Team 17 DVAPAC: Queensland Rail Maintenance and Track Visualisation Study

Marcus Salouk, Melvin Yin, Rahul Chowdhury & Jiajun Liu

Summary

Track Recording Car (TRC) data collected quarterly over 1m intervals by Queensland Rail (QR). This data details **track geometry** and is used to inform **ballast maintenance decisions.** In combination with Ground Penetrating Radar (GPR) and culvert location data, the team used **machine learning** algorithms to predict:

a) **Track geometry in the next quarter**

b) **Track maintenance needs in the upcoming quarter(s)**

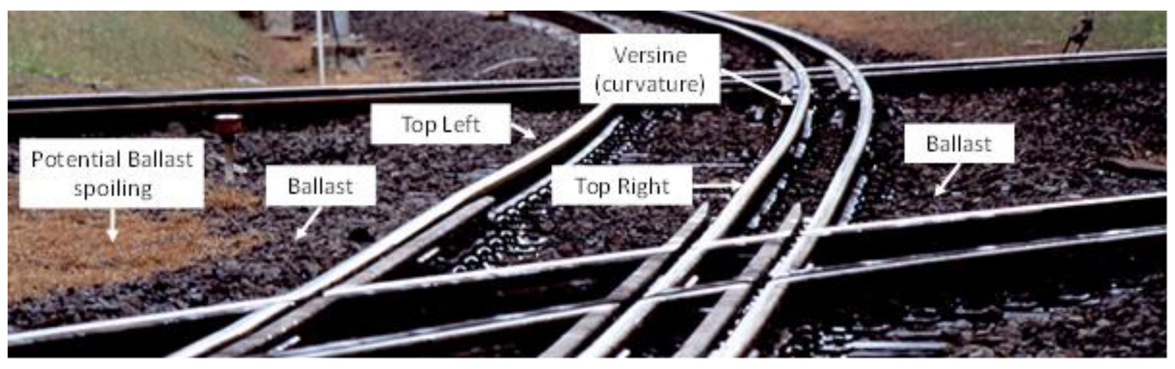


Figure 1. Illustration of Key terms

Data Exploration

- **TRC data:** [1m precision]
  - 3 track lines' worth of data over 22 quarters
  - 16 measures of track position and geometry
  - *Combined* measure follows Exponentiated Weibull distribution (fig 2.)
- **GPR data:** [5m precision]
  - Data for all lines available from 2015 run
  - 32 measures of ballast and track bed properties
- **Culvert data:** Geospatial drainage information
- **Work order data:** Time and type of maintenance work for 2 track lines. Certain areas are prone to higher work orders (fig 3.)

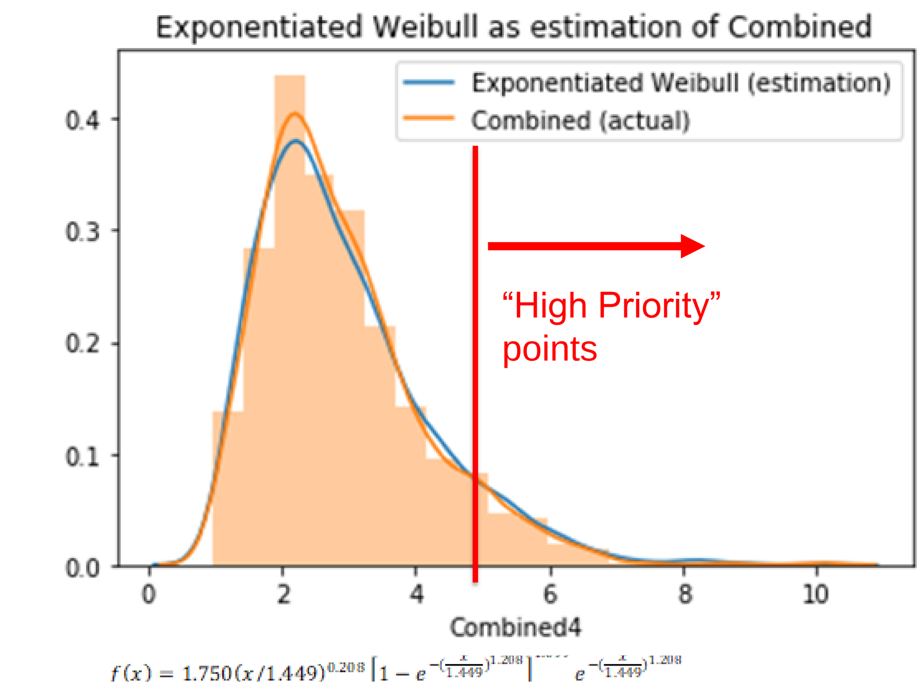


Figure 2. TRC combined measures follow a Weibull distribution. "High Priority" points above QR-set threshold of 4.9

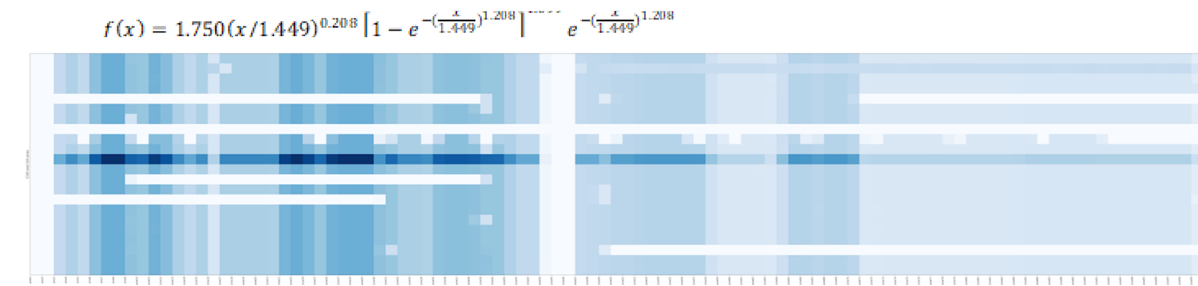
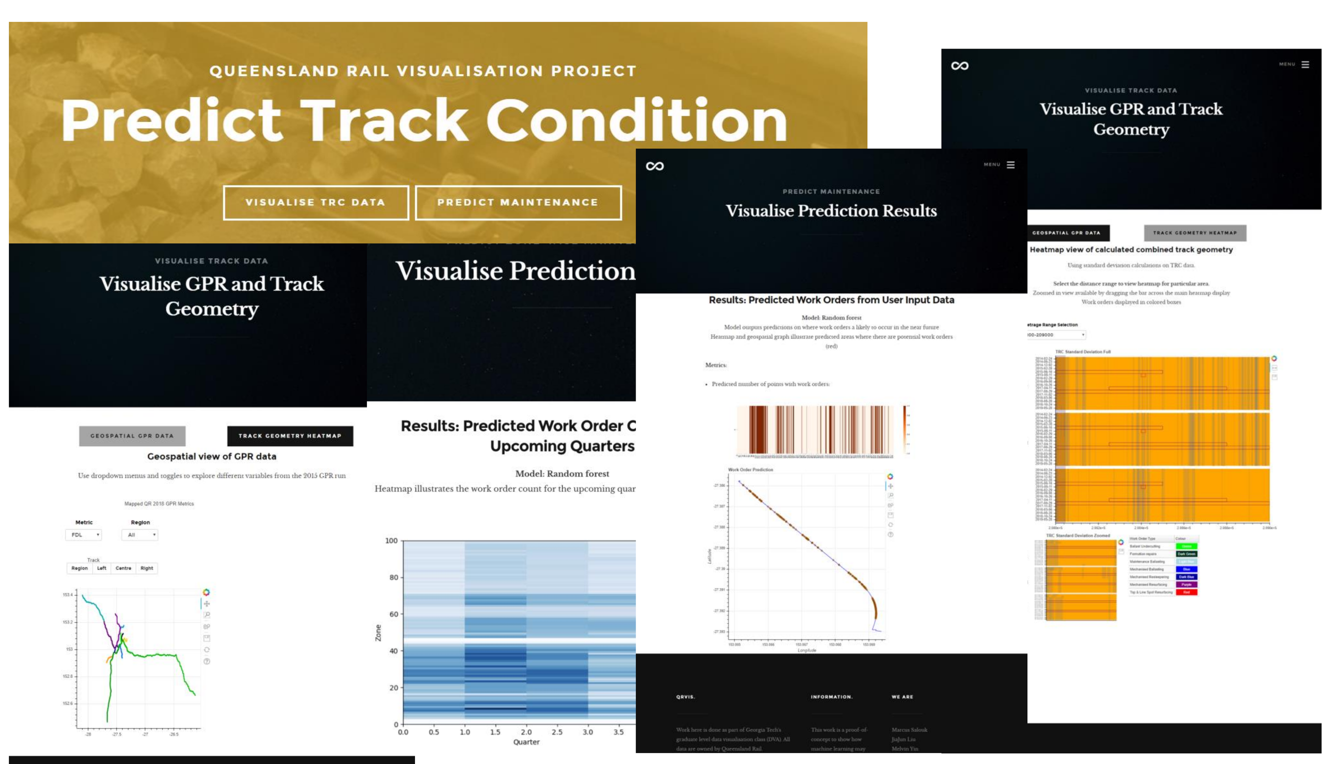


Figure 3. Uneven distribution of work orders over track

Web App with Interactive Visualisations

- Visualising GPR and TRC data – GPR data in **geospatial heatmap**, TRC calculated combined metric **scrolling heatmap** (fig 4)
- ML model predictions - **user uploaded TRC data** outputs model predictions on upcoming track maintenance, represented in heatmap and geospatial form



[Go To Repo] [Go To Site]

Figure 4. Screenshots of Django web application with interactive visualisations

Pre-processing data

- **TRC alignment tool:** Fix mis-alignment across time.
  - Comparing SUPER and GAUGE across time and minimising deviation for each TRC run
- **Joining datasets for analysis:**
  - Data joins performed using pandas library
  - Track codes, distance measures used for joining across datasets
- **Missing values Imputation:**
  - Missing values were imputed with either '0' or with surrounding values
- **Binning data into zones and quarters**: High speed data aggregation using parallel processing to handle huge quantity of data

a) Track Geometry Predictions

- Track geometry changes inferred via *combined metric* (fig 4) – indicative of **ballast degradation**.
- ML models trained on TRC + GPR + culvert data to predict future track geometry (table 1):
  - RF, SVR, KNN and ANN performed similarly in predicting 1 quarter ahead (table 1)
  - **RF outperformed base-model** using preceding Combined value as prediction by ~3% / ~12% accuracy in all / "high priority" points (fig 5)
  - **RF superior to base-model over 2 quarters prediction**: ~22% / ~42% accuracy in all / "high priority" points (fig 6)

$$Combined \equiv \frac{Std.DEV(TopLeft) + Std.Dev(TopRight)}{2} + Std.DEV(Twist3)$$

Figure 4. Equation of a combined measure representing changes in track geometry, most attributed to ballast condition.

Model	Best Test Accuracy	Gradient of best-fit	Best test Accuracy "high priority"	Best Correlation "high priority"
Baseline	80.92%	0.77	73.69%	0.06
Random Forest (9 features)	84.35%	0.99	86.2%	0.48
SVR (20 features, scaled)	85.14%	0.9	85.25%	0.2
KNN (9 features, scaled)	82.21%	1.02	81.64%	0.23
ANN (20 features, scaled)	85.88%	0.94	85.54%	0.34

Table1. Summary of ML results for predicting future track geometry 1 quarter ahead

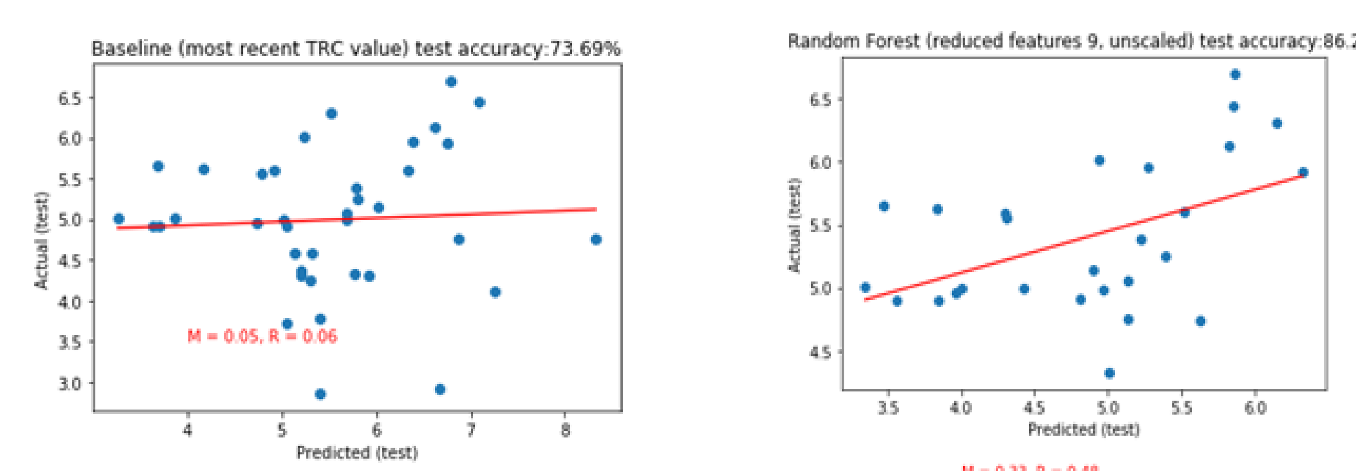


Figure 5. Predicting "high priority" points track geometry 1-quarter ahead. Random Forest (86.2% accuracy) outperformed baseline model which projected the preceding quarter response variable (73.69%).

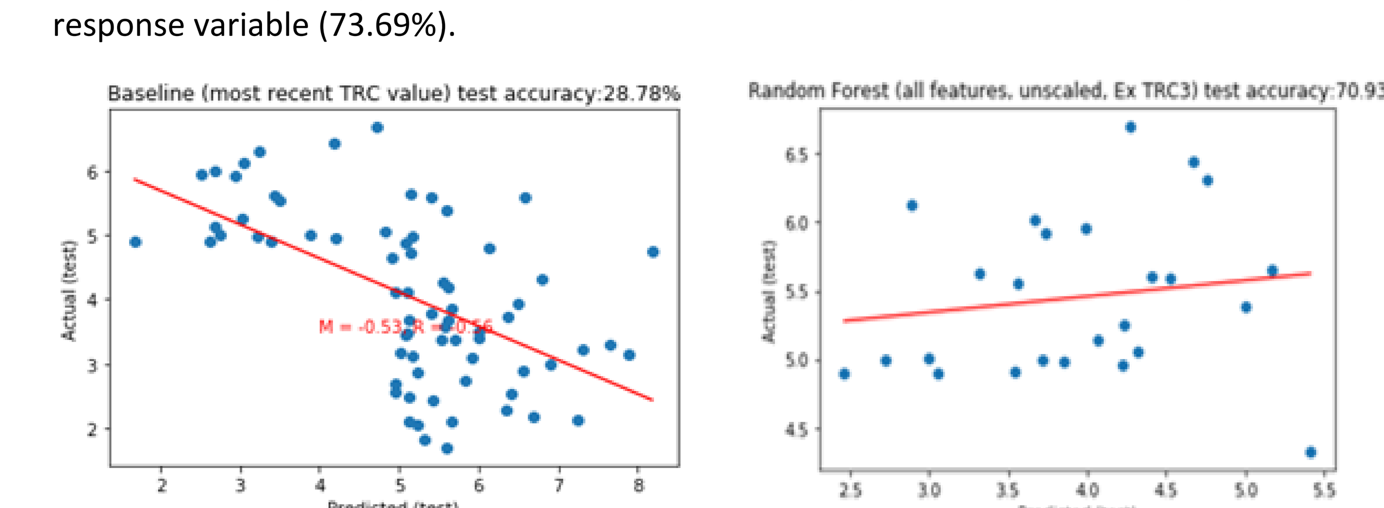


Figure 6. Predicting "high priority" track geometry points 2-quarters ahead. Baseline model unable to predict 2 quarters ahead (28.78% accuracy) vs. Random Forest (70.93% accuracy)

b) Track Maintenance Predictions

- **Zone-wise aggregation** of **quarterly** work order count gave rise to patterns with high correlation to track geometry (fig 8).
  - Track divided into 100 zones for analysis (fig 7)
  - Possible to identify troublesome zones, and predict for upcoming quarters.

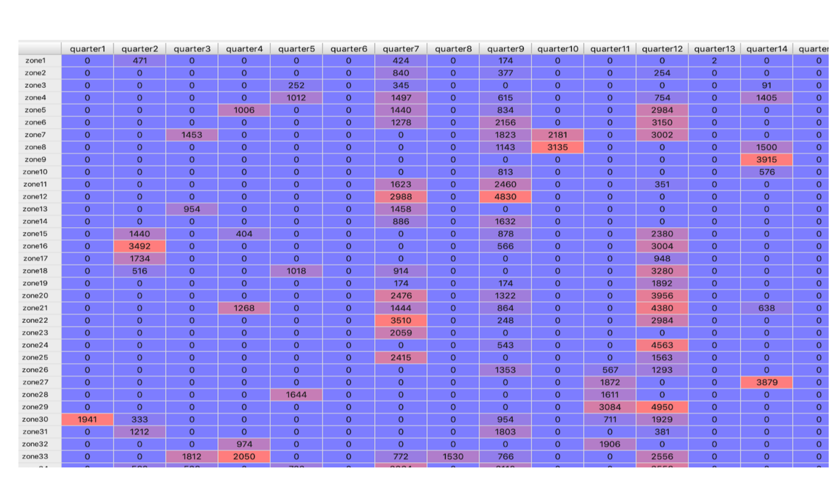


Figure 7. Maintenance work orders are binned into 100 zones and visualised as a heatmap. Some zones and quarters are more heavily maintained than others.

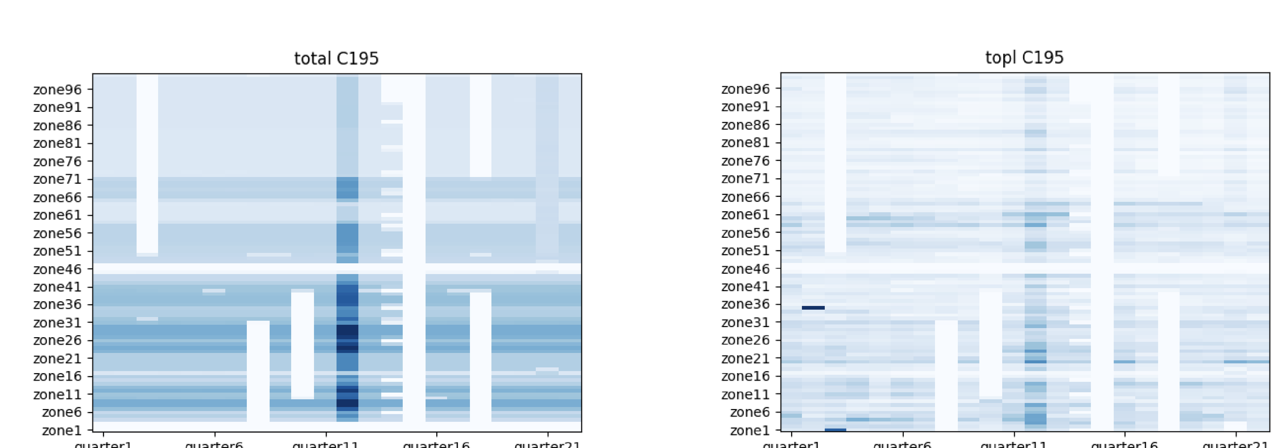


Figure 8. High correlation between work orders and track geometry indicates that maintenance carried out is effective and has immediate impact.

- Classify and predict presence of track maintenance each quarter using **GPR + TRC data** (fig 9,10).
  - Data was downsampled to balance classes
  - **Random Forest** displayed best test performance (table 2)
  - Accuracy reduction when using non-balanced data (whole TRC runs) indicating high variation within negative class not captured by model
  - Results visualised as an interactive geospatial plot (fig 11)

Model	Accuracy (bal.)	F1 Score (bal.)	Accuracy (unbal.)	F1 Score (unbal.)
Logistic Regression	57.4%	0.61	NA	NA
SVM	80.6%	0.38	NA	NA
Random Forest (base)	87.6%	0.44	65.3%	0.37
Random Forest + CV	89.5%	0.40	71.87%	0.80

Table2. Summary of ML result for predicting upcoming maintenance work. Random Forest performed the best in test accuracy with balanced class data, but suffered decrease in accuracy with unbalanced (raw) TRC data

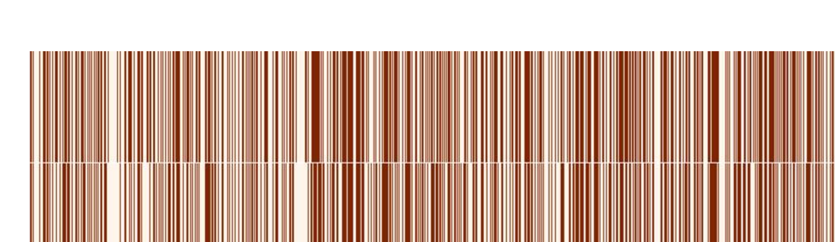


Figure 9. Heatmap of actual (top) vs predicted (bottom) maintenance work (balanced data; acc 87.6%): RF + CV

Figure 10. Heatmap of actual (top) vs predicted (bottom) maintenance work (unbalanced data; acc 72.6%) from C138 run in 2015-05-15: RF + CV model

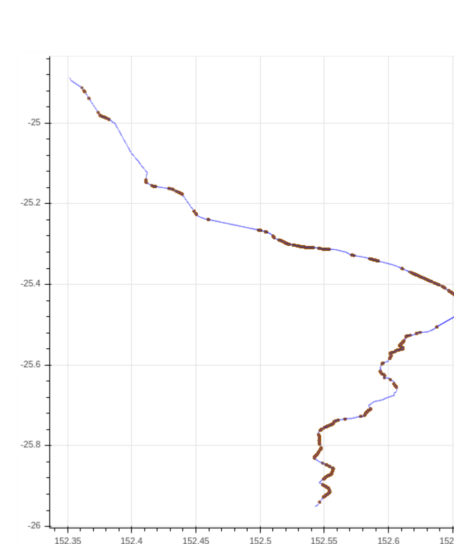


Figure 11. Geospatial visualisation of prediction output used in web app, generated in bokeh.