THE FUTURE OF TRANSPORT

The best approach to unlock the value from your datasets

Statistical modelling or machine learning?

Dr Ciaran Ellis , Caroline Wallbank, Dr Mark Bell

Introduction

The need to understand data in transport is not a new phenomenon; data on road casualties in Great Britain have been collected and published since 1926. This dataset, and others like it, have been used for many decades by road safety practitioners, governments and highway authorities to monitor developments in road safety.

Today, there is a huge volume of transport datasets in existence. These are generated by travel behaviour, opinions, sensors embedded in the environment, use of mobile apps and through surveys.

While data on road casualties are still collected for much the same purpose, many of the other datasets are collected for very different reasons. Each of these datasets plays a vital role in supporting transport system operation, but their use in combination, or for purposes other than the one for which they were originally collected is relatively uncommon. Restrictions on data sharing and large, unstructured data sources are just some of the issues limiting the insight which can be gained from these data.

The good news is that we are increasingly moving towards a world where everything is connected through the Internet of Things (IoT)[1], meaning an explosion in both the availability and richness of data. Computing power and bandwidth increases also mean that many datasets are available in real-time. With a wealth of information comes opportunity – access to big, open and connected data means that we have the ability to solve increasingly complex transport challenges.

Smart travel cards, sensor equipment embedded in the road environment, location tracking through Global Positioning Systems (GPS), mobile networks and data collected through social media and apps are some of the thousands of datasets being used every day to inform operation of the transport network.

1| A network of devices embedded with sensors which connect and exchange data using the existing Internet infrastructure.

Data are an increasingly valuable asset. Data are being used to improve public services, increase efficiency, reduce costs, develop new products and services, and for evaluation purposes. However, without sound analysis and interpretation, the value of data is likely to be lost. There is a growing demand for data science[2] capabilities to support analysis and interpretation. There are many tools and techniques that can be used for these activities.

Predictive models attempt to explain how historical data can be used to predict the future, or to evaluate outcomes. Within data science, two of the main approaches for predictive modelling are statistical modelling and machine learning. Traditionally most predictive models have been statistical models, chosen by humans to fit the data according to a rigorous set of assumptions.

However, machine learning is changing this and offers a framework for fitting models to data using algorithms, allowing models to be built from large and unstructured datasets. While statistical modelling and machine learning are both used in predictive analysis, the focus of these approaches is different.

Choosing the best approach for the question in hand will save a great deal of time and money, and lead to better predictions, whether the question relates road safety, in-car experience or journey planning. There are therefore important questions to ask when comparing the different approaches of statistical modelling and machine learning; how do we know which approach to use for the best results for predictive modelling, and what are the main factors that influence this decision?

This paper aims to answer these questions, and presents some case studies examining applications of these techniques in the transport sector.

2| Broadly considered to be the union of analytics, business and computer coding skills.

What is predictive modelling?

Predictive modelling is the use of a mathematical model to generalise a dataset to make predictions about potential outcomes. Predictive models can be used to inform what might happen on a future date (the weather next Tuesday) or how outcomes may vary under changing conditions (what happens to air quality if traffic increases by 150%?)

There are many transport problems where accurate predictive modelling is desirable. For some of these problems, the challenge lies in using data to understand the current situation, to predict a future network state and take the appropriate action.

Examples of predictive modelling

With IoT and Vehicle-to-Everything (V2X) communication[3], there is an increasing abundance of real-time data generated within the road environment. This is set to increase markedly over the next decade as connected and automated vehicles develop, and even as driver assistance systems proliferate. For example, Advanced Driver Assistance Systems (ADAS) have been developed to use sensors to detect hazards in the environment, predict the likely course of that hazard, and if necessary, alter the vehicle's behaviour in response. One example is Autonomous Emergency Braking (AEB) for vulnerable road users which can detect a pedestrian entering the carriageway, predict the movement of the pedestrian and apply the vehicle's brakes to avoid a collision if one is deemed to be likely.

In a future world where Vehicle-to-Pedestrian (V2P) communication is common, the vehicle could also provide a warning to the pedestrian through the use of a mobile device. All of these data exchanges require the systems to be interpreting the data in near real-time and generating predictions.

Automated vehicles are one example of IoT application: fully driverless vehicles will be embedded with technologies which allow them to communicate and interact over the internet. This interaction will include other vehicles, the infrastructure around the vehicle and other road users such as pedestrians.

Predictive analysis has many other use in transport. For example, predictive models have been used for:

- Forecasting road traffic casualties
- Predicting when and where the road network is likely to be busiest so that safety resources such as Traffic Officers or gritting can be distributed effectively
- Forecasting traffic levels in order to set appropriate signs and signals
- Modelling the impact that a new junction lay-out will have on vehicles on the surrounding roads.

3| A vehicular communication system that incorporates other types of communication including V2I (Vehicle-to-Infrastructure), V2V (Vehicle-to-Vehicle), V2P (Vehicle-to-Pedestrian), V2D (Vehicle-to-Device) and V2G (Vehicle-to-Grid).

Statistical modelling and predictive analysis

What is statistical modelling?

The statistical modelling process usually attempts to answer specific questions about a dataset that are set prior to analysis. For example, what are the main factors that contribute to road casualties? The main focus of the analysis is on inference – attempting to gain an understanding of the underlying process that generated the data, and then using this understanding to answer the specific question(s) of interest.

During each stage of the analysis, any decisions should be justified by sound evidence and fully documented. This includes reporting all model assumptions and providing evidence that they are valid. Some of these model assumptions may involve the shape of the data. Here, the data in question is fit to a theoretical distribution from a suite of distributions which are known to be representative of real-world processes. For example, if an assumption of the model is that the data has a Normal distribution (bell-shaped curve), diagnostic tests for this should be performed.

Overall the statistical analysis should rigorously demonstrate the suitability of the chosen model for that particular dataset. Fitting an accurate model to the dataset, and demonstrating the validity of this model, are key objectives. This model is then used to provide the statistician with an understanding of the process that generated the data, enabling them to answer the specific research question or questions of interest. For example in Case Study 1, it was crucial to understand the relationships that were present in the data, in order to understand how population changes may influence casualty trends in the future.

Statistical models are also able to perform prediction. However this is just one aspect of the modelling process, and it is not always of interest in the analysis.

Case study 1

An application of statistical modelling to predict road casualties

It may be desirable to understand the relationship between road casualties and the factors which contribute to them, predicting how the number of casualties may change if these factors change over time. Road casualties do not follow a normal distribution and as a result, it is necessary to fit these to a different distribution. Commonly the Poisson or Negative Binomial distributions are chosen instead, as these are a better fit for count data in this domain.

TRL has developed predictive models to understand the relationship between the number of casualties, the amount of travel and demographic variables including age and gender. These models have then been used to understand how changes to the population (and subsequently to the amount of travel) may influence casualty trends in the future, and to set casualty reduction targets.

Statistical modelling: the upsides

Due to its focus on inference, a key benefit of statistical modelling is the depth of insight that can be gained. A greater understanding of both the process that generated the data and any relationships that are present can be a powerful tool in decision making.

The outputs of a statistical model can be clearly linked to the inputs, so the relationship between the input parameters and the response variable (output) can be understood easily. This delivers insight into what drives changes in the variable of interest, leading to useful information to drive better policy or business outcomes. For example, in recent work for the MOVE_UK project[4], TRL evaluated signals collected by vehicles as they passed road signs, in order to determine whether they had detected the sign or not. From the structure and output of the model we could tell whether particular factors had a significant influence on detection ability, as well as the size and direction of that effect. This set of outcomes enables clear recommendations to be made on the basis of model outputs.

Statisticians are able to test hypotheses and check whether results are meaningful given the natural variation expected. This is crucial to the analysis of trial data, where limited data are collected in order to evaluate whether an intervention has worked or not.

The same approach works equally well when testing hypotheses derived from non-trial datasets. For example, for the MOVE_UK project, we were able to determine whether the likelihood of a road sign being detected was significantly higher on some road types compared with others, and what other factors increased the likelihood of detection.

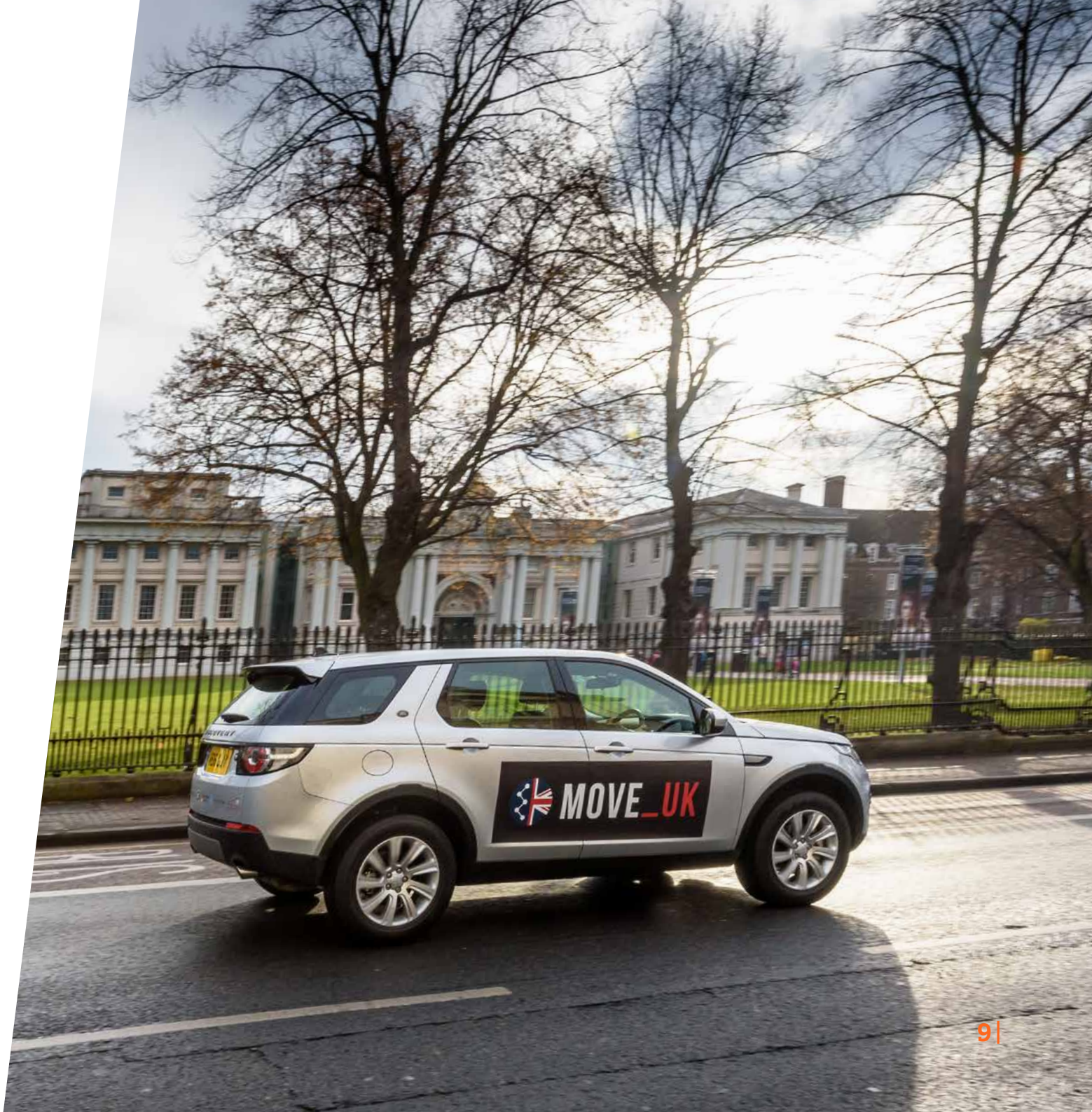
Any outputs of a predictive model are subject to random error, and in some cases potentially also bias. Statistical modelling is based on rigorous mathematical concepts, which aim to minimise these. The rigour required to generate working statistical models means that there is a strong requirement in statistical modelling to understand, examine and clean the data before use.

4| MOVE_UK is a collaborative project involving six leading UK organisations including TRL. It seeks to address key barriers to the development of autonomous vehicles through the development of methods to validate Automated Driving Systems (ADS). This involves advanced analysis of "event-based" data collected through real world testing.

This process can be very useful to:

- Discern whether data sources are generating usable data (or nonsense)
- To correct any poorly set up or incorrect data feeds and fields
- To see any initial patterns between the variables of interest

Following this process allows any errors to be fixed before modelling begins and determines whether the data used are up to the job.



Statistical modelling: the downsides

The application and checking of assumptions can make statistical modelling a laborious process that requires a high-level of expertise. In order to gain sufficient expertise, a number of years' experience and/or higher level qualifications are usually required.

Some modelling assumptions may be easy to validate, but others may not be. For example if the data fit a common distribution then the checking process is straightforward, but often real-world data does not fit neatly into a statistical definition. If this is the case, there are options available within statistical modelling, including transformations of the data, to provide a better fit. Nevertheless, an exact fit to statistical modelling assumptions may be difficult or impossible for a particular dataset.

Although the following is not recommended, to get around this issue, statistical models are sometimes used without validating all of the necessary assumptions. Some statistical tools (such as ANOVA[5]) are relatively robust to this, but others are not. The more modelling assumptions that are violated, the less reliable the results will be. If statistical assumptions are not fully met, there is a risk of overconfidence in results that do not warrant it, leading to incorrect policy or business decision being made.

Due to the numerous assumptions and rigorous approach that are required for statistical modelling, there can also be less flexibility in terms of the modelling choices that are available.

What other options are there?

Statistical modelling offers a rigorous approach to predictive modelling, with a rich range of possible insights, but there are some shortfalls. Statistical models can involve a high level of expertise and fail to deliver if a set of assumptions cannot be met, but what can be done to overcome these?

Most of the problems with statistical modelling stem from the rigid requirements to fit the data beforehand to a distribution. This was traditionally done as a way of using our prior knowledge of what big datasets (populations) look like, to infer what is happening with a small dataset. With the explosion in the size, storage and processing of datasets, there are a new suite of tools being developed which work on large datasets without reference to these same distributions. Many of these tools can be described as machine learning.

5| ANOVA or Analysis of Variance is used to compare the mean score on a continuous variable between three or more groups or conditions.

Machine learning and predictive analysis

What is machine learning?

The term machine learning refers to the process of using a computer algorithm to find patterns in data and generate predictions. Machine learning uses many tools from statistical modelling – so there is some overlap in terms of the techniques that are used. However, the key difference to statistical modelling is the focus of the approach. Machine learning algorithms have a strong focus on predictive power – and are assessed almost entirely on this. This gets around the necessity for a strict set of statistical assumptions to be met by using an alternative criterion to measure model success.

Investigating the predictive power of a model involves the following process. The data are split into three sets: the training, validation and test sets. Different models are fitted to the training data and these models are then tested on the validation data. When a final model has been selected, its predictive power is then assessed by fitting it to the (as yet unseen) test data. A commonly used formal method for this process is called cross-validation. This process is sometimes performed for statistical models too.

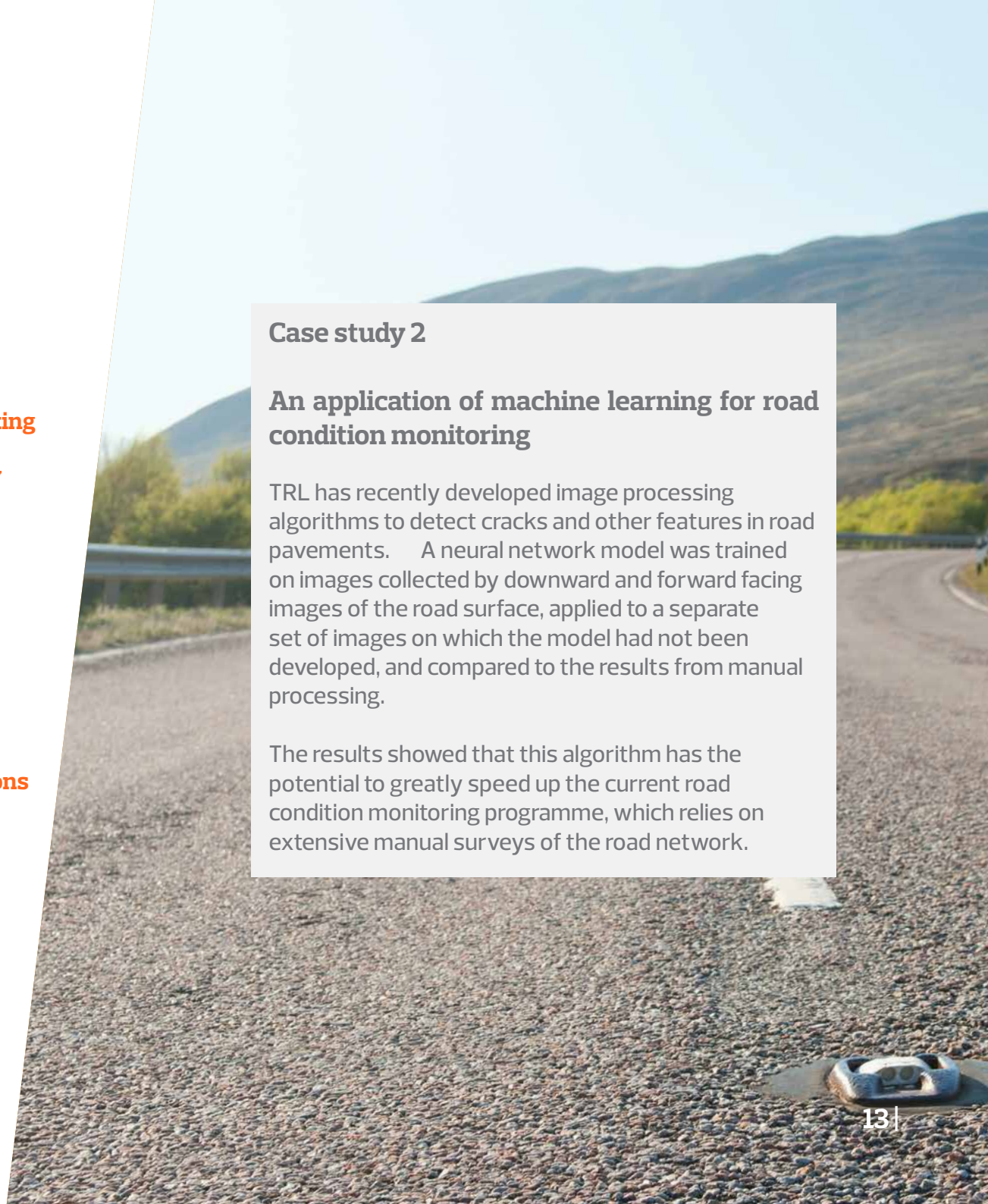
As in statistical modelling, exploratory analysis and pre-processing of data is also usually performed in machine learning, but the main objective of this is to get the data in the required form for the chosen algorithm. Unlike a statistical model, machine learning techniques do not attempt to understand the process that generated the data or the relationships that are present. In addition, there is no attempt to validate the model in the way that statistical models are validated. Success or failure of a machine learning algorithm is determined solely by its predictive power.

A model with a strong predictive ability can be a very powerful tool. In Case Study 2 for example, the ability of the algorithm to predict the road condition from the image data has the potential to make considerable savings in terms of time and costs. In this study, the predictive power of the algorithm was more important than attempting to understand the exact nature of the relationships between cracks and road condition, which would have been very difficult to quantify.

Practical examples of machine learning include:

- **medical diagnosis (through spotting patterns in images)**
- **object recognition such as licence plate reading and tracking**
- **natural language processing such as voice recognition by Amazon's 'Alexa' assistant**
- **lists of recommendations based on previous purchase or search history**
- **fraud detection.**

Many of these algorithms work in real-time, evaluating thousands of data points simultaneously and making decisions in a fraction of a second.

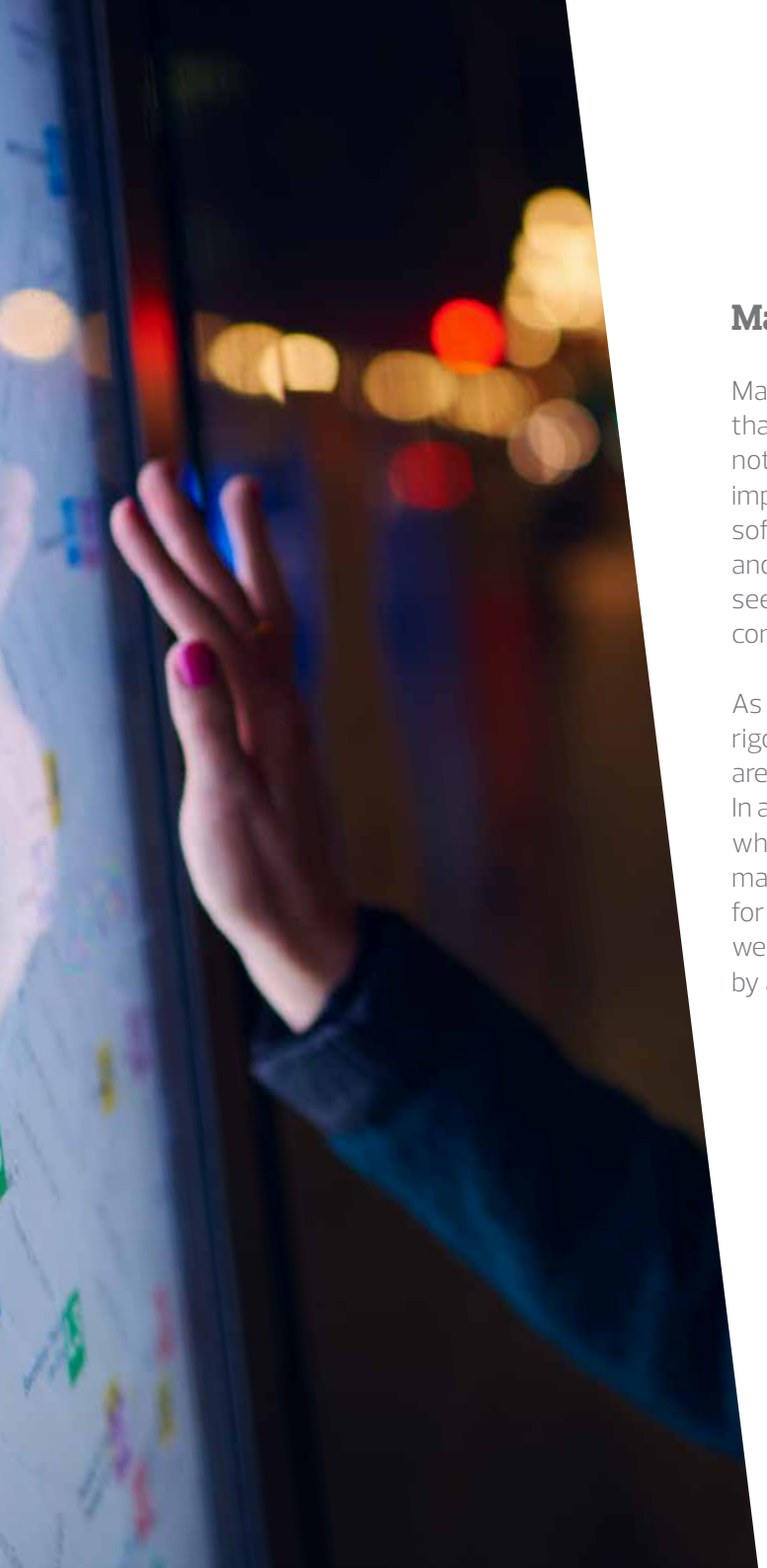


Case study 2

An application of machine learning for road condition monitoring

TRL has recently developed image processing algorithms to detect cracks and other features in road pavements. A neural network model was trained on images collected by downward and forward facing images of the road surface, applied to a separate set of images on which the model had not been developed, and compared to the results from manual processing.

The results showed that this algorithm has the potential to greatly speed up the current road condition monitoring programme, which relies on extensive manual surveys of the road network.



Machine learning: the upsides

Machine learning algorithms allow us to look into data that we don't have a good understanding of, and that do not fit into a traditional statistical framework. One area of important growth is in image processing. Image recognition software has improved markedly over the past few years and is being utilised in many areas of transport research – see Case Study 2 for an example of its application in road condition monitoring[6].

As machine learning algorithms do not require the rigorous assumptions required for statistical models, they are less restricted in terms of their range of applications. In addition to image processing, there are other problems which are suitable for these algorithms. Some processes may not fit any of the statistical distributions available; for example, any data driven by social influences such as website hits and words used on Twitter will be generated by a complex mix of factors.

There is no obvious physical model on which to base these data and so a machine learning algorithm may be the most appropriate place to start. Machine learning has enabled investigations of this type of data (sometimes called unstructured data) via Natural Language Processing and Deep Learning. There may be other complex processes where the general shape of the relationship is not known at the outset, and so it is less easy to fit a known statistical distribution or test.

Another consequence of requiring a less rigorous approach is that less input/expertise is required from the modeller. Strong programming skills are essential, but formal training or numerous years of experience in statistical modelling is not an essential requirement in machine learning. In addition, even when modelling assumptions are known to have been violated, certain algorithms have been shown to perform well in terms of predictive power.

Due to the focus of their approach, machine learning algorithms can have greater predictive power than statistical models. This is demonstrated in the results of the Kaggle [7] data science competitions, where machine learning algorithms often outperform statistical models.

6| Nemas K, Khatry R, Smirnov A, Peeling D, Mistry S, Crabtree M and Reeves S (2018). Applications of machine learning in transport. PPR863. TRL.

7| https://www.kaggle.com/competitions

Machine learning: the downsides

Some outputs from machine learning algorithms are more transparent than others, but many combine multiple models or use techniques such as neural networks which are difficult to interpret. This can lead machine learning models to be seen as a 'black box', delivering little insight into the process driving the outcomes. This may be troubling if the algorithm is being used to make complex and important decisions such as whether an autonomous vehicle should brake or not. In addition, if results do not make intuitive sense in terms of the real world setting, this can be difficult to justify and explain to stakeholders and clients.

Due to their 'black box' nature, machine learning algorithms have also been accused of being a significant contributor to the current reproducibility crisis in science[8]. This is the frequent number of occasions where research groups repeat the same experiment as previous groups, but obtain different results.

When the potential impacts of mistakes are very high, it would be an advantage to understand how decisions are arrived at, not least so that any unusual decisions can be checked and fixed. It is possible that suboptimal models are generated, where a more appropriate one could have been found with a more rigorous examination of the data. In addition, the relationships between variables is not explicitly checked during the process of training a machine learning algorithm, so less insight is provided into these.

Even if a suitable model is created for the first set of data, the nature of the data could still change over time. Any automatic update process could be vulnerable to accumulating errors over time. There should be strict checks and balances on the data and alerts to human researchers if extra checking is required.

In addition, the downside of the algorithms requiring fewer assumptions means that in general, large amounts of data are required to ''train'' them. For some techniques, vast amounts of data, consisting of millions of observations are required.

8| Dr Genevera Allen's comments on the current reproducibility crisis in science, reported by the BBC: https://www.bbc.co.uk/news/science-environment-47267081

How to select the most appropriate approach?

With so many new datasets becoming available and no shortage of problems to solve, all potential modelling solutions should be considered. Often this process will require the expert help of a data scientist or statistician. It is helpful to start by asking two important questions. The answers to these will inform the modeller on which approach to use.

Key questions to ask

1. Which type of data is available?

- Is the dataset small or large?
- Is the data "structured" (e.g. time series) or "unstructured" (e.g. images)

2. What insights are we attempting to extract from the data?

- Is our interest purely in the prediction power of the model?
- Or are we interested in deeper insights – such as gaining an understanding of the relationships involved and of the process that generated the data?

Small or experimental datasets

If the dataset is small or generated as part of an experiment, then statistical modelling will likely be the best way forward for both understanding and evaluating the data. Statistical modelling is particularly suited to the evaluation of trial or experimental data. It is often beneficial to get advice from a statistician about suitable analysis techniques at the beginning of a project, rather than waiting until the end.

Statistical testing can also be used to test different hypothesis; for example, data collected via in-vehicle sensors under different conditions can be used to test hypotheses around how these conditions affect driver behaviour.

Large datasets

If the dataset is large, but we hope to gain some understanding of the process being modelled, statistical modelling may still be the best route forward.

However, the emphasis changes if the data are unstructured, driven by unknown processes, or includes image data. In these cases a machine learning approach may be best.

Insights

If the primary interest is in the predictive power of the model, then depending on the data that we have, machine learning algorithms may be the most appropriate choice. However, if we are interested in a wider range of insights then statistical modelling can provide these.

Working hand in hand or the best approach

These guidelines are not hard and fast rules, and both statistical modelling and machine learning techniques are in constant development. The choice may not be clear-cut and in some cases multiple approaches will be suitable. A full discussion with a data expert will help identify the most suitable approach.

Any approach will be strongest if it takes the best features of both statistical modelling and machine learning. From statistical modelling this means cleaning, visualising and checking of the data before any modelling begins and the rigorous checking of any model assumptions. From machine learning the focus on prediction power and incorporation of cross-validation will benefit both machine learning and statistical models. Any model can be regularly updated and coded to do so in a straightforward fashion by using a coding language such as R or Python.

Whether the analysis technique chosen is statistical modelling or machine learning, the value of understanding a dataset prior to carrying out any analysis should not be overlooked. This often means going back as far as the data collection process to understand whether missing data, biases or measurement errors will influence the interpretation. It is critical to understand the limitations of the data in order to ensure that the resulting conclusions are robust and caveated appropriately.

TRL's capabilities for predictive modelling in transport

Whichever analysis technique is chosen, it is important to ensure that the results of the analysis are put into context. The term 'Data Science' has really taken off since 2014[9], but although the availability of skills in this area is increasing, real value will only be derived through the application of specialist transport knowledge to the interpretation of the results.

TRL's vision in this context is to be a Centre of Excellence for Data Analytics in Transport. Our experts have a vast array of knowledge across a range of core areas including infrastructure asset management and asset technologies, intelligent transport systems and traffic operations, sustainability and healthy mobility, vehicle safety engineering and technology research, major incident investigations, human factors, safety, and behavioural science. But we also recognise the importance of data science in these areas, and the need for continued skill development and learning. We are investing heavily in this area and are engaged in a variety of projects looking at the applications of machine learning, including funding for a University of Warwick PhD student to investigate efficient ways of linking large spatial–temporal datasets.

As a result, we can apply our expertise to answer a wide range of questions such as:

- What insight can be gained from data which have already been collected?
- How might these data be used as an evidence base to support strategic decisions, target setting or to answer specific research questions?
- What is the best method of processing, analysing and visualising data to provide these insights?
- How can the most important questions facing the transport industry help to shape data collection?
- How should data be collected and analysed to answer the salient questions?

9| A comparison of the use of this term on Google trends (https://trends.google.com/trends/?geo=US) shows that worldwide use of this term has increased five-fold between 2014 and 2018.

An example of a transport application which could benefit from applying these skills is the classification of automated vehicle behaviour, in order to assess vehicle performance. TRL is currently using statistical modelling as part of the MOVE_UK project to validate the performance of in-vehicle systems, such as traffic sign recognition, in the real-world. This application could easily be extended to other vehicle systems. This would help to identify the future challenges for those who manage the road network, in order to ensure that automated vehicles can safely operate in a live road environment.

A potential application of machine learning would be the development of improved algorithms for the setting of signals on Smart Motorways. Currently these are based on a threshold flow level for each site, but machine learning could be used to more accurately predict the onset of congestion from surrounding traffic data. This could improve road user perception of the signals thus improving customer satisfaction, a key priority for Highways England.

If you are interested in exploring opportunities to work with us to obtain more value from your datasets then please get in touch.

THE FUTURE OF TRANSPORT

There is a huge volume of transport datasets in existence. Data are an increasingly valuable asset, but this value is lost without sound analysis and interpretation, creating a growing demand for data science capabilities. Predictive models attempt to explain how historical data can be used to predict the future; two of the main approaches are statistical modelling and machine learning. This white paper compares both approaches, and for each presents a case study of its high-impact use at TRL, before providing guidance on when to select each approach.

Statistical modelling offers a rigorous approach and can provide a rich range of insights. This is most suitable for small and experimental datasets, but potentially also larger datasets – providing that these data have a known structure. However statistical modelling can fail to deliver if a set of assumptions cannot be met. Machine learning algorithms are less restricted in terms of applications and can handle unstructured data such as images. This approach focusses on the predictive power of the model and generally requires larger datasets. However the 'black box' nature of many techniques provides few insights. Any method will be strongest if it takes the best features from both approaches.

enquiries@trl.co.uk
www.trl.co.uk

TRL Crowthorne House, Nine Mile Ride,
Wokingham, Berks, UK, RG40 3GA

ACA011
ISSN: 2514-9695
ISBN: 978-1-912433-86-5