

|  |
| --- |
| TRAFFIC ACCIDENT SEVERITY PREDICTION –REPORT |
|  |
| November 15  Authored by: Andy Sin |

# TRAFFIC ACCIDENT SEVERITY PREDICTION

|  |
| --- |
| Introduction Drivers and insurers alike are keen to understand what is the probability and potential severity of accidents occurring on a given journey and under certain observable conditions. This model will use factors such as weather conditions, time of day and area to help derive predictions.  The output of this model will help users in determining the appetite for undertaking particular journeys or perhaps assist insurers in determining the premium level which should be applied to insurance. |

|  |
| --- |
| The DataThe dataset is based on publicly available collision data for the Seattle and as captured by Seattle Police Department. The data which is made available specifically for this study has been pre-filtered to only contain Severity 1 & 2 data, representing collisions with *property damage* or *injury respectively*. Collisions with Severity 2b and 3 data which denotes collision with *Serious Injury* or *Fatality* have been excluded.In terms of timeline, the dataset contains collision data from 2004 to present.The data is presented in a CSV format and originally contains 31 different attribute of data.**Data Cleansing**For the purpose of this evaluation, the dataset was consolidated down to focus on the following attributes which forms the main element of our modelling:Severity DescriptionsIncident DateIncident TimeWeather Conditions  * Light Conditions * Road Conditions  These attributes were selected as this analysis will focus on modelling severity based on the macro level conditions surrounding the journeys. Factors which are focused on the specific circumstances of the collision (such as location, drug influences, number of cars involved etc) are outside of the scope of this analysis.To further facilitate the analysis, the data has been further recategorized as follow:Conversion of Incident Date and Incident Time from Object Type to DateTime TypeRecategorisation of Incident Date into Calendar QuartersRecategorisation of Incident Time as follow:00:00 to 05:59 🡪 Early Morning06:00 to 11:59 🡪 Morning12:00 to 17:59 🡪 Afternoon18:00 to 23:59 🡪 Evening |

## Methodology

## The overall exercise was broken down into two distinct parts. The first part included an analysis focusing on getting an initial understanding of when collision events happen. Specifically, this involved looking at the distribution of collision across the selected key attributes which have been selected. The intent of this was to help in diagnosing the broad trends in the data, if any.

## The second part turned the attention to providing the modelling capability based on the attributes selected and using that as future predictor in determining if a collision were to occur, whether it will be a Severity 1 or 2. As we do not have associated data on the total journeys which were made, the model will not be able to predict the probability of a collision happening in the first instance.

## For the modelling, the “Decision Tree” technique was selected as this is deemed to be the best technique in dealing with the multiple discrete and independent nature of each individual attribute. Through building out this model, we will be able to model given a particular set of attributes and a collision happening, whether a Severity 1 or 2 collision is likely to occur.

## Results

## **Part 1 – Understanding the data**

Illustration 1 : Distribution of Collision by Quarter

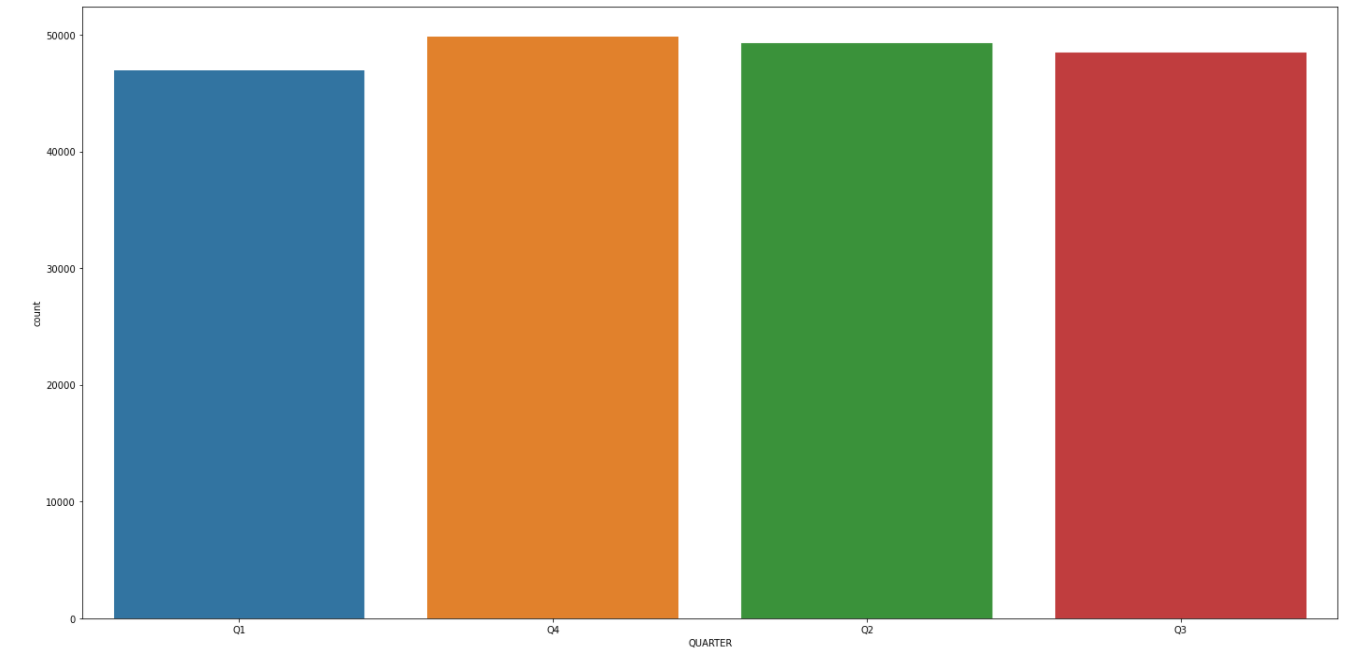


Illustration 1 shows the distribution of collisions by quarter. It can be observed that the collision is spread relatively evenly across all quarters although Quarter 1 does show a comparatively lower volume of Severity 1 and 2 collisions than the other quarters. Overall though, assuming that the number of journeys being taken is fairly constant throughout the year, it can be deduced that there are no major seasonal variations in the likelihood of a collision taking place across the year.

Illustration 2 : Distribution of Collision by Time of Day

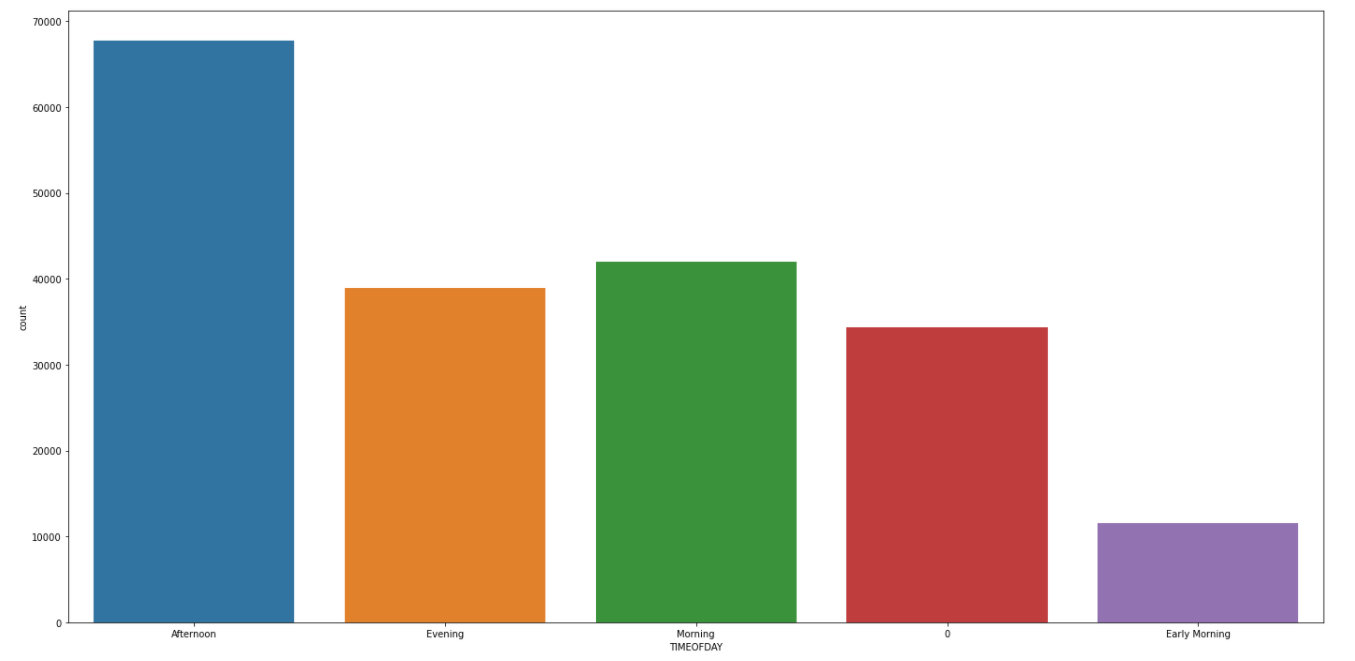
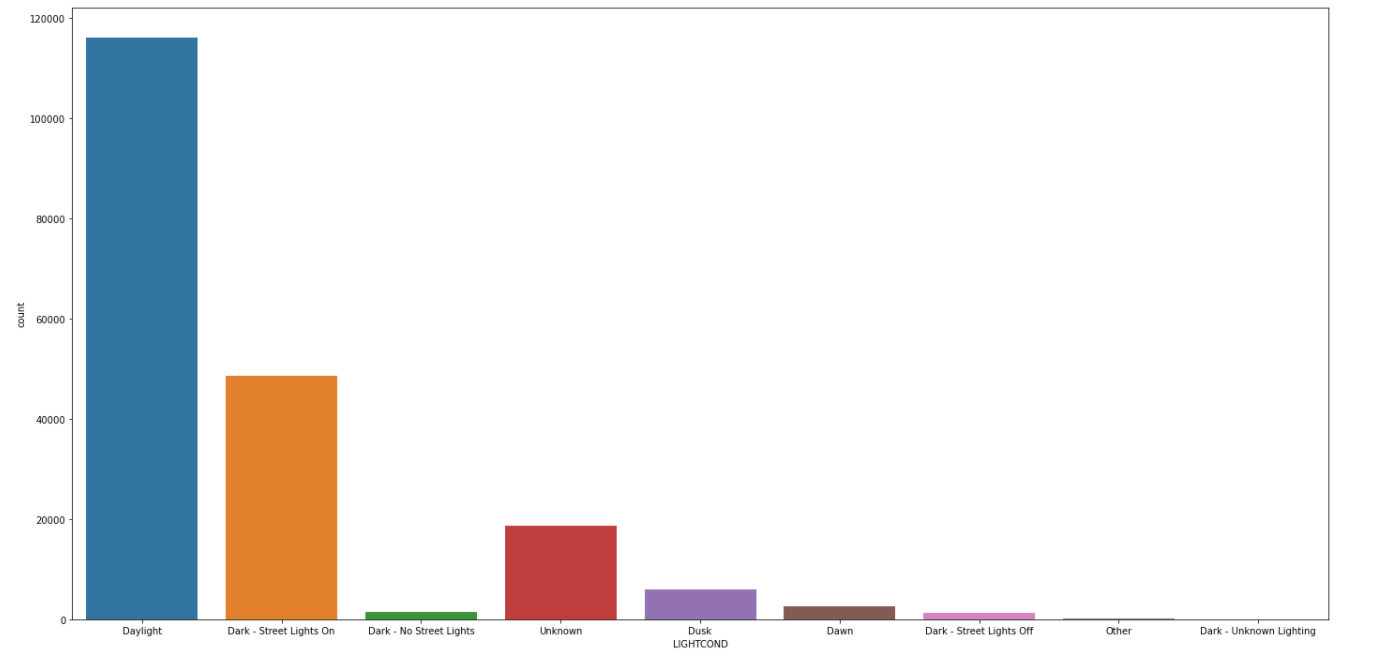


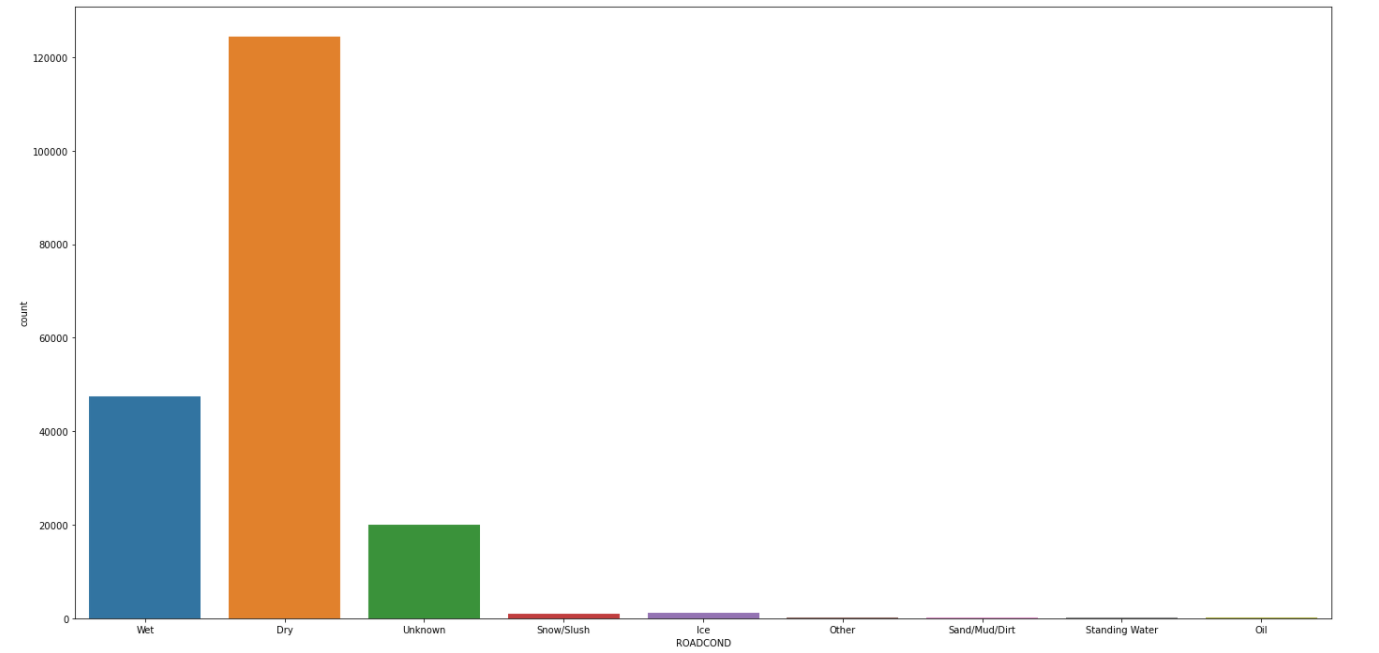
Illustration 2 shows an interesting insight in that there are materially more accidents taking place in the Afternoon than at any other time during the day. It is likely that this is driven by the rush hour period with increase traffic on the road. However, this compares unfavourably compared to the Morning period which also encapsulates the usual rush hour travels. A possible reason for this could be due to higher level of driver fatigue in the afternoon rush hour compared to the morning rush hour. Furthermore, by late evening, which is when majority of rush hour traffic occurs, light conditions are usually diminished. We can explore this correlation more in the next part of the analysis.

Illustration 3. Distribution of Collision by Light Condition



It can be observed from Illustration 3 that the large majority of the accidents happen during daylight hours. This is not surprisingly as it is likely the case that the large majority of journeys undertaken take place during the day. We do not have clear definition as to exactly the daylight conditions are defined but will assume that this includes twilight periods during dawn and dusk and that Daylight categorization is prevalent during the Morning and Afternoon periods. On the basis of these assumptions, this would correlate to the higher volume of accidents observed in the Morning and Afternoon periods.

Illustration 4 : Distribution of Collision by Road Conditions



Independently of the analysis above, this illustration shows that the large proportion of collisions do occur in dry conditions. This could lead to a simple conclusion that drivers are more likely to be involved in a collision in dry conditions. However, given consideration that the analysis is based on collision data from Seattle and that it is one of the five rainiest major U.S. cities as measured by the number of days with precipitation, that conclusion may not be accurate. Further analysis will likely be required to truly understand the likelihood of collisions happening in wet vs. dry conditions.

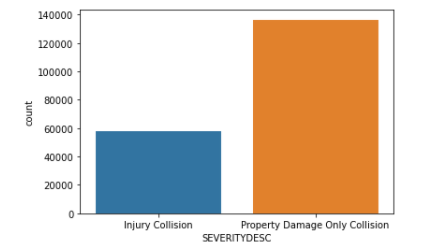
## **Part 2 : Predicting whether Severity 1 or 2 collision will occur**

As set out in the methodology, Decision Tree model was selected as a means to predict the likely Severity outcome of a collision given the following variables:

* Time of the year when it happened (by Quarters);
* Time of Day;
* Weather Condition; and
* Light Condition.

It is worth noting that based on the dataset, it can be observed that there a greater proportion of collision which are categorised Severity 1 – Property Damage Only Collision. The expectation is therefore that for the large proportion of overall conditions, it is more likely that a collision will be Severity 1 Collision.

Illustration 5 – Breakdown of Severity



However, the model is seeking to determine if there are certain scenarios will mean a Severity 2 - Injury collision is more prevalent than that of a Severity 1 – Property Damage Only Collision. The logic being that in more treacherous conditions, such as poor light or wet conditions, the nature of the collisions are likely to be more severe and thus potential for Severity 2 collision increases.

Illustration 6 – Full Decision Tree Diagram

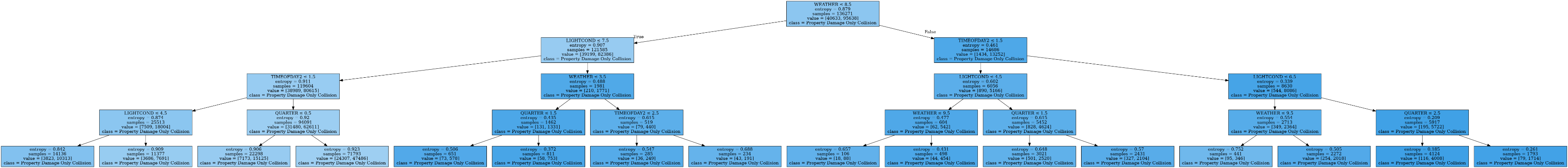


Illustration 6a – Scaled up Decision Tree (Left)

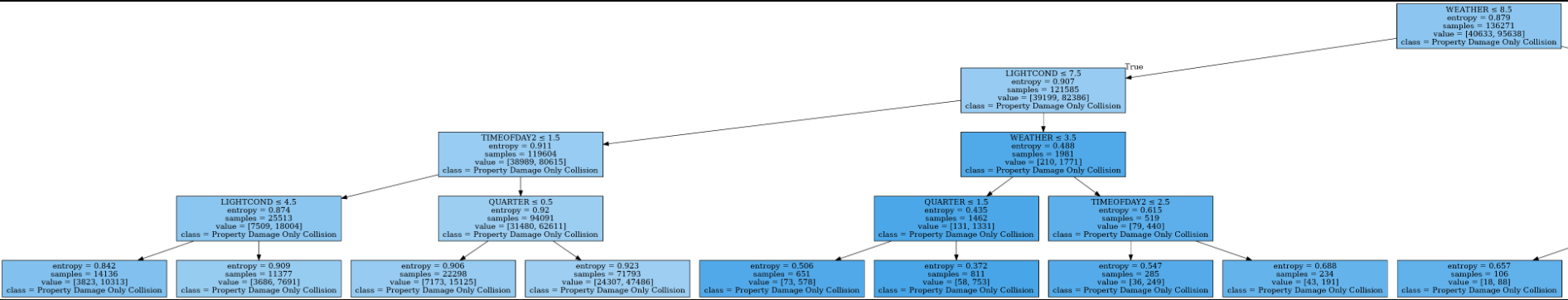
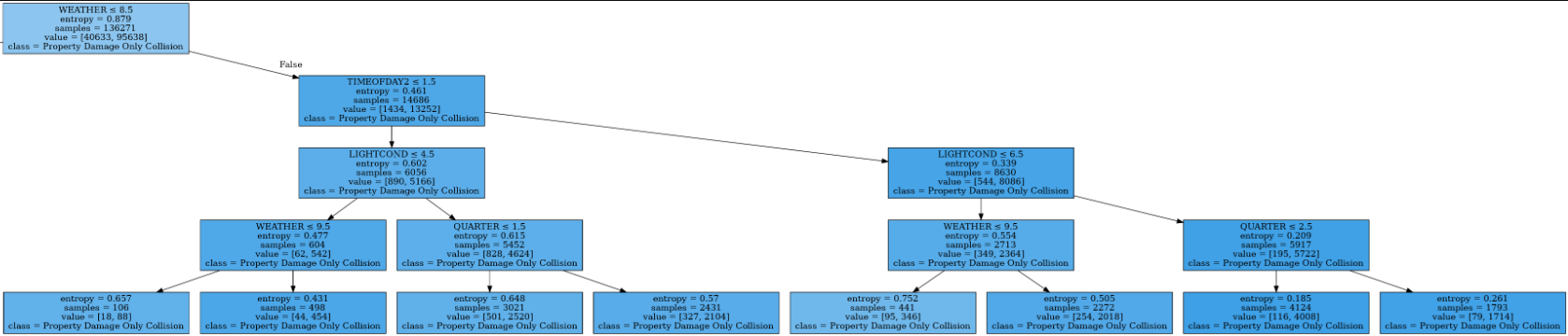


Illustration 6b – Scaled up Decision Tree (Right)



As can be seen from the Decision Tree Model above, based on the modelling technique used, it is not possible to attribute a particular set of conditions which will lead to Severity 2 – Injury Collision to be more likely than Severity 1 – Property Damage Only collision.

## Conclusion

The analysis has given us a good level of initial insight on the distribution of the collisions which occur during certain conditions. We can clearly observe that although there are no large differences between collisions happening during each calendar quarter. It has been clearly shown that collisions are not as evenly distributed across the other variables and it can be deduced that more collisions happen under certain conditions.

Nevertheless, in the absence of broader underlying journey dataset and road conditions across all journeys, it isn’t currently possible to definitively confirm if certain conditions are more likely to cause collisions when looking at all journeys despite a potential early indication that this might be the case.

In part 2 of the analysis, based on the initial Decision Tree modelling, it can be observed that there are no specific conditions which will lead to a Severity 2 collision being more likely to happen than a Severity 1 collision. Further analysis may yield different results and would merit further investigations.