**Trenord Diagnostic Data Processing**

We have .csv files containing diagnostic event information of rolling stocks. Each .csv file consists of a set of diagnostic events recorded for a duration of around 2.5 years for a train.

Python is used for this purpose. For cleaning and processing of diagnostic data from Trenord, we have focused on the following aspects:

1. **Mixing up of diagnostic data from different trains :** In the diagnostic events file, there is the possibility of having diagnostic events recorded from other train.
   1. **Objective:** Need to check whether all diagnostic events recorded in a file are from the same train.
   2. **Action taken** - check for the data source which represents the train number (e.g., field ‘Sorgente\_Dati’ in the .csv file).
2. **Time series information** –
   1. .csv files contain time information in the Unix timestamp format expressed in milliseconds. More information on the format is available at <https://www.unixtimestamp.com/>. We converted this number into the usual DD/MM/YYYY HH:MM:SS format with the formula: =A1/86400000+DATE(1970;1;1)
   2. Data is ensured to be ordered chronologically
3. **Alert type verification**: There is misalignment between the ‘Type’ of alert mentioned and the actual alert (i.e., an event is tagged as PDM while it is actually a PDO) for some diagnostic events.
   1. Action taken - We have verified the actual alert type by inspecting the alert code (e.g., ‘Gruppo’, ‘PdM’, ‘PdO’ field)
4. It is necessary to eliminate the diagnostic data generated within the maintenance systems. These are mostly associated with operations of the technicians carrying out the maintenance and not to actual faults that appear at that moment. These data are therefore not relevant for analysis purposes and can be easily identified as they show "1" in the "Deposit" column.
5. Consider only the data that are event = ‘ON’
6. **Consider only PDO for analysis :** In the file, each diagnostic event is associated with a PDM (engine personnel), i.e. an alarm visible to the driver. Each PDM is associated with K PDOs (workshop personnel). The PDOs specify more precisely the causes that determined the PDM. So for the next step of the analysis we only consider PDO. Here, we need to consider an issue where a PDM is not associated with any PDO. We need to take care of such cases.
   1. Action taken –
      1. From the alert code file, identify the PDMs that are not associated with any PDO (i.e., we have found 5 PDM alert for TCU which are not associated with any PDO)
      2. While considering the diagnostic events of a train we usually consider the PDO except we also consider PDMs which belongs to the 5 isolated identified PDMs (i.e., which are not associated with any PDO)
7. **Discard redundant events:** There are redundant events (PDMs and PDOs) in the diagnostic files.
   1. As regards the identification of diagnostics repeated following Trainset events, we have verified that these do not appear at exactly the same instant as the Trainset but in the moments immediately following: on some sample cases that we have analysed, the repetition occurs within 1-2 seconds from the “Trainset” event. There is only one diagnostic case related to the TCU, however it is part of another implant.
   2. Due to the IDU system structure :
      1. From a theoretical point of view :
         1. Data download from MD and MS of the 711-xxx are the same. These downloads both contain the PDOs generated by the control units forming part of the Train Set (1) of which the 711-xxx is part.
         2. Data download from MD and MS of the 711-yyy are the same. These downloads both contain the PDOs generated by the control units forming part of the Train Set (2) of which the 711-yyy is part.
         3. The PDMs, regardless of whether they are generated by the Train Set (1) or Train Set (2) control units, are memorized by all four monitors present on the train.
         4. Thus there may be two records of a PDO and four records of a PDM
      2. From a practical point of view :
         1. The download from MD and MS is not the same: the events are not temporally aligned (even for just fractions of a second) and some alarms, apparently randomly, are present only in one of the two monitors.
         2. Not all PdMs are stored on all four monitors and, for those stored, there is no exact time alignment. Some events are present only in the downloads made by the MCH in which the bank key is inserted.

Therefore, in order to have as complete a view as possible on the diagnostic status of the convoy, it will be necessary to merge the data downloads from all four monitors. This merge must necessarily be followed by an appropriate elimination of duplications. We have followed a simple heuristic to discard the duplicated data as it is extremely difficult to define rules to properly cleaning the data considering the above situation.

**Action taken:** As we are only considering the PDOs for the analysis, we consider the half of the number of alerts/critical alerts observed for a specific time-stamp.