, organisations can equip themselves to better understand how different events within a PAM system are interconnected. If done correctly this might simplify the process of detecting problems, which might exist within a system. This paper highlights the differences between correlation and causation. Potential pitfalls on how correlation can be mistaken for causation within a PAM environment are identify and explained. Recommendations are made on how to avoid these pitfalls.

* Corresponding Author

1 INTRODUCTION

Decision-making is a key component within any Physical Asset Management (PAM) environment. Ensuring efficient Asset Management expects organisations to make informed decisions quickly and thoroughly. Redman [1] suggests that this process often happens in the absence of adequate information or knowledge about the system. Understanding the difference between correlation and causation might help organisations to avoid some of these potential pitfalls which might occur during Asset Management decision making.

1.1 Background

Access to the right data can greatly facilitate decision making within a PAM environment. If used in accordance with ISO 55000 [2] and PAS 55 [3] asset data can be transformed into asset information, which might help organisations to make more knowledgeable decisions.

Although data may be of aid in asset management, Quinlan [4] believes it can also have negative consequences if not handled with care. Data used by organisations are often insufficient to substantiate important decisions [1]. When studying data in isolation it is difficult to determine whether all required information is available. Resultantly, organisations might base important decisions on vague interpretations of data, due to a lack of a better understanding of the problem at hand.

1.2 Problem Statement

According to Card [5] distinguishing between correlation and causation is not as intuitive as many might believe. For data to be of aid in decision making within an Asset Management environment, it is essential to ensure correct interpretation of this data. Many organisations mistake correlation for causation when analysing data [6]. Consequently decisions are often unknowingly based on self-invented problems, while the true problem is left unresolved.

Misinterpretations of data during root cause analysis may nullify any effort to develop substantial solutions to a specific problem. On the contrary, this can easily lead to creating an even bigger problem than before. It is therefore important for organisations to understand why it is important to distinguish between correlation and causation, know what this distinction is and ensure that they are capable of applying this knowledge for more efficient Asset Management decision making.

1.3 Objectives

This paper intends to

- Explain the difference between correlation and causation
- Identify and explain potential pitfalls where correlation can be mistaken for causation in accordance to a case study of which data is available
- Make recommendations on how to avoid these pitfalls

This paper is based on the philosophy behind root cause analysis, where the same approach has to be taken as in the legal world where a suspect is *innocent until proven guilty*.

2 CORRELATION VERSUS CAUSATION

In a statistical environment, *Correlation* describes a relation between different and separate events, where these events show a tendency to vary simultaneously. At a quick glance it is easy to assume that these events are linked and that the behaviour of one event has an effect on the behaviour of another. This however is not the case. Although it is important for such a correlation to exist, it is not possible to establish causality from correlation alone [7]. *Causation* on the other hand, describes a cause and effect relationship between events. The

behaviour of an event therefore directly affects the behaviour of other. *Root cause analysis* is the process of looking for the root cause which induced a specific effect.

2.1 Requirements for Causation

Many different approaches can be followed to prove causation [8] [9] [10] [11], but not all problems have the same nature. Therefore it is difficult to narrow root cause analysis down to one generic approach which would suite all problems. Card [5] identified three general requirements to facilitate this process.

- Correlation should exist between cause and effect
- Cause should precede effect
- Linking mechanism between cause and effect must be identified

If these requirements are satisfied, a causal relationship between events has been demonstrated.

2.2 Distinguishing between Correlation and Causation

Card [5] further explains the difference between correlation and causation efficiently through an example where people spends time reviewing a document to detect defects such as spelling or language errors. The results for this experiment are shown in Figure 1. It is seen that there exist a definite correlation between the hours spent reviewing and the defects found.

At first it might seem that there exist causality between the two events. Time has to be spent to find defects. The results suggest that if this time is spent then defects will be found. Can this therefore be defined as a cause and effect relationship?

In reality the reviewers will reach a threshold where after they will not likely find more defects, even if they spend infinite time trying. It is also important to note that the action of reviewing and finding defects happens simultaneously. This example fails to adhere to the second requirement and therefore it is evident that there exist no cause and effect relationship between the time spent reviewing the document and the amount of defects found.

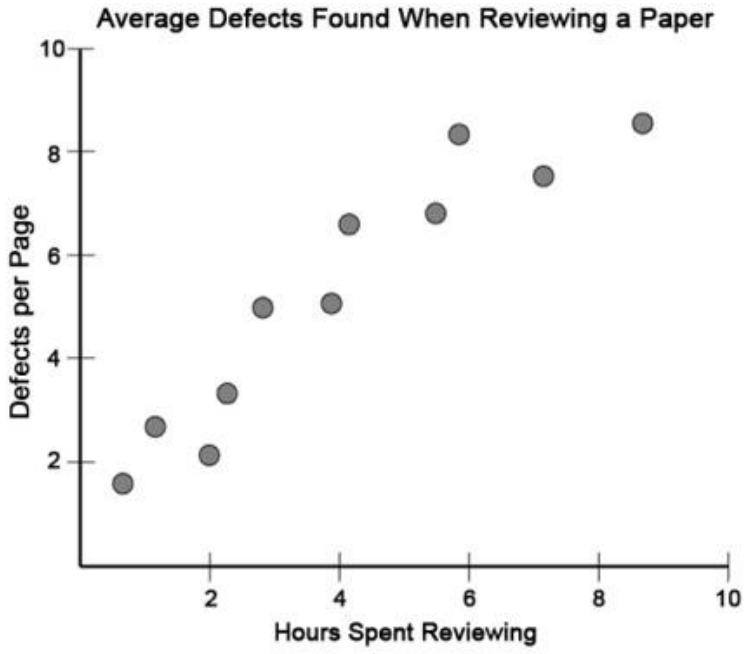

Figure 1: Example of correlation between variables as described by Card [5]

Once explained it is easy to understand that there is no causal relationship between the events described above. Note that this is a very simple example and in reality PAM systems tends to be much more complex. Therefore, root cause analysis requires an in depth understanding of the system to be able to properly define a problem and hopefully help find the root cause for this problem [12]. Only once the root cause has been established can solutions be generated on how to solve these problems.

2.3 Root Cause Analysis within a PAM Environment

All problems are defined by the events which caused the problem. Therefore to be able to solve a problem, the cause which created the problem has to be identified and then strategies can be developed to prevent these causes from reoccurring [12]. Root cause analysis is therefore a very important tool which can be used in a Physical Asset Management environment to find efficient solutions to problems which exist in complex systems.

The following sections will take a systematic approach on how a mining company, practising Asset Management, might go about looking for the root cause for component failure within large mining machinery used on different mines. Pitfalls, where correlation might be mistaken for causation during this process, are identified and explained. Recommendations are made on how to avoid these pitfalls.

3 IDENTIFYING CORRELATION

The first step to prove causality between specific events requires ensuring that a correlation exist between the events which are investigated. Rodgers and Nicewander [13] suggest various methods to establish correlation using the famous Pearson's product-moment correlation coefficient, but regardless of the methods used to establish correlation, this step requires access to relevant data.

3.1 Ensure Data Integrity

Data is an important tool used during decision making within a PAM system [3], but the use of incomplete data for decision making is very dangerous. Crucial information might be missing from the data and therefore basing decisions on this data can lead to inferior solutions to be developed. When this is the case, a problem can become data specific and correlations between events may easily be mistaken for causation. Data integrity must therefore be ensured before data mining can begin.

Organisations should try to ensure, as far as possible, that all needed data is obtained to efficiently help find the root cause for a specific problem [12]. This process might be made easier if sound strategies are followed to record and store data. The use of appropriate unique keys and identification codes can help to better structure data and help facilitate future data analysis. When planning which data will be recorded within a PAM system it might help to implement a virtual root cause analysis on critical components within the system. This should ensure that the organisation have access to all the required data for when a problem do occur.

3.2 Pattern Recognition

Data can be analysed to find reoccurring patterns or specific trends. It might be possible to argue that a repeating pattern within a dataset might also be seen as correlation of some kind. For this to make sense within this context the data will have to be split in two or more parts to be able to compare the repeating data from different time intervals with one another. Note that although this is a very important part of data mining, it is very difficult to prove that the behaviour of data in one interval directly influence a following pattern even if correlation does exist. Usually trends are the results of other external factors and are not specifically dependant of previous behaviour. Therefore it is important to take note of these trends, but they cannot be used to prove causality.

3.3 Matrix Scatterplots

When analysing data with the purpose to find a root cause for a specific PAM problem it is important to start by looking for correlation between events. This will help identifying areas which should to be further investigated for potential causality. Usually this includes filtering through large datasets, which can easily become a tiresome and frustrating process if all variables are investigated separately.

The use of matrix scatterplots is a very effective method to easily filter through large datasets in search of correlation. Figure 2 shows a matrix scatterplot of data which recorded the fitting and defitting information for components during maintenance on large mining machinery. As highlighted, this diagram highlights various locations where potential correlation might occur. These regions show that one set of data change as another varies. Simultaneously it shows which variables can be ignored in the search for correlation.

This is also an effective method to identify whether there are data entries which might be faulty. For instance Var 3, the component code and Var 9, the component description from Figure 2 should show a perfect linear correlation. This is not the case and therefore it is reasonable to assume that there are potentially faulty entries present in the data.

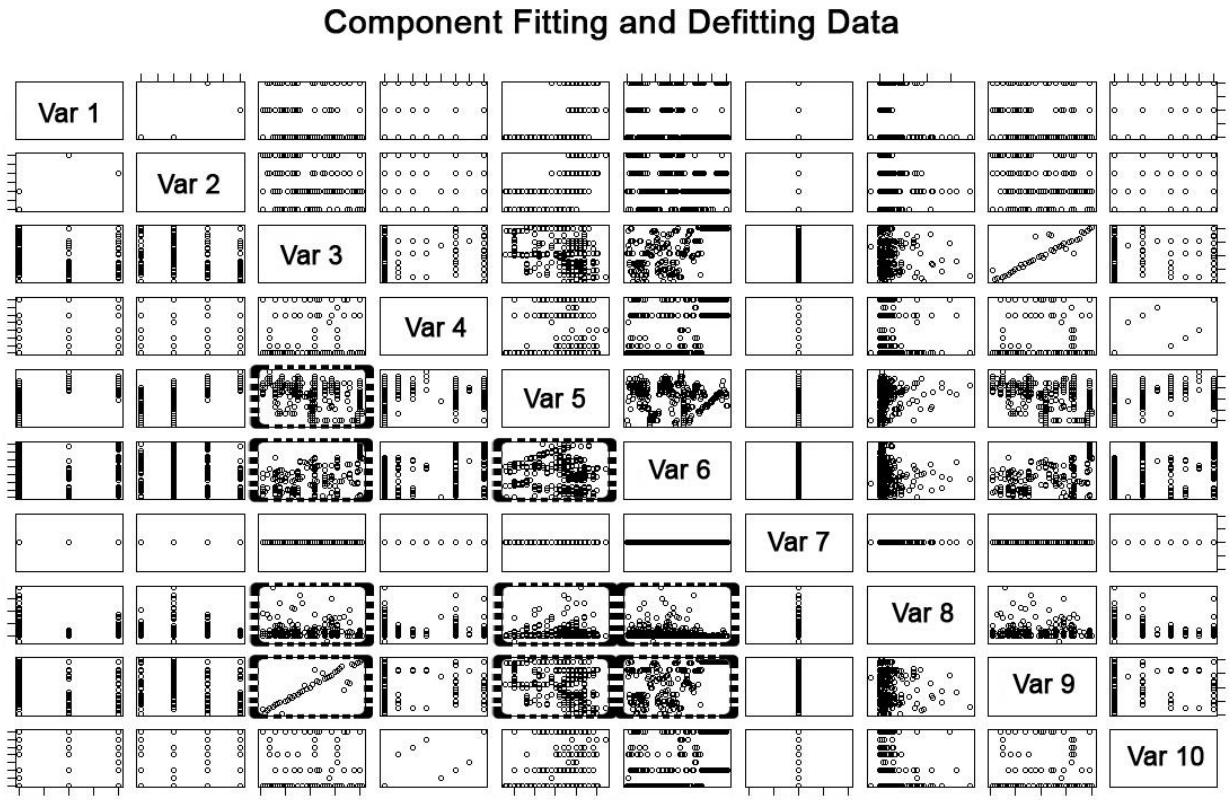

Figure 2: Matrix scatterplot for data from specific machine at specific mine

3.4 Data Comparison

Once events where potential correlations might exist have being identified, these events can be isolated for a further investigation. Figure 3 illustrates the life span of a specific component, from a specific mining machine, on four different mines. If the component fails it is replaced or repaired. Each repair or replacement is represented as a dot on the graph. Wherever a component life of zero is indicated, it is evident that the data is corrupt at this point. It is difficult to pick up any specific pattern in the component lifetime via inspection alone. Nonetheless, it is evident that the components on different mines varied more or less within the same range.

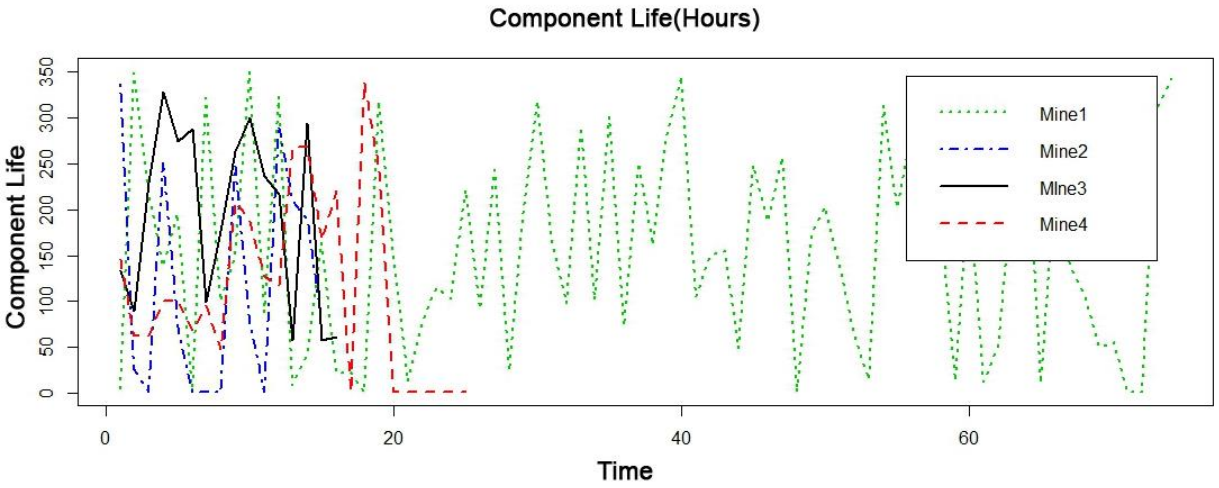

Figure 3: Comparing component life from different mines

This suggests that the component failure is most likely not location specific and is therefore independent of a mine's climate and the product handled by the machines. This might rule out or enforce the possibility that a component failure might be due to material failure such as corrosion. Data does not always provide answers in the format which is expected. It is the responsibility of the organisation to develop sound strategies to help obtain relevant information. To ensure efficient PAM strategies an organisation should be able to think of the right questions to ask and apply the required expertise to answer these questions sufficiently.

4 CAUSE SHOULD PRECEDE EFFECT

Once correlation between events within a PAM environment has been established, the next step to prove causality is to ensure that the cause precedes the effect.

4.1 Causal Tree

An efficient technique to ensure that the cause precedes the effect is by identifying the potential causes which might have caused the problem. A causal tree is an efficient method to visualise this information [14]. Figure 4 shows an initial causal tree showing potential causes which might have caused component failure in a mining machine. Such a tree has to be expanded through an iterative process as far as possible. During root cause analysis new findings should enable an organisation to improve such causal trees. This does not only help during root cause analysis on the current problem which is investigated but also for future problem detection and should also help to improve future designs for future components.

4.2 Cause Before Effect

As seen in Figure 4 the component failure can potentially be caused by six different events defined by Bloch [15].Each of these events are then further expanded into the possible causes which might have induced these six events. All events are repeatedly expanded until all possible causes for the component failure have been identified.

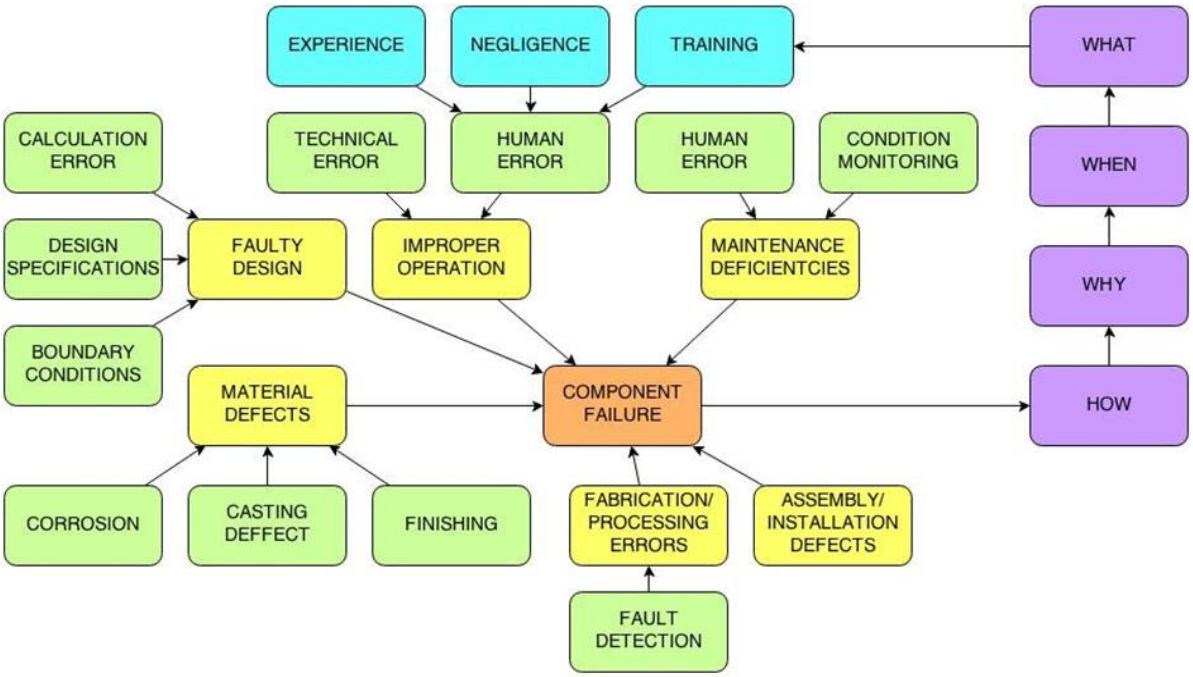

Figure 4: Causal tree for component failure

When two events are investigated to determine the order in which they occur, the causal tree can be used. The higher up the hierarchy an event occurs, the later it happened. For instance, if component failure is compared to the training a driver received before operating a mining machine, it can be seen that component failure may occur due to lack of training, this is a human error resulting in improper operation and therefore component failure occurred. It is not possible to argue that a component failed, which resulted in the driver being inadequately trained for the job, since the driver was trained before the component failed.

5 LINKING MECHANISM BETWEEN CAUSE AND EFFECT

After correlations between events have been established and it can be shown that the effect precedes the potential cause, a linking mechanism has to be identified to prove causality.

In reality identifying the exact root for an event within a complex PAM system is not always an easy task. Initially the only know certainty is that a problem has occurred, but the nature of the problem is not necessarily known. To find the root cause, this problem has to be characterised. Simultaneously potential causes have to be identified and linked to the problem [12]. This involves an iterative process of investigating all potential causes to establish what the problem is, when it happened, why it happened, how it happened and if it contributed to realising the problem. In many aspects this requires completely the opposite approach as was taken in the previous section where the causal tree in Figure 4 was developed.

5.1 Forcing a Cause to fit an Effect

At this stage a common mistake which might be made during root cause analysis involves identifying a potential cause for an effect. This cause is then isolated for further investigation. It is easy to argue that a predefined potential cause is destined to cause a specific effect. Therefore organisations might argue that since the specific effect did happen, it is due to the predefined potential cause. For instance, if you strike a match, there will be fire. Does this mean that if there is fire, it was caused by a match? It might even have been caused by a combination of different events, maybe a match and wind.

A potential cause can easily be forced unintentionally to match an effect. Especially when in a large and complex PAM system where there are little data available to support decision making. It is therefore important to note that root cause analysis tries to mine useful information from complex and interconnected systems, where different events can rarely be completely isolated from one another.

5.2 Understanding the Bigger Picture

Another mistake which might be made when analysing problems in a PAM system, is to isolate subsystems from the goals and philosophies which defines the greater system. Within complex systems there are many different factors at play which indirectly influence a problem and therefore its causes [16].

Consider a component failing on some large mining machinery. This component belongs to a machine, which has a certain purpose on a mine. The mine is part of a specific industry which again is driven by a country's economy. The country's economy forms part of the global economy, which is defined by the global availability of resources and the human need for certain commodities. All of these factors are interrelated and therefore important to consider when making decisions to implement sound PAM.

Although the current position of the global economy do not influence whether or not a gearbox fails within a machine, it might influence the strategies which can be implemented to fix the problem. If the global economy is down and the stock market drops, it is likely that the mine may be financially influenced by the situation.

If a specific mining machine breaks down and the machine is crucial in ensuring the mine's performance, it is important to resolve the problem as quickly as possible. This might involve applying more or less resources to root cause analysis, depending on the criticality of the problem. Root cause analysis strategies might vary and therefore also influences PAM strategies. For efficient Asset Management it is therefore important to know how all events within a PAM system is interconnected and whether there exist a correlating or causing relationship between these events.

It is a good idea to visualise these connections for a better understanding of the relationships between events. Figure 5 visualises the internal and external relations between events and parties involved within a mine. Once the workings of the larger system are known, problems occurring in different locations can be identified, defined and further investigated

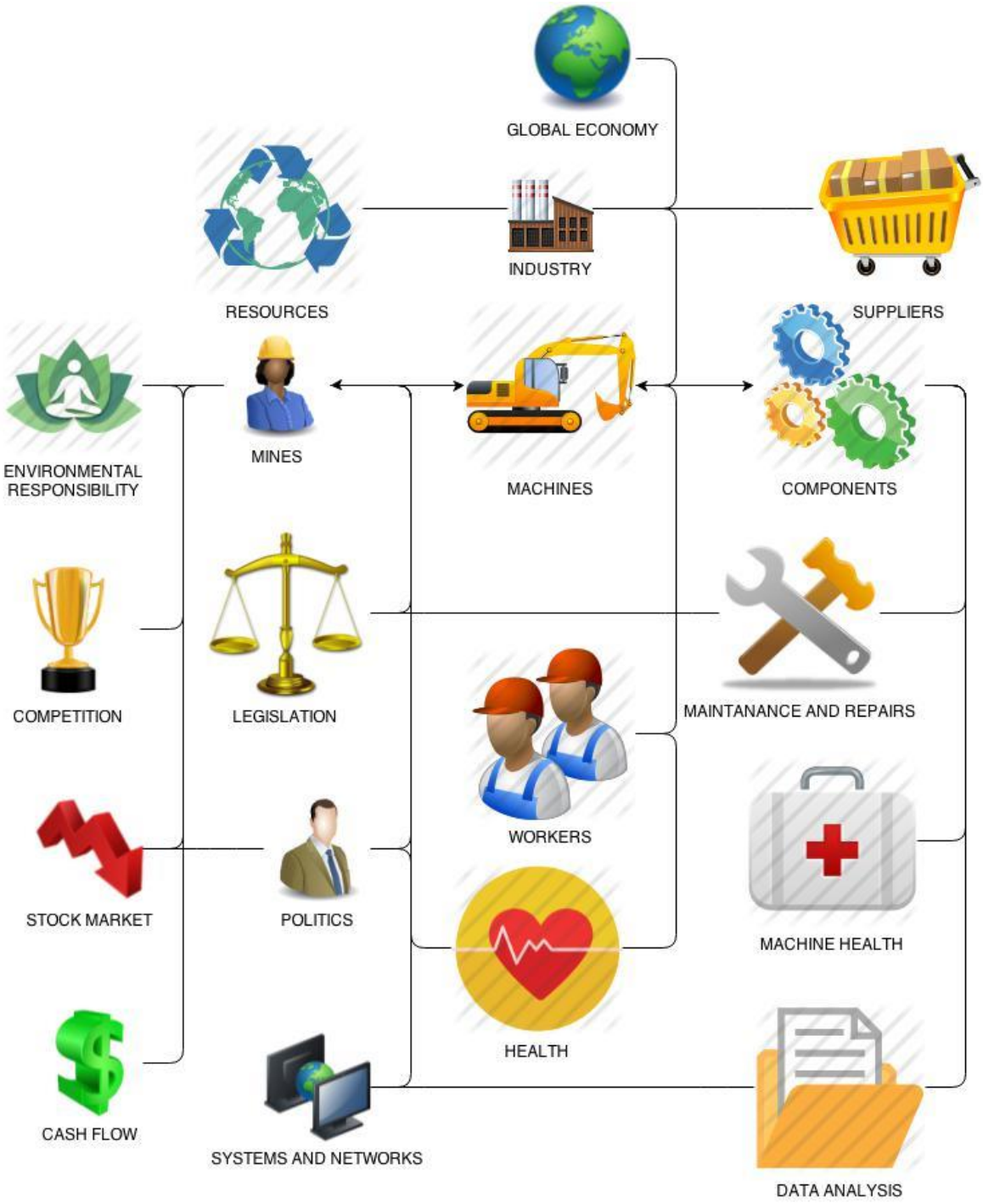

Figure 5: External and internal influences of a mine

5.3 Finding the Root Cause

Only once a problem, which may arise within a PAM environment, is fully characterised and understood, may the search for the root cause begin. As mentioned earlier there are various techniques which can be used to find linking mechanisms to fully prove causation.

Despite the various techniques and tools available it remains important for organisations to truly know the ins and out of a system and simultaneously be able to implement critical thinking, to help find the linking mechanism which will prove causality for a problem within a PAM system.

6 RECOMMENDATIONS

With the complexity of PAM systems ever increasing, it might be important for organisations who take Asset Management seriously, to design their systems for improved root cause analysis. This includes identifying connections between different subsystems and predicting how potentially unwanted behaviour within a system might influence behaviour in other subsystems.

Designing and building a root cause analysis strategy into a PAM system will aid failure detection when problems occur. Root cause analysis is not only meant to be implemented as a maintenance strategy, but can also help preventing potentially unwanted behaviour in the future.

Due to the large infrastructure of interconnected subsystems and events present in a PAM system, it is easy to confuse correlating events for causality. It is therefore recommended to keep the three general requirements as described by Card [5] in mind when trying to demonstrate a causal relationship between events.

Relevant and accurate data is an enormous aid for efficient decision making. Ensure that all data used is correct and factual. Develop sound strategies to ensure that data is properly recorded and user friendly as proposed by Baker [17]. Once the integrity of the data can be proven the search for correlation between events can start.

Use causal trees to break events up into their potential causes as far as possible. This will help determining whether a cause precede an effect and furthermore help to better understand where to investigate when looking for the root cause of an event. Potential weak points within a system can also be identified for later improvement of the system.

When searching for a linking mechanism to prove causality, the goals and philosophies of the larger system should guide the strategies which are developed for root cause analysis. No event which forms part of a complex PAM system, can be viewed in complete isolation.

Root cause analysis requires an organisation to have the same mentality as in law where a suspect is *innocent until proven guilty*. During the process of root cause analysis information is the greatest tool to assist in finding solutions to problems. Use this information wisely, within the systems context, to narrow down problems and find their causes. Be aware of the potential pitfalls when mistaking correlation for causation and ensure PAM strategies which will avoid these pitfalls.

7 CONCLUSION

Due to its wide application, root cause analysis is a topic, which is intensively researched by many. Organisations strive towards improved efficiency, simultaneously systems tend to be more complex than ever before and modern legislation adds to the complications experienced in these systems. Despite trying very hard, no generic tool has been developed to search for and identify causality within a system. For this reason it is important for organisations practising Physical Asset Management to avoid mistaking correlation for causation during root cause analysis.

As shown in this paper the process of finding the root cause to a specific problem is not always easy. Complex systems tend to disguise crucial information which is needed for efficient decision making. The combined knowledge of how events within a PAM system are connected, guidelines for sound PAM such as PAS 55 and ISO 55000, relevant data and knowing how to interpret the data accurately, should help organisations avoid potential pitfalls which might occur from mistaking correlation for causation and to ultimately practise more efficient Physical Asset Management.

8 REFERENCES

[1] **Redman, T.C.** 1998. The Impact of Poor Data Quality on the Typical Enterprise *Communications of the ACM,* 41(2), pp 79-82.

[2] **ISO 55000,** 2013. International Standard, *Asset management - Overview, principles and terminology,* pp 8.

[3] **PAS 55,** 2008. Publicly Available Specification, *Part 1: Specification for the optimized management of physical assets,* pp 16.

[4] **Quinlan, J.R.** 1990. Decision Trees and Decisionmaking, *IEEE Transactions on Systems, Man and Cybernetics,* 20(2), pp 339-346.

[5] **Card, D.N.** 2006. Myths and Strategies of Defect Causal Analysis, Pacific Northwest.

[6] **Saghaian, S.H.** 2010. The Impact of the Oil Sector on Commodity Prices: Correlation or Causation?, *Journal of Agricultural and Applied Economics,* 42(3), pp 477–485.

[7] **Yee, A. S.** 1996. The causal effects of ideas on policies, *International Organization,* 50(1), pp 69-108.

[8] **Jayswal, A., Li, X. Zanwara, A., Loua, H.H. and Huangb, Y.** 2011. A sustainability root cause analysis methodology and its application, *Computers and Chemical Engineering,* Vol. 35, pp 2786-2798.

[9] **Lehtinen, T.O., Mäntylä, M.V. and Vanhanen, J.** 2011. Development and evaluation of a lightweight root cause analysis method (ARCA method) – Field studies at four software companies, *Information and Software Technology,* Vol. 53, pp 1045-1061.

[10] **Wright, R.W.** 1985. Actual Causation vs. Probabilistic Linkage: The Bane of Economic Analysis, *Journal of Legal Studies,* Vol. 14, pp 435-456.

[11] **Pearl, J.** 2003. Causality: Models, Reasoning and Inference, *Econometric Theory,* Vol. 19, pp 675-685.

[12] **Andersen, B. and Fagerhaug, T.** 2006. *Root Cause Analysis: Simplified Tools and Techniques*, 2nd Edition, Quality Press, Milwaukee.

[13] **Rodgers, J. and Nicewander, W.** 1988. Thirteen Ways to Look at the Correlation Coefficient, *The American Statistician,* 42(1), pp 59-66.

[14] **Andre, B.** 1991. Computer-aided fault tree synthesis I (system modeling and causal trees), *Reliability Engineering & System Safety,* 32(3), pp 217-241.

[15] **Bloch, H. P.** 2005. Successful Failure Analysis Strategies, *Reliability Advantage Training Bulletin,* Vol. 3.

[16] **Urry, J.** 2002. *Global Complexity*, Polity Press, Cambridge.

[17] **Baker, R.** 1998. *Managinga Data Warehouse*, Veritas Software Corporation, Chertsey.