SAPIENZA
UNIVERSITÀ DI ROMA

A Data-Analysis Approach for Driver Behavior Characterization

Ingegneria dell'informazione, informatica e statistica

Corso di Laurea Magistrale in Ingegneria Informatica- Engineering in Computer Science

Candidate

Falesiedi Flavio Massimo

ID number 1595746

Thesis Advisor

Prof. Aristidis Anagnostopoulos

Co-Advisor

Prof. Ioannis Chatzigiannakis

Academic Year 2019/2020

Thesis not yet defended

A Data-Analysis Approach for Driver Behavior Characterization
Master's thesis. Sapienza – University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Version: March 16, 2020

Author's email: Flaviomax94.fmf@gmail.com

Contents

Abstract

Despite the continuous evolution of vehicles' fine-tuning developed by the various car companies in terms of on-board instrumentation and safety systems, the world of daily personal transport still presents some problems that do not guarantee perfect safety on the road. The question, then, arises spontaneously; are these new mechanisms sufficient for granting safety to those who are driving and for those around them? The answer is: generally yes. These systems are well made and fulfill their work. The real problem is the human factor. In this study our goal is to catalog the drivers' behaviors through the set of data collected from different trips along with the cities, to outline a profile, monitoring attitudes in the road and improve driving and safety. However, those things are based on the use of the innovations that are present in the vehicles or, in parallel, through the use of information from the car's control unit and from data obtainable from smartphones. Employing such information, we have outlined an algorithm that, based on what are the main and secondary factors influencing a bad driving style, provides a group of thresholds for these values which, near some scores, contribute to classify the driver. In support of this identification, there is then a cross-check based on the analysis of the maximum peaks of the main factors, which is used to confirm the driver profiling, or eventually dispel the erroneous classifications characterized by altered data. This system, therefore, can be safely used as a driving tutor who comes to scold us in case of dangerous behavior on the road.

Chapter 1

Introduction

This work aims to enhance the theme of road safety, to help motorists to have a higher understanding of their driving approach, indicating a bad use of the vehicle and trying to understand when, dictated by altered or agitated moods, the driver is engaging in dangerous behavior towards himself and others. We know in fact how the human component, despite the continuous evolution of vehicle instrumentation, is the most determining factor as regards road risks, since failure to comply with road rules cannot be identified solely with those mechanisms inside the vehicle. This is why the help of the analysis of aggressive or unsuitable driving attitudes takes over.

In this study, these behaviors are not only related to speed limits, but concern numerous factors such as the level of fuel consumption and therefore the level of pollution, the impetuosity of pedal pressures and steering, and the identification of further values above certain thresholds. Conversely, a calmer and safer behavior is identified in those drivers who, in addition to respecting the road limits, apply gradual pressures on the pedals, gradual steering and have low consumption.

Several scientific presentations demonstrated the close relationship between driving impetuosity, compliance with speed limits, and the level of fuel consumption. One of these is [8], which comes to show that the fuel consumption rate of crash-involved vehicles was higher than that of vehicles not involved in crashes, due to their higher speeds. Another example can be found in the study of [15] which, through the table 1.2 following a road safety program, shows that there has been a decrease in both accidents and emissions of CO_2. Close to them also a further study ([23]) has identified how an unsuitable and aggressive driving develops higher consumption in the different vehicles (1.1), an example of it can be seen in [21] in which is stated that an aggressive driving approach has, in average, 25% of additional fuel consumption for short trips, 22% for medium-length trips, and 20% for long trips. Finally, numerous European tests have identified how rapid accelerations and sudden braking can increase consumption by 40 % as well as the percentage of accidents.

All the work starts from some analysis introduced in some related studies, where, through electronic devices (OBD II-On Board Diagnostic), the possibility of accessing many of the vehicle's values by communicating directly with its control unit, has

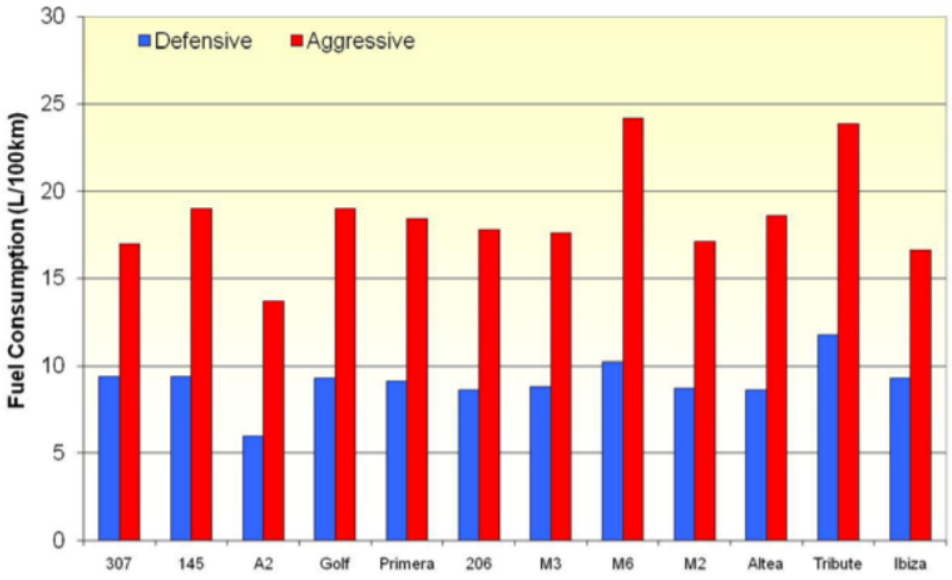


Figure 1.1. Consumption comparison between aggressive and defensive driving styles. From Tzirakis, Zannikos and Stournas (2007)

Road safety program	Fatal crashes saved per annum	CO_2-e reduction factor	CO_2-e saved per annum (k tonnes)
Random road watch	80	More consistent driving behaviour, lower speeds	40
Random breath testing	210 [1]	More consistent driving behaviour	40
Speed cameras	82	10% average speed reduction	400
50 km/h local street speed limit	19	10% average speed reduction on 50km/h routes	33
Fatal 4 public education campaign	20 [2]	More consistent driving behaviour	67

Figure 1.2. Fatality and CO_2 savings from road safety programs in Queensland 1998-2000. From Meers and Roth (2001)

been introduced.

At the same time, the possibility of combining this information with the ones from the tools and sensors present in our smartphones was then exploited. This, to assist those who are the various alerts and safety mechanisms that modern cars offer to drivers. Through this set of information, a group of standards has therefore been developed over the years; they roughly drew the boundary between sporty driving, characterized by higher fuel consumption, and a calmer driving, that allowed you to have more control over this consumption.

As a consequence, some scores have been introduced to help the end-user to have a conception of how unsuitable his guide was.

In this study, we concentrated to design an algorithm both looking at thresholds and scores mentioned above, and also introducing others of them based on further and new mechanisms. Among these new elements, the first is the identification of the main features (4.2), (4.3) that reflect the distinction between a good driving approach, from a dangerous one in a general way. To do this, several datasets containing very heterogeneous information were taken into account, to find those factors that in most of the situations could be used as wake-up calls for classification driving style.

After that the main candidates were identified as "main features", through other data correlation processes in the different road routes, we concentrated on finding those secondary characteristics (5.2.2) which, if used in support of the former, could lead to a higher level of precision in the classification. Thanks to these further studies, it was possible to find many elements supporting the main identifiers, since in the various dataset the information was different and most of the elements identified in one of them did not apply to the others. Once the identification of these factors was completed, the aforementioned scores (5.1) and thresholds (5.2.1) were introduced and applied on these main and secondary features to make a more detailed identification of the different driving styles. Subsequently, there was the introduction of a new mechanism of classification based on the analysis of the primary features' peaks number. It will be used as a support but also as a verification element to ascertain whether the classifications based on scores and thresholds had been done correctly or not (5.4.1).

These peaks were also used as an identifier of the route's type, concentrating on the coincidence of them between the various primary features (5.4.3).

Then, we have used an unsupervised classification approach that in other studied seems to provide good clustering results for the driver characterization, (6) trying to understand if it could, at least partially, emulate the classifications that we obtained. As for the last sections, an attempt was made to partially classify those trips, using firstly the information obtainable only via smartphones (7.1), and then relying on the ones only obtainable from the communication with the machine's control unit (7.2).

Chapter 2

Related works

As we can imagine, the great importance of the topic means that the studies applied to it are in significant number. It was, therefore, possible to analyze different approaches related to the subject, to obtain basic information to propose an innovative and different work. Among the numerous treaties relating to the problem of driving safety and vehicle consumption, we identify a group of them strongly related, to the approach developed in this document.

An example is an article written by Carmona J., Garcìa F. et al. (2015) [4], in which through a combination of software and hardware components, like Raspberry Pi 2 with a shield as CAN-Bus adapter and a GPS with IMU, aims to provide real-time behavior detection both in urban, interurban and highways scenarios. The results highlight some elements, like the presence of higher values in the standard deviation for the vehicle velocity in aggressive driver behavior (especially in extra-urban contexts), a great variation in the use of the engine with high changes in the revolutions and so a higher standard deviation even in the vehicle engine RPM, and a higher standard deviation value related to the steering wheel movement (angular velocity) for the aggressive behavior, not only in an urban environment but in all the environments. A strengthening of this approach is proposed by Oussama Derbel et al. (20115) in [6] in which the number of parameters taken into consideration, obtained through an analysis carried out on different types of routes, increases introducing the jerk, i.e. the variation of acceleration during a period. Here too, the results confirm that the incorrect driving is strongly characterized by fairly high jerking values which follow the trend of an aggression indicator also used in our study, the Aggressiveness Score.

Another work that focuses on analyzing the various accelerations made by drivers, is that of Chen Y. et al. (2018) [20]. In particular, it takes care of obtaining information from the acceleration sensor present in smartphones, integrating it with an OBD connected to the vehicle, to analyze the driver's behavior, classify events of various kinds, and establish thresholds useful to affirm whether the driver applies aggressive driving. Here, the distinction of the acceleration events is carried out through a specific module called Records Analysis Module, which receives information on the accelerations on the different axes such as Acceleration-X, Acceleration-Y, and Acceleration-Z. Once that this information has been obtained it transfers them in a data structure. From this data structure, the module will then process and

analyze the information by defining a model for the specific driver. The result of this approach emphasizes that the rapid decelerations in braking are accompanied by a value beyond the negative threshold of Acceleration-Z, and that the rapid steering and maneuvers are accompanied by values of Acceleration-X beyond the positive threshold. On the same side we found the study carried out by Antoniou C., Gikas V. et al. (2014) [2], in which, through information relating to the accelerations on the different axes X, Y and Z, are created some clusters containing the different types of drivers, classified through the K-means approach.

Similar to this approach is the one developed by Martin S., Van Li M. et al. (2013) [24], which uses the vehicle's inertial sensors, obtainable via CAN-bus communication, to keep track of all braking and steering events related to accelerations. Through these events, the study creates a classification of the driver into two distinct groups. This is done, firstly using an unsupervised learning technique such as K-means clustering, and then applying a supervised learning technique such as the SVM (support vector machine), based on the subdivision of the starting dataset into two subsets of routes, one in which the trips with the drivers already labeled with suitable or less suitable behavior (training subset) are inserted, and one in which the algorithm independently labels the driver (subset test). This use of supervised and unsupervised approaches is also strongly present in the work of Wang Z., Liu F. et al. (2018) [25]. In this study, once the various information like speed, acceleration, and steering angle are obtained, the K-nearest neighbor, artificial neural network, and support vector machine are then applied. The first approach focuses on the overlapping between average values of the trip features belonging to different sets, treating them as proximity points, the second simulates the communications of human neurotransmissions by constantly inserting elements to obtain an update of the traversing weight and connections between them, and the third divides the information into spaces with wide dimensionality, through the calculation of the categorization distances and specific kernel functions.

Along the same lines as the previous ones, also the work of Fugiglando U., Massaro E. et al. (2018), proposes to classify, through an unsupervised approach, the different drivers based on their attitude on the road. To do this, the study carefully looks at elements such as the gas pedal position, brake pedal pressure, steering wheel angle, steering wheel momentum, velocity, RPM, longitudinal and lateral acceleration obtained through CAN-bus technology, on which it applies two different classification methodologies. The first approach is based on the identification of the signal values in the different ranges, focusing on the local maxima, averages, and variances of these samples, the second one instead uses the K-means approach, applied to the normalized datasets, experimenting with different numbers of clusters.

Of other veins, are the studies presented by Nousias S., Moustakas K., Bressan E., et al. (2018). Some of them ([10]) focus on unifying the various existing real-time classification methods, with some innovative methods to create algorithms that manage to outline a driver's score. These scores consist of various parameters, such as the braking parameter, shift-up parameter, cruising parameter, RPM parameter, and many others, used to create identifiers that are as complete and classifying as possible. Alongside the creation of these scores, in [11] a particular human factor is also addressed, the driver's heartbeat. In this study, the heartbeat of the various drivers is analyzed, submitting them to drive on different types of roads,

with various climatic conditions, under stresses of different nature such as cognitive stress, emotional stress, and sensorimotor stress. This shows how the mean heart rate coming from a stressful situation is generally higher than the mean heart rate presented in the normal driving condition, and defines an estimation of the risk of a particular situation through the mechanism called Risk Assessment module.

A further study, that deals with human factors in-depth, is the one of Haworth N., and Simmons M., ([8]), in which all those human components obtainable both via smartphone and OBD adapter are analyzed. Because human factors are mainly responsible for the high emissions of polluting factors and road accidents. All this information, obtained in disparate ways, is then used by a smart algorithm that produces a score for the driver by classifying his driving attitude. Focusing on the pollution factor, we cannot fail to mention the work of Heijne V., Ligterink N. and Stelwagen U. (2017) [9], in which, from a dataset consisting of over 13500 driving hours, different approaches have been developed both regarding the distinction of the type of road faced, and regarding the classification of eco-driving, based on averages, deviations, and correlations of elements like the speed limits, velocity, engine load, braking, country infrastructure, gear-shifting, and road type.

The document by Martinelli F., Mercaldo F., et al. (2018) [13] deals with a completely different methodology. In particular, a driver profiling mechanism is introduced, not to identify recklessness attitudes in him, but to identify any vehicle theft by noticing a different driving style than usual. This profiling is carried out through a machine learning algorithm that works on a set of over 50 features, some of which have already been mentioned, and others completely new such as Intake Air Pressure, Engine Soaking Time, Torque of Friction, Fuel Pressure and many others. This big set, then, will be reduced to a group of main features through the PCA approach. From these selected features, the mechanism will then offer a model that identifies the individual driver. Subsequently, once the model has been developed, a second predictive phase will analyze the feature vector of the most recent guide, managing to distinguish whether the driving behavior is the one of the owner or a stranger's one.

Finally, we identify a study similar to the initial phase of our work, the paper of Massoud R., Bellotti F. et al. (2019) [14]. It comes to provide a driver identification mechanism through the individuation of changes in RPM, throttle position, jerks, car speed, and other features sampled through OBD smartphone sensors, GPS location, and vehicle sensors. Once the driver's attitude has been classified, the mechanism provides feedback to counteract risky attitudes, such as sudden maneuvers and steering, and expensive attitudes from the fuel point of view, introducing suggestions regarding the fuel economy and road safety.

Chapter 3

Dataset description and preprocessing

In this chapter we will present all the datasets taken into account in our analysis, indicating which study they come from and highlighting the features that characterize them.

In a second moment, we will then focus on the cleaning processes carried out, each ad hoc for the specific dataset, indicating which information will be set aside and for what reason. This, to reduce errors due to the composition of the information, and to obtain datasets that are as structured and compact as possible.

3.1 Dataset description

3.1.1 Sparkworks' Dataset

The first dataset that we introduce was obtained from a SparkWorks Console's secured section, and is related to the studies applied in [10],[11] and [12], [22],[1], [18], [19], [17]. It is formed by 826 files of different length trips, performed by many users in different cities among all the European area. Each of these trips contains a set of features that can be very useful to understand some behavioral patterns in the driving style of the drivers. The most important elements present in these trips are Latitude, Longitude, Acceleration on X-axis, Acceleration on Y-axis, Acceleration on Z-axis, Heartrate, Vehicle Speed, Engine RPM, Throttle Position, Fuel type, Fuel Level, Fuel Consumption Rate, Air Intake Temperature, Gear, Road type, Traffic Confidence, Humidity, Pressure, and Weather.

From these values, first of all, we have to understand how much these data are really useful, taking into consideration their real presence in the dataset and their values variation during the trip time. For this reason, we have to classify them to understand which of them is effectively helpful for our study. Here we present two methods of classification:

- **Data quality of each trip**

  The first approach of data quality individuation could be related to the following elements

- Number of values of a specific element
- Values' update frequency within a window of 1 minute
- Analysis of values that effectively change against the ones that are fixed

Developing a score for each trip based on a formula similar to the following one:
$TripQuality = (RPMValues * \alpha_1) + (SpeedValues * \alpha_2) +$
$(ThrottleValues * \alpha_3) + (RPMFrequency * \beta_1) + (SpeedFrequency * \beta_2) +$
$(ThrottleFrequency * \beta_4) + (DifferentRPMValues * \gamma_1) + (DifferentSpeedValues *$
$\gamma_2) + (DifferentThrottleValues * \gamma_3)$

It is a good starting idea, but we must notice that in this way we cannot obtain a specific trip's quality score since the weights $(\alpha_1, ..., \gamma_3)$ are not scientifically determined. To classify a trip of the dataset we should understand that we need a threshold for each specific value, but we have to find these thresholds.

An idea is to identify the minimum of sampling frequency for each determined value type as a threshold, and use this threshold to give a score to the specific value type, where the sum of the scores gives the general rating of the trip. After having discarded all the files containing paths of less than 10 minutes of driving, due to their shortness, we can create a score for each value as follows.

- For Speed, RPM, Throttle and Gear values we base our score on their frequency of sampling, starting, from a minimum of 1Hz to a specific value that is different for each of them depending on the general distribution of the sampling frequency of all the trips' value. For example, for Speed and RPM the maximum frequency with score 10 is fixed about to 6Hz, while Throttle and Gear obtain a maximum around 5Hz.
- For the Heartrate we have identified the score from 1 to 10 for a sampling frequency between 0,033Hz and 0,33Hz.
- Finally we have two negative scores, one related to the Heartrate values over 180 bpm that has a score from 0 to -5 for the magnitude of this erroneous value (from 2 to 25 erroneous values in the same trip), and one related to Throttle values. This because many trips have a fixed value of the throttle position so they have to be penalized with a score from 0 to -5 depending on how many times the same values repeat themselves.

This allows us to analyze the different distribution of values to catch some particular correlation between features that are presented in 4.2,4.3, and 5.2.2. In the image 3.1 we present the distribution of the value, taking into consideration only the scores having a value greater than 0, of the trip longer than 10 minutes.

- **Geographical classification per type of paths**

  For what regards the Geographical classification we have used the latitude and longitude values of each path, to recreate through the Google Maps API the trip done by each user. This allows us to understand that most of these trips are performed in Greece, England (in particular in the Leeds area) and France (in particular in the city of Versailles).

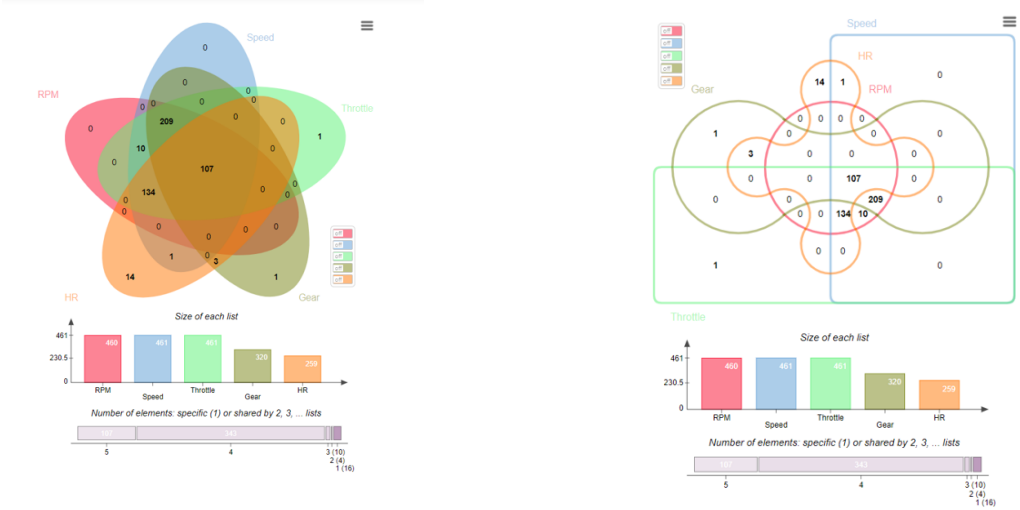


Figure 3.1. Venn representations of the SparkWorks dataset

Figure 3.2. An example of a trip in the Seoul Dataset

3.1.2 Seoul Dataset

In this dataset, we already know that all the trips are performed in the city of Seoul (3.2), with the same car, executed on four different round-trip paths (i.e., between Korea University and SANGAM World Cup Stadium) for about 23 hours of total driving time. [13]

Then we also have that each of these paths is accomplished by a specific driver with an ID from A to J for a total of 10 drivers, where each of them has two groups of paths divided in group 1 and group 2.

The work done on this dataset was one of splitting the general file creating a set of subfiles divided per driver and per path to understand better the distribution of the values and to eventually find a lack of information for specific features.

However, contrary to the previous dataset, each of these trips is filled with readable values with no empty values, that allow us to say that it is no needing of developing a classification on the quality of the trips except for their length, that is,

#	Feature	Description
2	*Accelerator_Pedal_value*	This sensor registers the movement of the accelerator pedal: Accelerator pedal opening angle percentage as determined by the accelerator position sensor.
3	*Throttle_position_signal*	The relative throttle position sensor is used to monitor the throttle position of a vehicle
4	*Short_Term_Fuel_Trim_Bank1*	Fuel trims are the percentage of change in fuel over time in short term
5	*Intake_air_pressure*	This data is used to calculate air density and determine the engine's air mass flow rate
6	*Filtered_Accelerator_Pedal_value*	ECU's filtered accelerator pedal opening angle percentage as determined by the accelerator position sensor.
7	*Absolute_throttle_position*	Actual position of the throttle
8	*Engine_soacking_time*	Duration of time a vehicle's engine is at rest prior to being started.
9	*Inhibition_of_engine_fuel_cut_off*	The fuel cut-off control system is responsive to a brake switch signal and an engine speed signal having a value above a fuel recovery threshold to decrease the value of a fuel cut-off threshold to again perform the fuel cut-off even in the normal fuel recovery range. This value represents the inhibition of engine fuel cut off
10	*Engine_in_fuel_cut_off*	This value represents the inhibition of engine fuel cut off, i.e., fuel cut-off threshold.
11	*Fuel_Pressure*	Effective pressure is the actual applied pressure for the injector, and is the pressure differential across the injector.
12	*Long_Term_Fuel_Trim_Bank1*	Fuel trims are the percentage of change in fuel over time in long term.
13	*Engine_speed*	It is also called engine's RPM, i.e., Revolutions Per Minute. In other words it is the number of revolutions the crankshaft makes per minute.
14	*Engine_torque_after_correction*	The value after correcting the torque to which an engine is adjusted before a gear disengagement.
15	*Torque_of_friction*	Friction torque is the torque caused by the frictional force that occurs when two objects in contact move.

Figure 3.3. Seoul Dataset Features (1)

in average, around 1700 samplings.

This set of trips is very useful because contains a lot of features that are described in the images 3.3, 3.4, 3.5.

Another interesting factor is that among them, many features are not present in the other datasets like:

- Accelerator Pedal value
- Wheel velocity (for each wheel)
- Engine coolant temperature
- Transmission Oil Speed
- Torque of Friction
- Intake Air Pressure

#	Feature	Description
16	*Flywheel_torque_interventions*	The flywheel stores energy when torque is applied by the energy source, and it releases stored energy when the energy source is not applying torque to it. The value represent the flywheel torque after torque interventions.
17	*Current_spark_timing*	The time to set the angle relative to piston position and crankshaft angular velocity that a spark will occur in the combustion chamber near the end of the compression stroke.
18	*Engine_coolant_temperature*	The temperature of the engine coolant of the internal combustion engine
19	*Engine_Idle_Target_Speed*	The desired idle RPM in relation to coolant temp.
20	*Engine_torque*	Engine torque is also related to the gearing. The lower the gear, greater is the pulling ability of an engine and hence greater the torque that this value represents.
21	*Calculated_LOAD_value*	This value indicates a percentage of peak available torque.
22	*Min_indicated_engine_torque*	Minimum *Engine_torque* value
23	*Max_indicated_engine_torque*	Maximum *Engine_torque* value
24	*Flywheel_torque*	The value represent the flywheel torque
25	*Torque_scaling_factor*	This value is described as how flexible or how much force can be expressed in a given gear when the driver scales the gear.
26	*Standard_Torque_Ratio*	This value is described as how flexible or how much force can be expressed in a given gear.
27	*Requested_spark_retard_angle*	The transmission control unit (TCU) controls modern electronic automatic transmissions. This value computes the requested spark retard angle from TCU.
28	*Requests_engine_torque_limit*	This parameter monitors the request to engine torque limits (ETL) by TCU
29	*Requested_engine_RPM_increase*	This parameter monitors the TCU requests related to the RPM engine increasing
30	*Target_engine_speed_used_in_lock-up_module*	It monitors the lock-up valve, used to shut off the signal pressure line of pneumatic actuators.
31	*Glow_plug_control_request*	It monitors the request to check the glow plug
32	*Activation_of_Air_compressor*	The value of the air compressor's working.

Figure 3.4. Seoul Dataset Features (2)

#	Feature	Description
33	*Torque_converter_speed*	A particular kind of fluid coupling that is used to transfer rotating power from a prime mover
34	*Current_Gear*	The engaged gear
35	*Transmission_oil_temperature*	The value of the temperature of the fluid inside the transmission.
36	*Wheel_velocity_front_left-hand*	The speed of the front left hand wheel
37	*Wheel_velocity_rear_right-hand*	The speed of the rear right hand wheel
38	*Wheel_velocity_front_right-hand*	The speed of the front right hand wheel
39	*Wheel_velocity_rear_left-hand*	The speed of the rear left hand wheel
40	*Torque_converter_turbine_speed_-_Unfiltered*	A torque converter is a type of fluid coupling that is used to transfer rotating power from a prime mover, such as an internal turbines in this case.
41	*Clutch_operation_acknowledge*	It is responsible to signalize when a clutch operation happens.
42	*Converter_clutch*	It is responsible for activating the torque converter clutch to prevent slipping at highway speeds
43	*Gear_Selection*	It represents the gear selected by the sensor
44	*Vehicle_speed*	It represents the current speed of the vehicle
45	*Acceleration_speed_-_Longitudinal*	It represents the value related to the acceleration speed longitudinal
46	*Indication_of_brake_switch_ON/OFF*	It indicates whether the brake indication is on or off
47	*Master_cylinder_pressure*	The pressure of the master cylinder, a control device that converts non-hydraulic pressure into hydraulic one.
48	*Calculated_road_gradient*	This value computes the slope of the currently traveled road
49	*Acceleration_speed_-_Lateral*	It consists in the acceleration value that a curving car manifest
50	*Steering_wheel_speed*	This value represents the wheel speed when steering
51	*Steering_wheel_angle*	This value represents the wheel angle when steering

Figure 3.5. Seoul Dataset Features (3)

- Engine soaking Time
- Acceleration speed Longitudinal
- Acceleration speed Lateral
- Steering wheel speed
- Steering wheel angle

They are helpful to develop different studies regarding the driver behavior's classification and to understand new correlations between the features.

Anyhow, we have a little problem with this dataset, even if we know the general paths environment we don't have any GPS location, neither latitude nor longitude values, and this does not allow us to compare the different sections of the paths.

3.1.3 Cephas Dataset

This dataset ([3]) is composed of three main files that are used to develop three different analyses on the data. The *first one* contains a set of drivers that have carried out different paths with different cars, the *second one* is formed by a little number of users, performing a trip with the same car on the same path, and the *last one* is, again, formed by a group of users that, with the same car, have realized the same route, with the difference that here the drivers set is way larger than the previous one.

To work on these dataframes, we take the general trips containing all the data and then we split them in a subset of trips divided for the type of experiments.

- **The first "Sub-dataset"** contains 14 trips based on a set of interesting features like *Engine Power, Barometric Pressure (KPA), Engine Coolant System, Fuel Level, Engine Load, Engine RPM, Intake Manifold Pressure, Fuel Type, Fuel Pressure, Speed, Throttle Position*
- **The second one** formed by 4 trips presents other features like *Latitude, Longitude, Altitude, Barometric Pressure, Fuel Level, Engine Load, Engine RPM, Manifold Pressure, Engine Coolant Temperature, Speed, Throttle Position*
- **The last "Sub-dataset"** is probably the most useful of the three because it is formed by 19 trips with all the interesting features of the previous sub-dataset.

The first sub-dataset can be used for strengthening the SparkWorks dataset, where can be used to add other paths to the already big set of them, increasing the possibility of discovering relationships among features. The one formed by 19 trips, instead, can be used side by side with the Seoul one for determining the behavioral characteristics of the drivers, comparing the different driving styles on the same road trip.

There is only a problem with this Cephas dataset, after the splitting in different trips based on the drivers' ID we have some very small paths, and this makes them useless, so in the successive phase of cleaning and deletion, some of them will be discarded.

3.2 Data cleaning and preprocessing

As we can imagine, having a lot of information to analyze, makes it probable to find some data that is not correct due to a bad-sampling, transmission error, error on data-type reporting and other reasons. Near the non-correct information, we have also to distinguish the missing information that can affect the trips. This scenario is even more important because we are talking about three different datasets that have different sampling mechanisms and different features sampled. Let's see in detail which type of deletion and cleaning are necessary for each of these datasets.

Regarding the **SparkWorks dataset**, we have said that it is formed by a huge number of trips developed in Europe. We, firstly, need to discard all that trips having a duration time less than 10 minutes, then, we have to clean up all the erroneous values, like the ones in the Heartrate feature, where the Bpm values higher than 180 are substituted with the first previous good value.

Another important action is the deletion of some dataset columns, in particular the empty ones, the ones that are not so useful for our study, and the Throttle Position's one if it has less than 5 different values during all the samplings. The less useful features deleted in this dataset are the following:

- RR Interval
- Short and Long Term Fuel Trims Bank
- Fuel Pressure
- Air Intake Temperature
- Road OpenRL
- Current Flow Speed & Travel Time
- Free Flow Speed & Travel Time
- Risk Probability & DIL

For what concerns the **Seoul dataset**, the clean-up performed is based on the deletion of unneeded features and the deletion of the features that are empty or always with the same value as:

- Short and Long Term Fuel Trims Bank
- Engine Soaking Time
- Engine in Fuel Cut-Off
- Current Spark Timing
- Flywheel Torque
- Master Cylinder Pressure
- Activation of Air compressor

- Class
- Torque Converter Turbine Speed

In **Cephas dataset**, we have applied the same approach of the Seoul dataset, deleting all the unneeded features, empty columns, and features that have always the same value. Near to this, we have dropped the trips that have a small number of samplings. In particular for the set of 14 trips, once cleaned, we've discarded the user's trips with ID: 5,10, due to a too-small dimension (less than 15 Kb), and for the set of 19 trips, we've discarded the paths with ID: 3,9,17.

The deleted features for the **Cephas 14** subset were:

- Ambient Air Temperature
- Long Term Fuel Trim Bank
- Vehicle ID

While for the **Cephas 19** subset were:

- Ambient Air Temperature
- Air Intake Temperature
- Long and Short Term Fuel Trim Bank
- Vehicle ID
- DTC Number
- Trouble Codes

Chapter 4

Preliminary data analysis

In this chapter, we will present in detail all the types of features that from now on will be used during the study, distinguishing them both from the methodology used to obtain them, and from their strength and presence in the various datasets.

Alongside this, two different methodologies applied to these features will be addressed, to identify links and correlations between them, based on their performance in the various paths taken by drivers. In the first approach, the variance, the standard deviation and the average of the values of the individual features will be calculated for each path. After this, will be applied to them the approaches indicated in 4.2, to identify similar trends among them factors. In the next method, instead, we will focus on a double correlation, the first, concerning the vector components of the features in their entirety, concerning the single path, and the second applied on the entirety of the paths based on the values obtained in the previous step.

4.1 Categorization of features

As we mentioned earlier, different types of features were taken into consideration during our study. These differences are related to two main factors; the first concerns the methodology applied to carry out the samplings, while the second focuses on their presence in the datasets and on their classification strength showed in the various correlation methods.

The first categorization divides the features between the set of those obtainable via any fairly recent smartphone, and the set of those are reachable only through direct communication with the central unit of the vehicle. In our case, this communication is made through an OBD (On-Board Diagnostic) electronic device which, once inserted in the respective plug inside the vehicle, provides some data coming from the central unit.

So, there is an initial distinction between:

- **Smartphone Features**
- **OBD Features**

Which are dealt with in detail in 7.1 and 7.2

The second categorization divides the features between the group of those that are present in each dataset and that offer interesting correlations in both the methods addressed in 4.2, and 4.3, from the ones that have a variable presence, depending on the dataset taken into consideration, but which have a particular correlative strength.

The following are then presented:

- **Primary Features**
- **Secondary Features**

This distinction is not only formal, the partial dominance of the former over the latter, due to their constant presence in each dataset, is, in fact, flanked by the fact that these features characterize mostly the driving behavior (like the RPM, the vehicle speed and the throttle position). However, we must not underestimate the secondary features, which can, thanks to their correlative strength, be used alongside those of the primaries to enhance the classification of drivers.

4.2 Features' normalization and coefficient correlations

Let's look at the most interesting features of each dataset, to calculate their variance, standard deviation, and average value [4]. After that, we will apply different methods of correlation to discover if some factors are strictly connected.

The main correlation approaches will be

- **Pearson Correlation coefficient**

  Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name. It can be used to summarize the strength of the *linear relationship* between two data samples. In other words, it is the normalization of the covariance between the two variables that gives an interpretable score. The coefficient returns a value between -1 (full negative correlation) and 1 (full positive correlation) where a value of 0 means no correlation. [27]

- **Kendall Correlation coefficient** In statistics, the Kendall rank correlation coefficient, commonly referred to as Kendall's τ coefficient is a statistic used to measure the ordinal association between two measured quantities. A τ test is a non-parametric hypothesis test for statistical dependence based on the τ coefficient.

  It is a measure of rank correlation: the similarity of the orderings of the data when ranked by each of the quantities. Intuitively, the Kendall correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low

when observations have a dissimilar (or fully different for a correlation of -1) rank between the two variables. Both Kendall's τ and Spearman's ρ can be formulated as special cases of a more general correlation coefficient [26]

- **Spearman Correlation coefficient**

  In statistics, Spearman's rank correlation coefficient, or Spearman's ρ, is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function. The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses *monotonic relationships (whether linear or not)*. If there are no repeated data values, a perfect Spearman correlation of $+1$ or -1 occurs when each of the variables is a perfect monotone function of the other. Intuitively, the Spearman correlation between two variables will be high when observations have a similar rank between the two variables and low when observations have a dissimilar (or fully opposed for a correlation of -1) rank between the two variables. Spearman's coefficient is appropriate for both continuous and discrete ordinal variables. [28]

Let's start focusing on the ***SparkWorks dataset***, here we've evaluated the above cited calculations related to Acceleration on both X,Y and Z axes, Throttle, Gear, Speed, Heartrate and RPM. From these values we obtain the correlations showed in 4.1, 4.2,and 4.3.

The interesting elements coming out from these correlations are, obviously, different, depending on the applied method.

For example, in Pearson's approach (4.1) we can notice

- The correlation among the axes of the different accelerations
- The quite strong relationship between variance and standard deviation of the Speed factor with the RPM factor's ones (red square)
- The quite strong relationship between the average Speed and average Gear values with the average Throttle values (purple rectangle)
- The less strong relationship between the Heartrate feature's values with the Acceleration on X-axis' and the Throttle's one standard deviation (orange rectangle)

In the Kendall's one (4.2) we find other interesting elements

- Again, the correlation among the Accelerations different axes' values
- A weaker relationship among variance and standard deviation of the Speed and RPM factors (red square)
- The not so strong relationship between the Acceleration X and Y features with the average Throttle values (black rectangle)

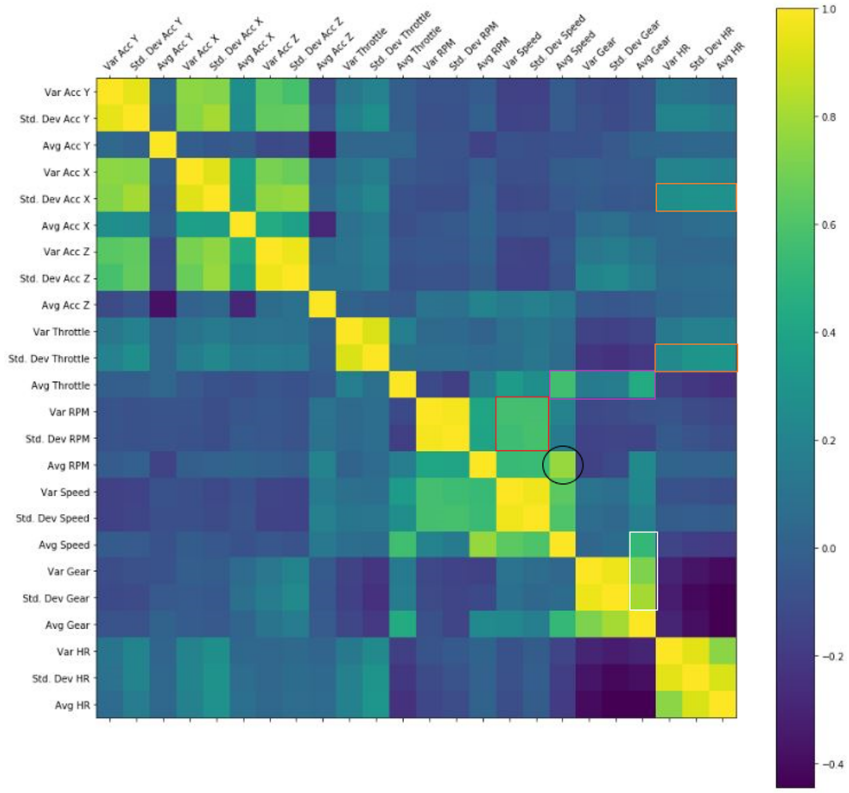


Figure 4.1. Pearson Coefficient on SparkWorks dataset features' normalization

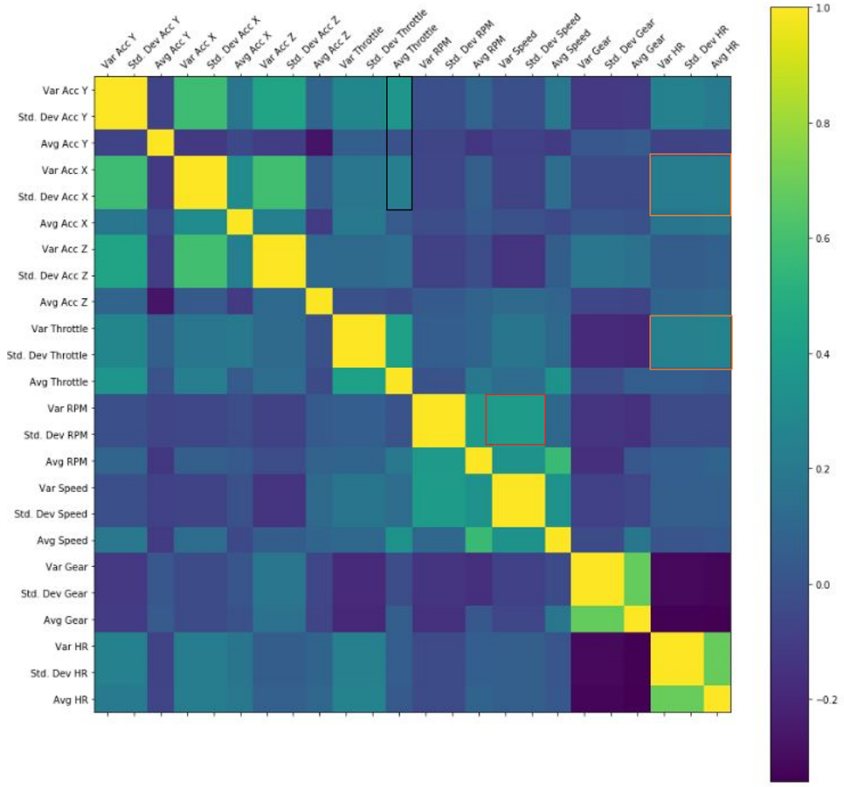


Figure 4.2. Kendall Coefficient on SparkWorks dataset features' normalization

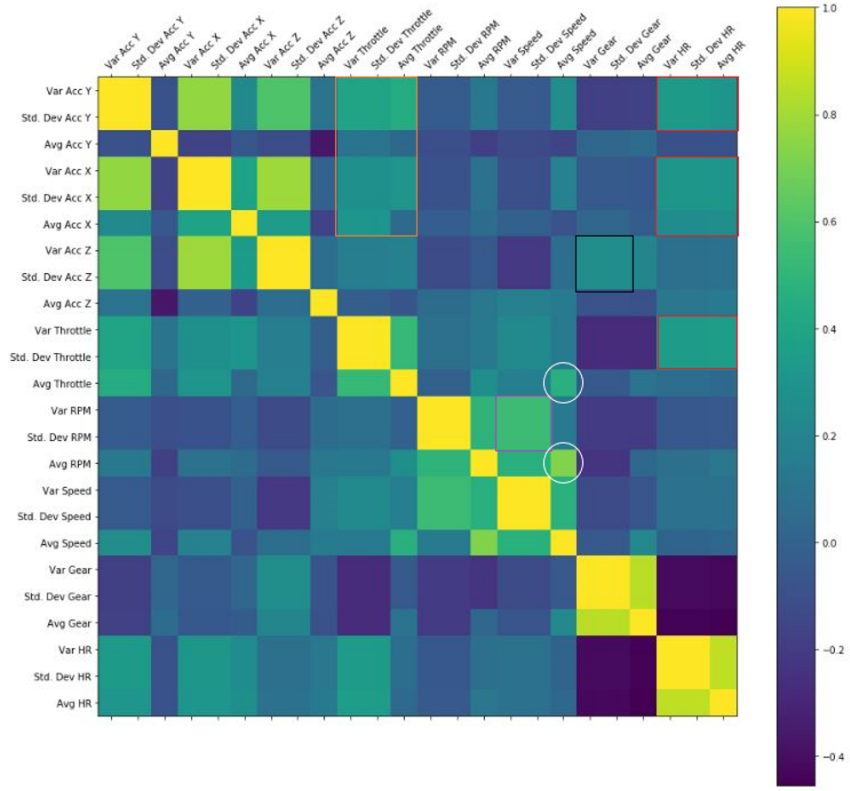


Figure 4.3. Spearman Coefficient on SparkWorks dataset features' normalization

– The quite strong relationship between Heartrate features with the Acceleration X and Throttle features (orange rectangle)

– A quite strong relationship between Gear features and variance and standard deviation of the Acceleration on Z-axis

Then, we switch to the result coming from the Spearman approach (4.3), in which we can notice

– Again, the correlation among the Accelerations different axes' values

– The quite strong relationship variance and standard deviation of the Speed and RPM factors (purple square)

– The quite strong relationship between the average Speed values and average Gear values with the average Throttle values (white circles)

– The less strong relationship between Heartrate feature's values with the Acceleration Y, X and Throttle features' values (red rectangles)

– The not so weak relationship between Throttle values and Acceleration on Z-axis ones

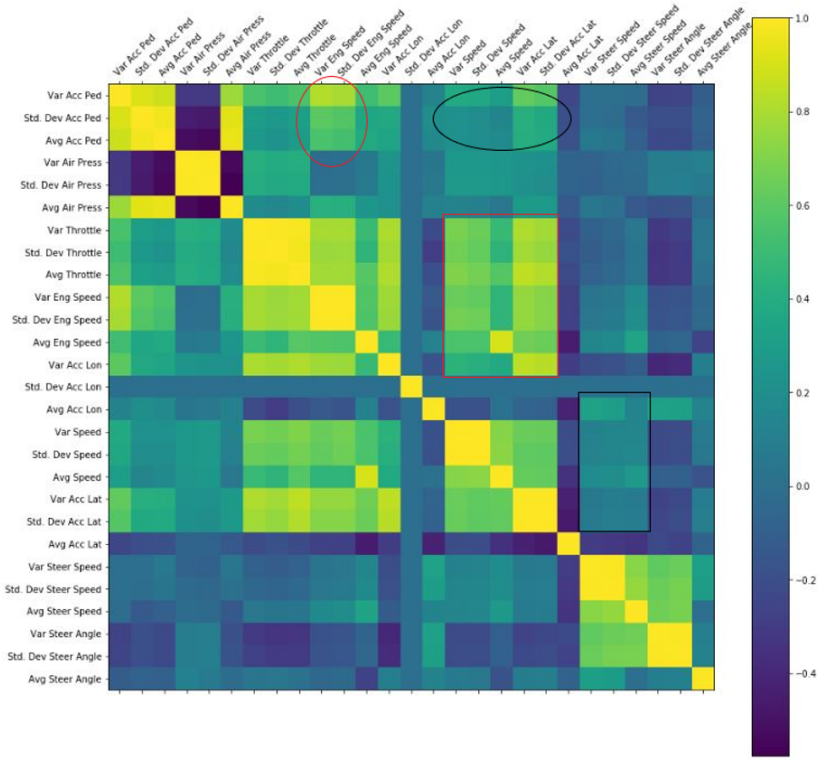


Figure 4.4. Pearson Coefficient on Seoul dataset features' normalization

Once that we have analyzed quite deeply the various scenarios related to the SparkWorks dataset is time to focus on the others, in particular to the ***Seoul*** one.

In this dataset, we have to take into consideration some new types of features, like the accelerator pedal value, intake air pressure, absolute throttle position, steering wheel speed, steering wheel angle. Near to them, however, there is a set of common elements formed by the engine speed, vehicle speed, acceleration speed Longitudinal, and acceleration speed Lateral.

Let's start looking at the results coming from the Pearson approach(4.4),

- We notice a strong correlation between Acceleration Pedal and Engine Speed features(red circle)
- A quite strong correlation of the Acceleration Pedal with the Speed feature, and a strong correlation with the Acceleration Latitude feature (black circle)
- There is also a huge correlation between the Acceleration Longitudinal's standard deviation and variance values with the Throttle and Engine Speed features, and a quite strong among the Speed and the Throttle and Engine Speed features (red square)
- A small correlation between the Speed feature and the Steering Wheel feature (black square)

In Kendall scenario (4.5), instead, we have

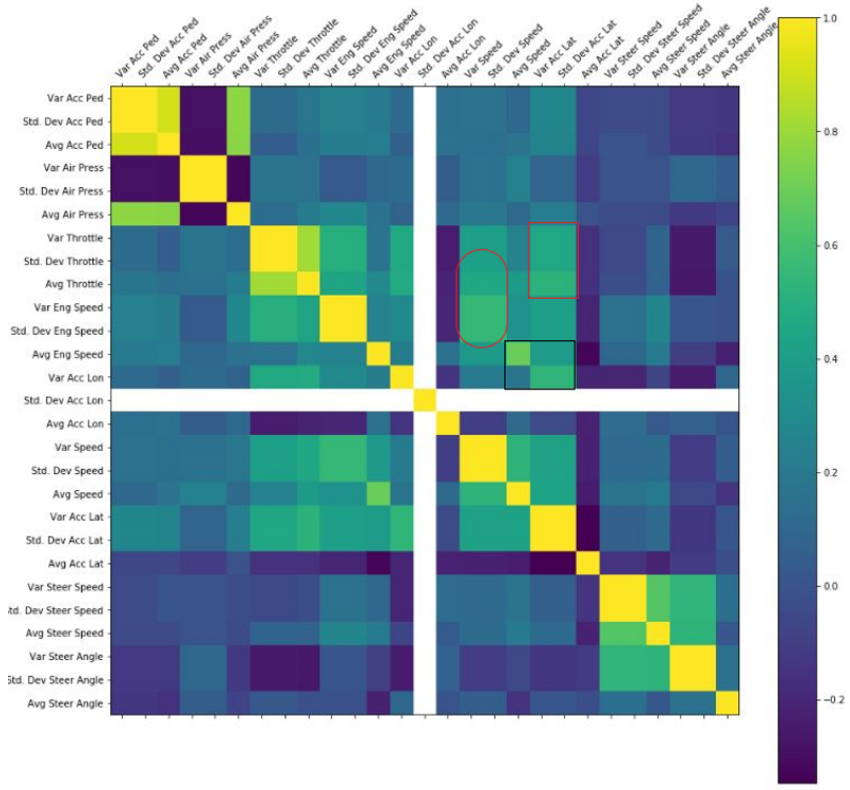


Figure 4.5. Kendall Coefficient on Seoul dataset features' normalization

– A quite strong correlation between the Speed and Engine Speed, and a little mediumly strong correlation between Speed and Throttle feature (red oval)

– A not so weak correlation between Longitudinal Acceleration and Latitudinal Acceleration features (black rectangle)

– A quite strong correlation between Throttle feature and Latitude Acceleration feature

Finally, switching to the Spearman correlation results (4.6), we can notice a lot of elements, so we divide them into two main groups of related features, the one characterized by strong relationships, and the one characterized by less strong relationships.

Strong Relationships

– Between Throttle and Engine Speed features (black square)

– Aong Speed, Throttle and Engine Speed features (red rectangle)

– Between the average Engine Speed values and the average Speed values (black circle)

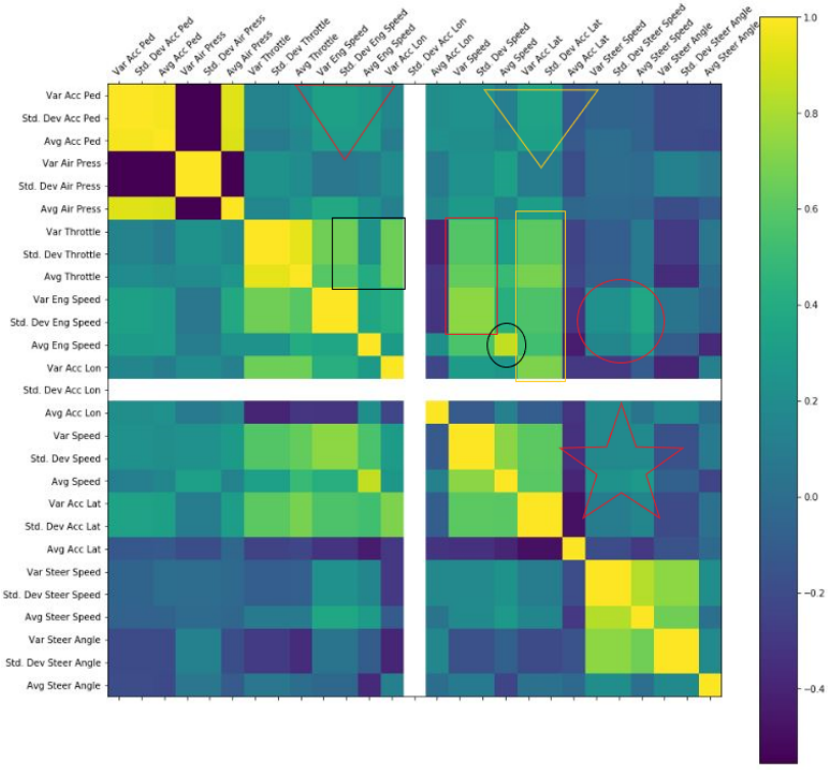


Figure 4.6. Spearman Coefficient on Seoul dataset features' normalization

– Among Latitude Acceleration, Throttle and Engine Speed features(yellow rectangle)

Less strong Relationships

– Between Steering Speed and Engine Speed features' values (red circle)

– The Engine Speed and Acceleration Pedal features' correlation(red triangle)

– The Latitude Acceleration and Acceleration Pedal features' relationship(yellow triangle)

– Between Speed and Steering Speed features' values (red star)

Now is the turn of the ***Cephas dataset***. Here we have to apply two different analyses since the two main sections, the one formed by 19 samples, and the one formed by 14 samples, have some different features that we want to look at.

Firstly we will look at the dataset formed by ***19 drivers*** executing trips on the same path, looking in particular at features like Engine Coolant Temperature, Engine Load, Engine RPM, Intake Manifold Pressure, Speed, Throttle Position. However, the result is not completely satisfying, in fact, from the Pearson approach, we obtain a correlation matrix that is too confused, where it seems that each feature is connected with the other, so, for this reason, we have to discard it.

Different is the situation in the Kendall scenario (4.7), here we have a lot of interesting elements that we will split as before into two sets, quite strong correlations and less strong correlations.

Quite Strong Correlations

- Engine Load feature with Intake Pressure feature (black circle)
- Between Engine Load and Throttle features (red rectangle)
- Intake Pressure feature and Throttle feature(red circle)
- Average Speed values with average Throttle values(white circle)
- Between average Engine RPM values and average Throttle values(white square)

Less Strong Correlations

- Between the Engine Coolant Temperature element and the Engine RPM feature (red oval)
- Between Speed and Engine Coolant Temperature feature (black oval)
- Between Engine RPM and Speed values(black square)
- Among Throttle feature, Speed feature and Average Intake Pressure element(red triangle)

Anyhow, this subdivision of features connection is needed even in the Spearman section (4.8), obtaining two sets of correlations.

Strong Relationships

- Intake Pressure feature with Engine Load feature(black circle)
- Engine Load feature with Throttle feature (red square)
- Intake Pressure element with Throttle feature (red circle)
- Average Speed values and average Throttle values(white circle)
- Average Engine RPM values and average Throttle values(white square)
- Average Engine RPM values and average Speed values (black star)

Less Strong Relationships

- Engine Coolant Temperature element with Engine RPM feature (red oval)
- Speed feature and Engine Coolant Temperature feature (black oval)
- Intake Pressure and Speed features (red star)

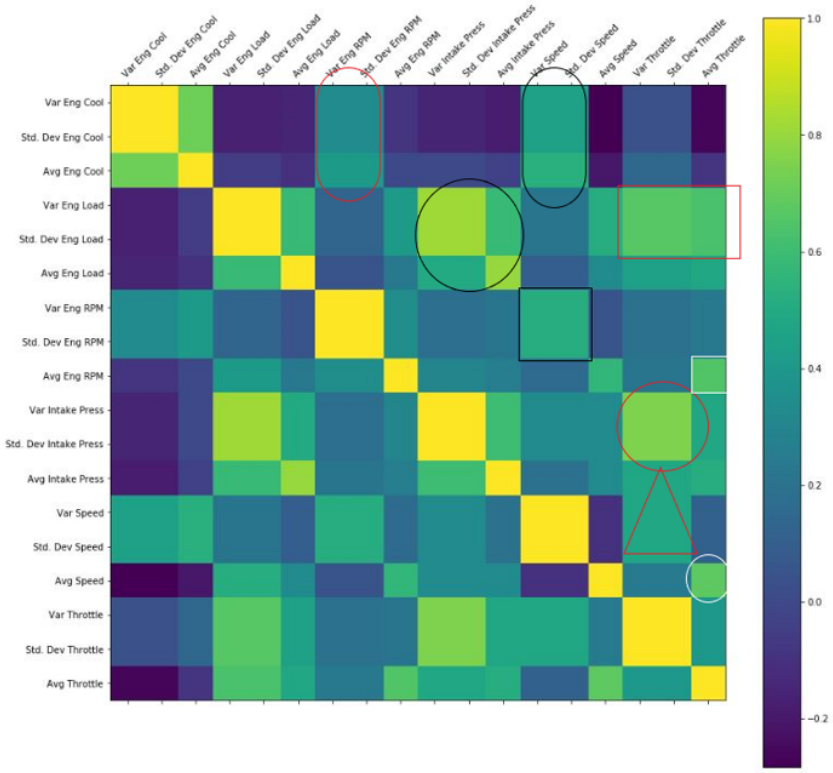


Figure 4.7. Kendall Coefficient on Cephas 19 dataset features' normalization

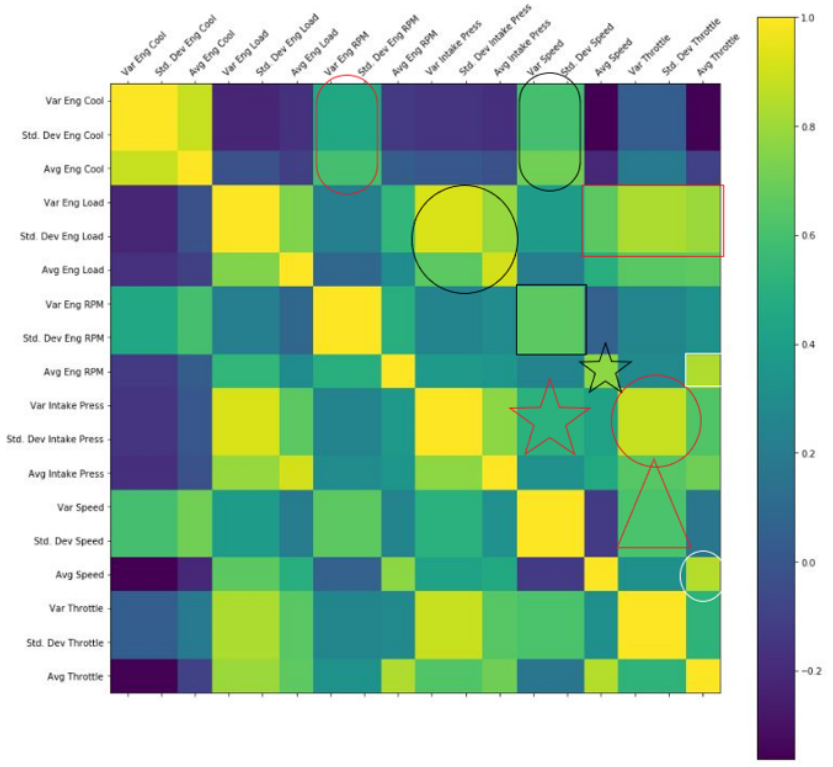


Figure 4.8. Spearman Coefficients on Cephas 19 dataset features' normalization

– Throttle element and Speed features (red triangle)

In the other scenery, we will look at the dataset based on the ***14 drivers' trips*** made up on the different paths, looking in particular at some features like Engine Coolant Temperature, Engine Load, Engine RPM, Intake Manifold Pressure, Speed, Throttle Position and Timing Advance.

The first results that we look at, are the ones related to the Pearson correlation approach(4.9). Here we have some strong correlations like

– The Engine RPM feature with Speed feature (black square)

– A relationship among the Engine Load feature and the Speed, Engine RPM and Intake Pressure features (red square)

– The Engine RPM and Intake Pressure Manifold features(red oval)

And some less strong correlations like

– The Intake Pressure Manifold with the Speed feature's values (white rectangle)

– The Intake Pressure Manifold's elements with the Timing Advance's elements (black circles)

– The average Engine RPM's values with the average Throttle RPM's values (black triangle)

– The Engine Coolant Temperature feature with Timing Advance feature (black star)

Regarding Kendall's Correlation results, except for the strong relationship between the Engine RPM and the Speed factors, all the other elements interconnections are quite weak so we avoid to insert that correlation matrix.

Quite different, instead, is the state of Spearman's correlation (4.10) results, where we have a quite balanced relationship subdivision, between strong correlations and less strong correlations.

In particular, the Strong Correlations are

– The Engine Load feature with the Speed, Engine RPM and Intake Pressure Manifold features (red rectangle)

– The Engine Coolant Temperature element with Timing Advance feature (red circle)

– The Speed factor with Engine RPM and Average Engine Load features (black square)

And the Less Strong ones are

– The relationship between Engine RPM feature and Engine Coolant Temperature feature (red triangle)

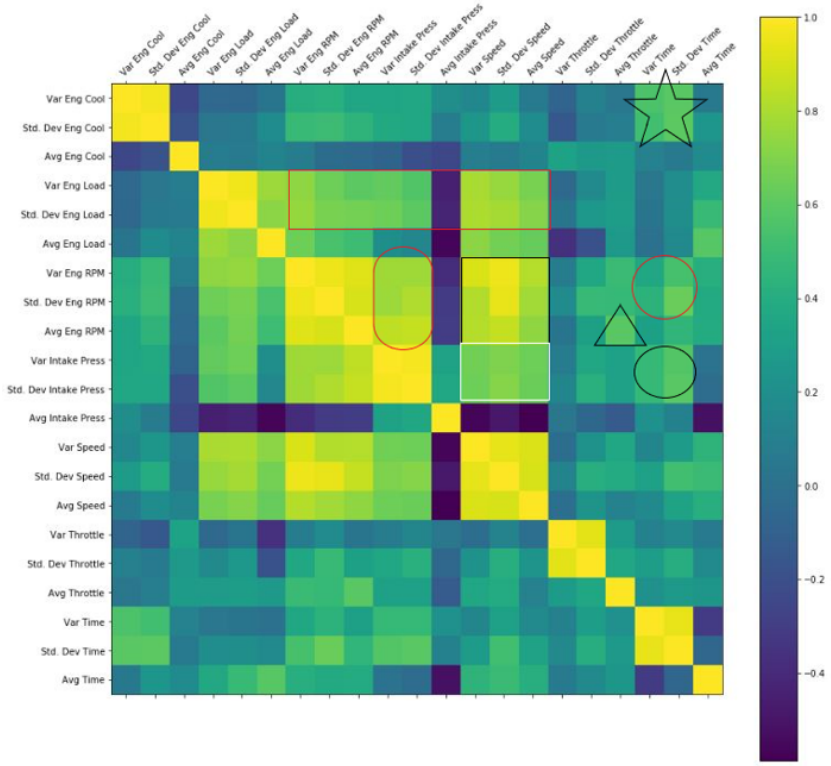


Figure 4.9. Pearson Coefficients on Cephas 14 dataset features' normalization

– The one between Engine RPM and Throttle (black triangle)

– The one between Intake Pressure Manifold and the Speed feature (black circle)

– The average Engine RPM's values with the average Throttle's values (white circle)

From these various analyses, we then draw up a recap in which are contained the strongest and the most commonly seen relationships. We start underlining the high bounds' strength of the *Speed feature, Throttle feature and Engine RPM* (also called Engine Speed), that are present in all the different datasets.

Another important connection can be found the acceleration features like the Accelerator Pedal, Longitudinal Acceleration, Latitudinal Acceleration with the Throttle and Engine Load features, this relationship is very strong in the Seoul dataset, in which the acceleration information is very frequent.

Finally, we have two other strong bounds, the first in the Cephas set of trips, in which the Intake Air Pressure (IAP) is closely related with the triple Speed-Throttle-Engine RPM, and the second in the SparkWorks dataset, where, the Gear factor is tight with the Throttle and Speed features.

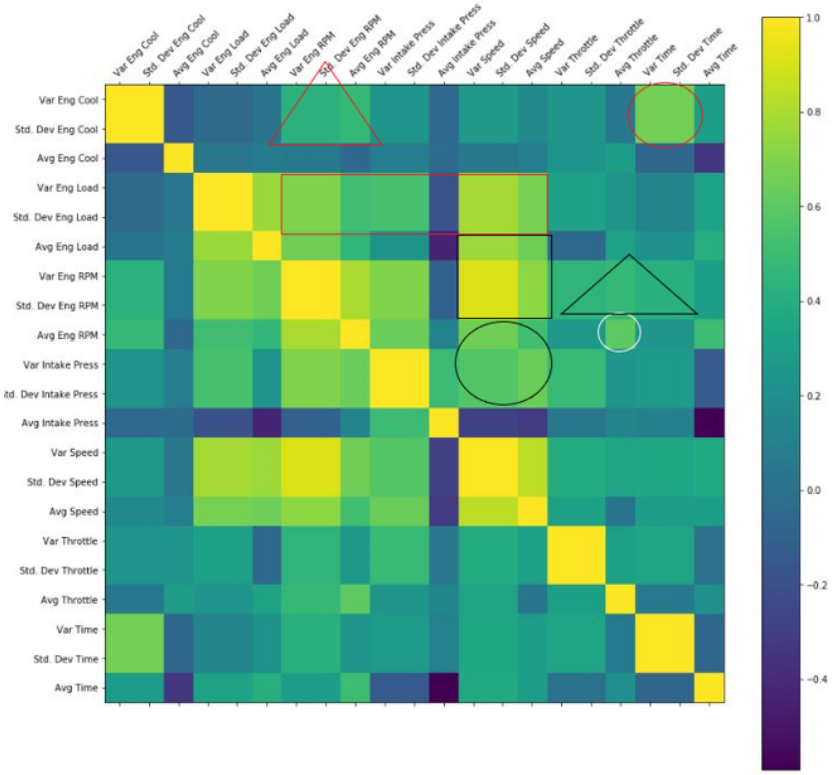


Figure 4.10. Spearman Coefficients on Cephas 14 dataset features' normalization

4.3 Coefficients' correlation in features entirety

However, we must understand if there is a good alternative from a correlation based on the average values or, in general, starting from a unique value for each path's feature.

For this reason, the new intuition is to create a triangular matrix in which, for each trip, we relate each feature's array with the others, obtaining a quantifier of the correlation among these values. Starting from these multiple triangular matrixes, each one related to a specific trip, we will sum up them in a final structure, dividing the obtained values by the number of considered paths.

The only bond that we have to take into account is that each couple of selected arrays must have the same length and no empty values, for this reason, is it not possible to apply this method on the SparkWorks dataset in which we have features that are sampled with different frequencies generating arrays of different length. The only way to solve this problem is to apply normalization on this dataset's vectors, but even in this way, the only thing that we obtain for all the three methodologies is a bad-structured matrix (4.11).

This probably happens, because the data in the source, even after cleaning and normalization, are structured in a non-optimal way so we can obtain only some main hints, that are the same strong correlations that we will find beneath in the other dataset correlations.

We have, then, applied this approach to the other two datasets, reminding that

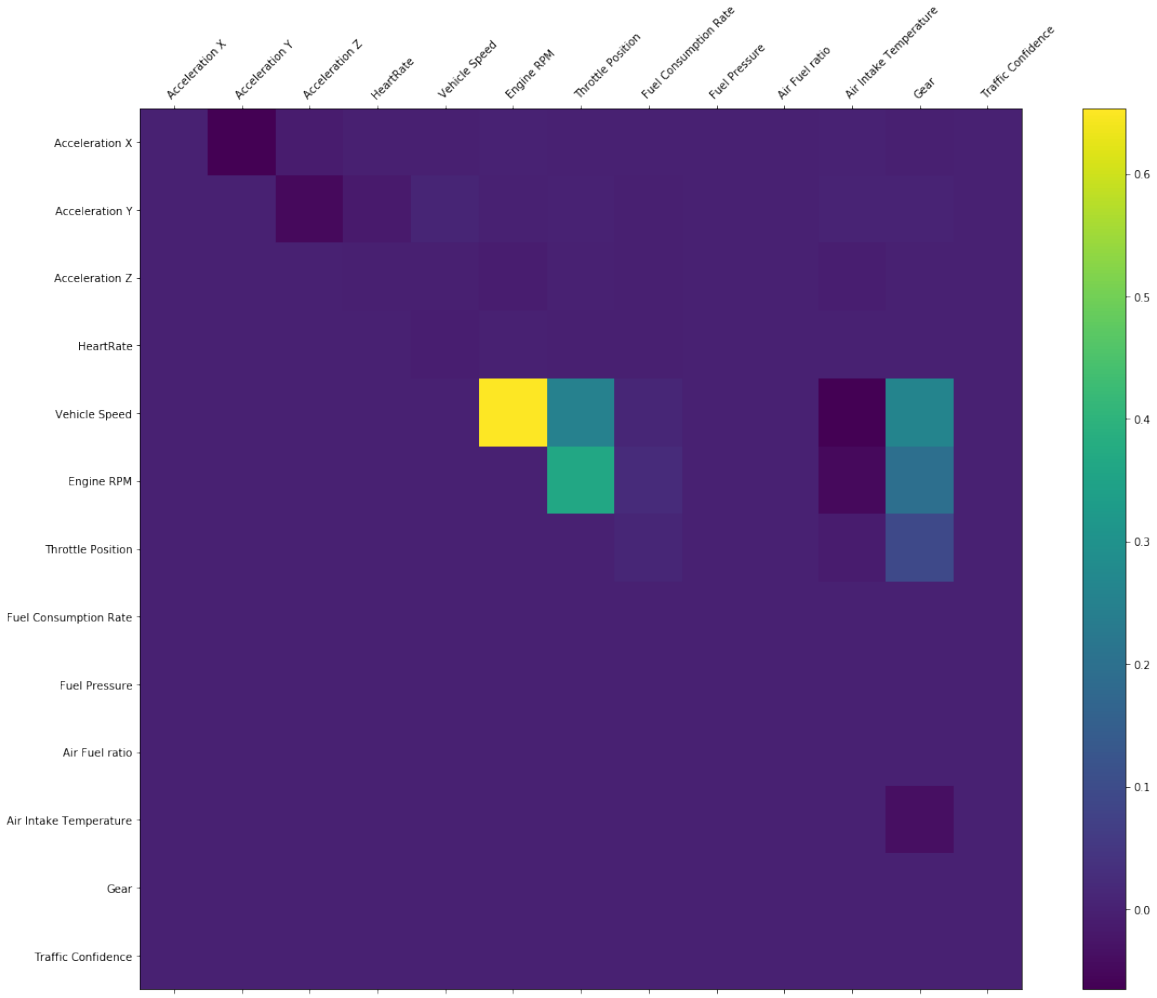


Figure 4.11. SparkWorks dataset's features normalization attempt

Figure 4.12. Pearson correlation on Seoul complete features

the Cephas one is characterized by many features having some null values that we have replaced with 0 values. This, to see if they develop some new correlation coefficient result that is similar to the ones obtained from the correlations applied to the values' averages.

In the ***Seoul dataset***, we obtain two different results, one related to all the fields and another one related to a selection of them.

Let's start with the analysis done on all the fields, in particular with Pearson's approach (4.12).

Here the most interesting relationships that can be seen are

– Absolute Throttle Position feature with the Engine Speed and Engine Torque features (black oval)

– Engine Speed feature with Wheels velocity (all of them) and Vehicle Speed features (black rectangle and red triangle)

– Maximum indicated Engine Torque element with Torque Converter Speed and

Figure 4.13. Kendall correlation on Seoul complete features

Wheels velocity features (red circle and red rhombus)

– Maximum indicated Engine Torque with Vehicle Speed (black triangle)

– Vehicle Speed with Wheels Velocity, Current Gear and Torque Converter Speed (white oval)

But also in Kendall's scenario (4.13) we obtain a similar result formed by

– Relationship among Absolute Throttle Position, Engine Speed and Engine Torque features (white rectangle)

– Relationship among Accelerator Pedal Value, Engine Speed and Engine Torque (black oval)

– Correlation between Intake Air Pressure and Acceleration Speed Longitudinal (white square)

Figure 4.14. Spearman correlation on Seoul complete features

– Correlation among Engine Speed, Wheels velocity (all of them) and Vehicle Speed features(black rectangle and red triangle)

– The one between Maximum indicated Engine Torque feature with Vehicle Speed feature(black triangle)

– The one among Vehicle Speed, Wheels Velocity, Current Gear and Torque Converter Speed features(red oval)

Finally, for what concerns the Spearman's correlation (4.14), we find a mix of the previous relationships found with the above approaches, with some adding elements like

– Absolute Throttle Position with the Engine Speed and Engine Torque (white rectangle)

– Throttle Position Signal with the Engine Speed and Engine Torque (black oval)

- Intake Air Pressure and Throttle Position Signal with Acceleration Speed Longitudinal (white oval)
- Engine Speed with Wheels velocity (all of them) and Vehicle Speed (black rectangle and red triangle)
- Maximum indicated Engine Torque with Vehicle Speed (black triangle)
- Maximum indicated Engine Torque with Torque Converter Speed, Wheels Velocity and Torque Converter Turbine Speed (red circle and red rhombus)
- Vehicle Speed with Wheels Velocity, Current Gear, Torque Converter Speed and the Torque Converter Turbine Speed Unfiltered (not underlined, but in the same position of the previous matrixes)

The next step is to reduce the number of features on what we are focusing on, proposing the usual three classification methods on these smaller groups. As usual, we start with the Pearson approach, looking at the main relationships among them (4.15)

- Relationship among the Fuel Consumption, Throttle Position Signal, Intake Air Pressure, Absolute Throttle Position, and Engine Speed features (black rhombus)
- Correlation between the Acceleration Speed Longitudinal feature and Throttle Position Signal and Intake Air Pressure features (red circle and black triangle)
- Relationship of Engine Speed with Wheels Velocity and Vehicle Speed features (red rectangle, white oval, and black oval)
- Correlation between Engine Torque feature and Acceleration Speed Longitudinal feature (white square)

Regarding the Kendall's one, we have a duplication of the Pearson's results, with the only difference that here we have a weaker relationship between the Fuel Consumption feature and the others elements, and between the Acceleration Speed Longitudinal feature and the Intake Air Pressure feature. (4.16)

While for Spearman's matrix representation we have the same relationships discovered in Pearson's approach.

Now is the moment to analyze the ***Cephas dataset***. In this section, the complete features' number and the selected ones' number is, more or less, the same, this because a lot of the features present irrelevant or not useful values. As usual, we divide the visualization into two blocks, the one related to 19 drivers on the same path, and the one related to 14 drivers on different paths. The most noticeable thing is that the interesting relationships in these blocks are the same for each of Pearson, Kendall and Spearman's approach.

In particular, the analysis of the trips related to the ***19 drivers*** (4.17) on the same path proposes these attractive correlations.

- Among MAF (4.3) feature and Engine Load and Engine RPM features (red oval)

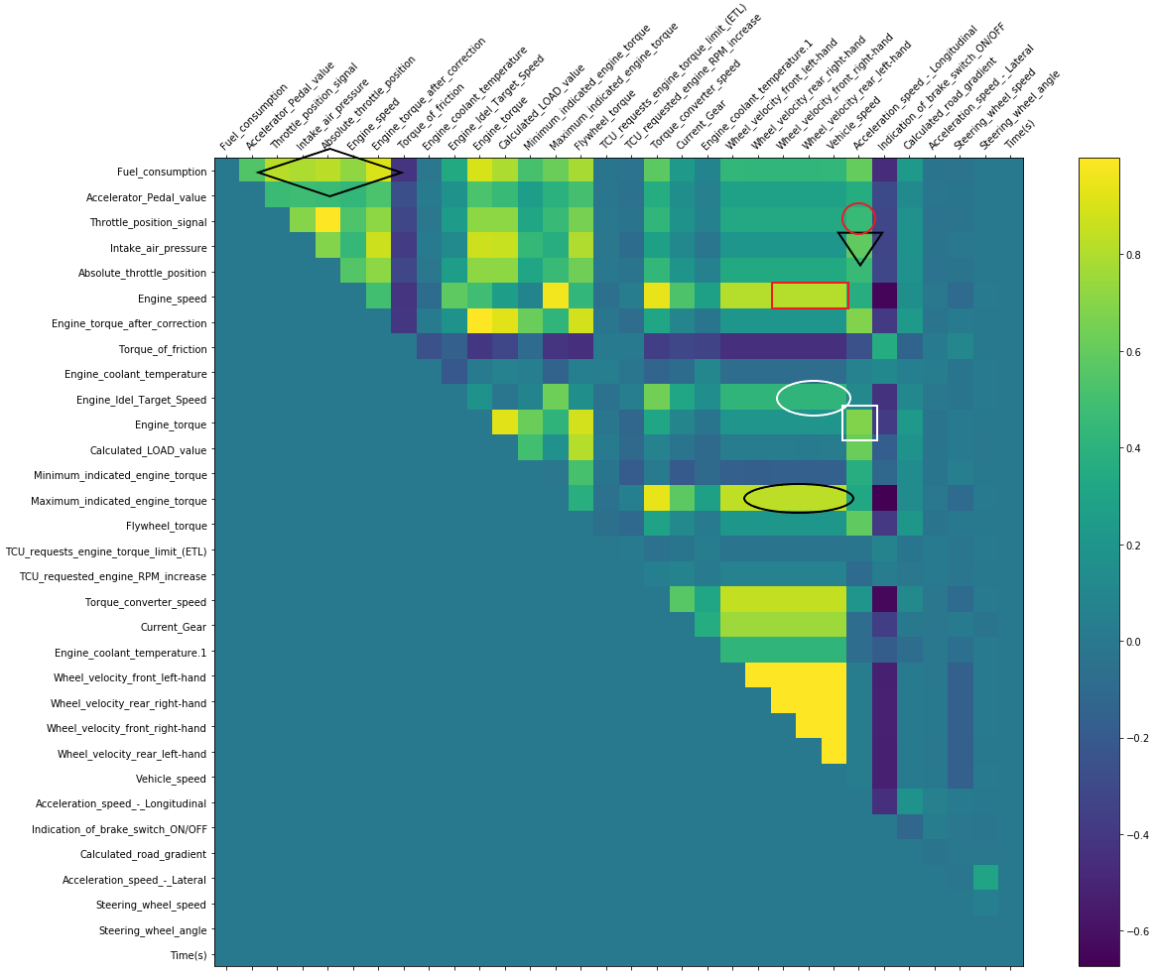


Figure 4.15. Pearson correlation on Seoul selected features

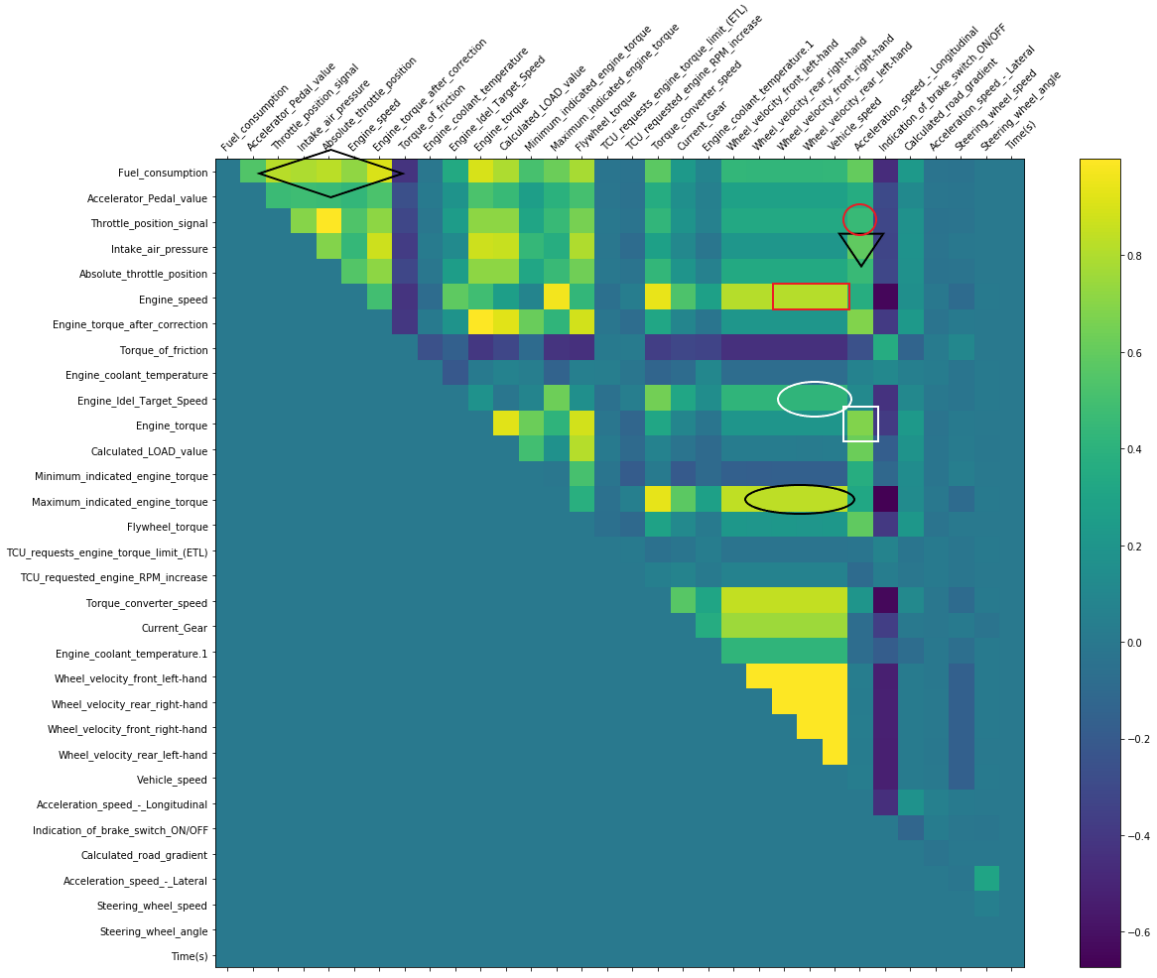


Figure 4.16. Kendall correlation on Seoul selected features

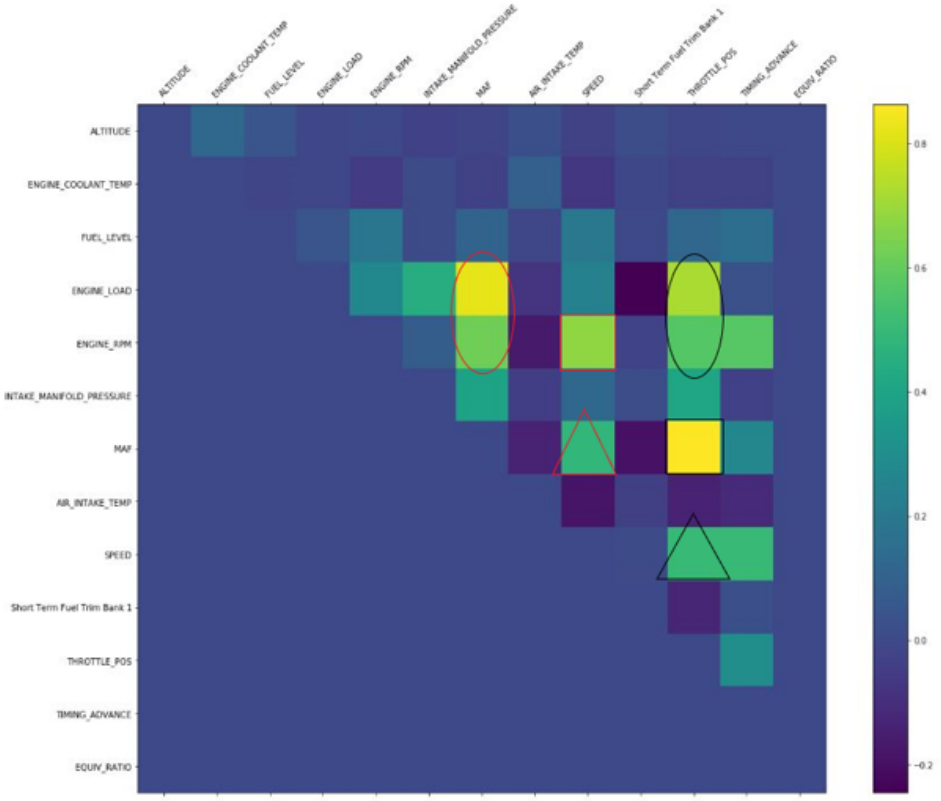


Figure 4.17. Pearson, Kendall and Spearman correlation on Cephas 19 features

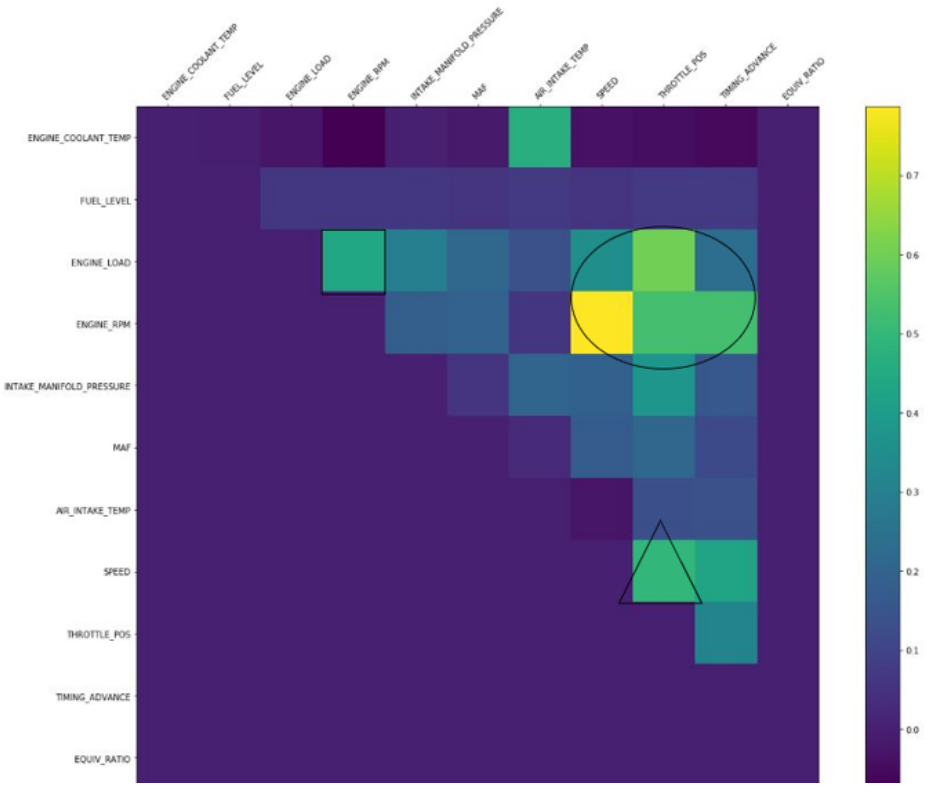


Figure 4.18. Pearson, Kendall and Spearman correlation on Cephas 14 features

– Among MAF feature and Speed and Throttle Position features (red triangle and black square)

– Among Speed element and Throttle Position and Engine RPM features(red square and black triangle)

– Among Throttle Position feature and Engine Load and Engine RPM features(black oval)

Then, we have the block of the **14 drivers' trips**, performed on different roads, where there are some noticeable interconnections (4.18) like:

– Engine Load with Engine RPM (black square)

– Speed with Throttle Position (black triangle)

– Engine RPM with Speed Throttle Position and Timing Advance (black oval)

– Engine Load with Throttle Position (black oval)

At this point, we have to compare the two methods, putting on one side the one developing correlations from the values' averages, standard deviations and variances calculations, and on the other side the one developing correlations starting from the features' vector relationships, to understand if they present the same results.

The answer can be obtained by looking at the Seoul and Cephas dataset, in which we have managed to apply both these methods. The analysis of their results tell us that usually, these relations are not coherent between the two approaches, however, looking carefully, we can find some relationships that are in common with the two methods.

In particular, they are:

- Speed and Throttle
- RPM and Speed
- Engine Load and Throttle
- Engine Load and RPM
- RPM and Throttle

Anyway, some of the correlations presented in only one of these two approaches could be used as a secondary element to figure out the driver categorization and the individuation of reckless driving styles. These secondary elements can be resumed with:

- Steering Wheel Speed
- Steering Wheel Angles
- Air Intake Pressure
- MAF
- Acceleration Speed Longitudinal
- Acceleration Speed Latitudinal

So concluding, we can affirm that now on we will study all the different trips paying special attention to the variation in time of the **"Primary Features"**:

Engine RPM, Speed, Throttle Position, Engine Load

and the **"Secondary Features"**: *Steering Wheel Speed, Steering Wheel Angles, Air Intake Pressure, MAF, Acceleration Speed Longitudinal, Acceleration Speed Latitudinal*

N.B. For MAF we indicate the Mass Air Flow sensor, which is one of the key components of an electronic fuel injection system in your car. It is installed between the air filter and the intake manifold of the engine. The mass airflow sensor measures the amount of air entering the engine or the airflow.

Chapter 5

Features and driving characterization

In this chapter some of the most important elements of drivers' behavioral characterization will be introduced, paying attention to their application on datasets and their role within the general classification algorithm addressed in this study. Among the most prominent factors, we certainly find the scores and thresholds, both selected by different scientific studies to be used in this new approach.

The thresholds will be used as initial indicators, to divide the values of the features into different groups, developing the first categorization between aggressive or unsuitable behaviors and calm attitudes. This initial classification will then be enhanced through the use of scores. The values of these scores will cover two roles, the first related to the creation of further thresholds that will form the superior limit and inferior limit set, employed for further classification, and the second, related to the verification of the effective relationship. of the values classified beyond the features' thresholds. This, to effectively quantify their power and to identify which of the secondary features are most related to the scores performances in the various paths.

Near to this, the clustering of the values obtained through the secondary thresholds will also be used as additional classifiers to enhance the initial subdivision carried out through the primary features and the scores.

Finally, the concept of peaks analysis will be introduced as a verification of the classification method, and therefore, as a mechanism for cleaning up any categorization errors, such as false positives or false negatives.

5.1 Driving scores

Here we want to introduce the set of elements, cited before, that will be useful for the general classification of the drivers' profiles. These elements are the **scores** and they are presented, as for the thresholds, in many works related to this subject. However, from their large group, we have selected only some of them basing on the fact that they were related to the same set of elements that we find in our datasets.

These scores are

- The Aggressiveness Score
- The Gear ShiftUp Score
- The Eco-Score

5.1.1 The Aggressiveness Score

The Aggressiveness Score is the simplest one [10], [11]. It is a score related to the variance of the Engine RPM (or Engine Speed) values present in every single trip. It gives a quantification of the high work of the engine that is considered to be unsafe and expensive for the drivers.

$$Aggressiveness_{RPMScore} = \frac{\sum_{i=1}^{n}(RPM_i - \mu)^2}{n}$$

Where n is the number of RPM samplings done in the specific trip on which the score is calculated.

5.1.2 The Gear ShiftUp Score

The Gear ShiftUp is introduced in the paper [9]. Here is put beside the Acceleration Factor the gears that are used, to understand how the behavior of the acceleration changes during the various shifts. However, we must take into account precise information:

> *The velocity and the Acceleration just before the gear changing moment are an indication of driving style. Investigation showed that the values two seconds before the gear change are most representative of the driving style. The acceleration at the gear changing moment itself is usually already lower because the driver releases the accelerator pedal to change gear.*

This means, that this approach is applied to a temporal window that dwells between 6 seconds and 2 seconds before the shift event. The focusing on these acceleration values is done because it is strongly probable that each driving style has a specific pattern on the accelerator pedal while a shift is performed, that could be used as a helping factor of driver categorization.

In particular, this score takes into account all the acceleration values sampled form 2 to 6 seconds before a shift event, calculating the average of them. From these averages is then calculated the general average of these values and then compared with the standard threshold of acceleration, to understand if the various shifting is done with an aggressive approach or a calm one.

5.1.3 The Eco-Score

For what concerns the Eco-Score, we are introducing a quite complex concept that is well defined in the document [10]. The Eco-Score is a factor obtained from many

calculations applied on the trips. It is formed by RPM Eco-Score, Cruising Eco-Score and Shifting Eco-Score, each of these factors analyze different elements of the driving performance, so let's see them in detail after the presentation of these elements needed for understand the different algorithms.

- $sftup[i]$ is a shift-up event with index i
- $sftup[i]_{rpm}$ is the peak RPM during the event
- $sftup[i]_t$ is the time required to complete a shift-up event
- u is the vehicle speed
- $\overline{u}$ is the mean vehicle speed
- $\overline{th}$ is the mean throttle position
- s_u^2 is the speed variance
- s_{th}^2 is the throttle variance
- w is the temporal window

The **ShiftUp Eco-Score** parameter describes the evaluation of shift-up events as a factor to calculate Eco-Score. The general idea is to punish shift up events with RPM values higher than 2500, and reward shift up events with RPM values between 2000-2500. The reward is less if the driver shifted up after several seconds above 2000 rpm. The punishment is more when the duration is several seconds after 2000 rpm.

The algorithm is described in the following steps

1. Assume a 30 seconds temporal window
2. For each shift-up event $sftup[i]$ calculate the peak RPM, denoted as $sftup[i]_{rpm}$ during the event and the duration $sftup[i]_t$
3. Create the histogram $\mathbf{H}$, where $\mathbf{H}[n]$ is the set of indices of shift-up event assigned to bin n
4. Assign each shift-up event index to a histogram bin $\mathbf{H}[n]$ using the following expressions:

$$\left[\frac{(sftup_{rpm} - 2000)}{100.0}\right], sftup_{rpm} \in [2000, 3100]$$

5. Calculate parameter penalty for each bin $f_p(n)$.

$$f_p(n) = \left(6 - \frac{12}{(1 + e^{4-n})}\right)$$

6. Given histogram **H** of length N the penalty or bonus parameter, expressed by vector **P**, the value that is added to the histogram bins is calculated from the average duration of all the shift up events of the same bin with the following expression. The first bin corresponds to shift up events below 2000 rpm and is multiplied with constant factor of 5:

$$P(n) = \begin{cases} 5 \cdot f_p(n) \cdot \frac{\#H(n)}{\sum_{i=0}^{N} \#H(n)} \cdot (-0.0002024) \cdot \frac{1}{\#H(n)} \sum_{i \in H(n)} sftup(i)_t + 1.045 & \text{if } n = 0 \\ f_p(n) \cdot \frac{\#H(n)}{\sum_{i=0}^{N} \#H(n)} \cdot (-0.0002024) \cdot \frac{1}{\#H(n)} \sum_{i \in H(n)} sftup(i)_t + 1.045 & \text{if } 0 < n \leq 4 \\ f_p(n) \cdot \frac{\#H(n)}{\sum_{i=0}^{N} \#H(n)} \cdot (-0.0002024) \cdot \frac{1}{\#H(n)} \sum_{i \in H(n)} sftup(i)_t + 0.4 & \text{if } n > 4. \end{cases}$$

7. Finally, the ecoscore is calculated by summing up the values of vector **P**

$$ECOSCORE_{SHIFTUP} = \sum_{n=1}^{N} P(n)$$

Different is the approach for the **RPM Eco-Score**. The RPM ecoscore parameter describes the evaluation of acceleration events as a factor to calculate ecoscore. The calculation of acceleration ecoscore parameter employes the following notation:

- $RPM[i]$ is an RPM event with index i
- An $RPM[i]$ event occurs when $RPM > 2500$. The events are also validated from Throttle position values id available.
- $RPM[i]_PEAK$ maximum RPM value during RPM event
- $RPM[i]_t$ is the duration of an RPM event measured in seconds
- w is the duration of the temporal window.

The algorithm is described in the following steps:

1. Assume a 30 seconds time window, $w = 30$
2. Events of high RPMs are detected.
3. For each event $RPM[i]$ the duration of the event ad the maximum RPM value is utilized to construct the RPM ecoscore event factor parameter.

$$EventFactor = -\left(1 + \frac{RPM[i]_t}{w}\right) \cdot \left(5 - \frac{10}{1 + e^{\frac{RPM[i]_{PEAK} - 2500}{100}}}\right)$$

4. The sum of all event factors construcs the RPM Ecoscore Parameter

Then we have the **Cruising Eco-Score** parameter, that evaluates the ecoscore during the vehicle's cruising. The algorithm is described as following:

1. Assume 30 seconds temporal window
2. Calculate speed variance s_u^2 above 40 Km/h
3. Calculate throttle position variance s_{th}^2 above 40 Km/h
4. Employ overlapping variable size windows within the 30 seconds temporal window is created with according to speed
5. Calculate speed factor f_u

$$f_u = \frac{\overline{u}}{100} \cdot \left(0.1 - \frac{0.3}{1 + e^{3.5 - s_u^2}}\right)$$

6. Calculate throttle position factor of temporal window

$$f_t h = \frac{\overline{th}}{10} \cdot \left(0.1 - \frac{0.2}{1 + e^{4 \cdot (1 - s_{th}^2)}}\right)$$

7. Calculate ecoscore

$$ECOSCORE_{CRUISING} = \begin{cases} 1.5 - \frac{1.2}{1+e^{\frac{53.3647 - (f_u + f_{th})}{10}}}, & (f_u + f_{th}) > 0 \\ 1.5 - \frac{1.0}{1+e^{\frac{(f_u + f_{th}) - 50}{10}}}, & (f_u + f_{th}) < 0 \end{cases}$$

The **Final Eco-Score** is nothing else than the sum of these scores, higher it is, worst the driving approach is developed.

5.2 Features characterization

This section aims to detect some approaches that can be useful to create some thresholds and scores for the trips' values. This is done to start to cluster the various features' values in some sets and to give a score to each trip that can help us in classifying the drivers' attitudes as "Good driver behavior" or "Bad driver behavior", in a unified manner. For example, a clustering that can be useful for our purpose could be the creation of three sets, "Good", "Medium", and "Bad" where we're going to insert the different features' values of the drivers' paths.

5.2.1 Primary thresholds

First of all we select some approaches found in the documents [14],[5],[6], in order to create a clustering of the "main" features.

The most interesting ideas discovered in those papers are

1. Throttle thresholds
    - Calm Throttle: 0-39% inclination

- Moderate Throttle: 49-59% inclination
- Aggressive Throttle: 60-100% inclination

2. RPM thresholds

- Calm RPM: Rpm<2000
- Medium RPM: 2000<Rpm<2999
- Aggressive RPM: Rpm>3000

3. Acceleration thresholds (first)

- Moderate acceleration: Ax>1.5
- Aggressive acceleration: Ax>3

4. Acceleration thresholds (second)

- Good Acceleration: -3<Ax<4
- Bad Acceleration: Ax>4
- Bad Deceleration: Ax<-3

5. Vehicle Speed thresholds

- Medium Speed: Speed Val≤ Speed Limit
- Aggressive Speed: Speed Val>Speed Limit

These thresholds introduction is inserted near the scores selection previously introduced to obtain a general overview of the drivers' classification and to identify some relationship between them and the many features.

For this reason, let's now see, focusing on the datasets one by one, which threshold has been used on which features, and successively, which score has been applied through them.

For the **Seoul dataset** we don't have GPS locations but we can still apply the RPM Eco-Score, the Gear ShiftUp score, and the Aggressiveness Score. The missing information related to the GPS locations creates to us some problem in the usage of the Speed threshold defined above. If we don't know the exact position of the vehicle we cannot obtain the information related to the specific road's speed limit; however, to circumvent this problem, we have used the following substitutive threshold

$$\begin{aligned} \text{Good if} &\quad \text{Speed value}<50 \text{ km/h} \\ \text{Medium if} &\quad 50\text{km/h} < \text{Speed value} < 80 \text{ km/h} \\ \text{Bad if} &\quad \text{Speed value}>80 \text{ km/h} \end{aligned}$$

based on the fact that all the routes identified by the drivers were in the urban area of Seoul.

Regarding the other main features, we have applied all the thresholds defined in section 5.2.1 to obtain a clustering of the features' value, the result is showed in figure 5.1.

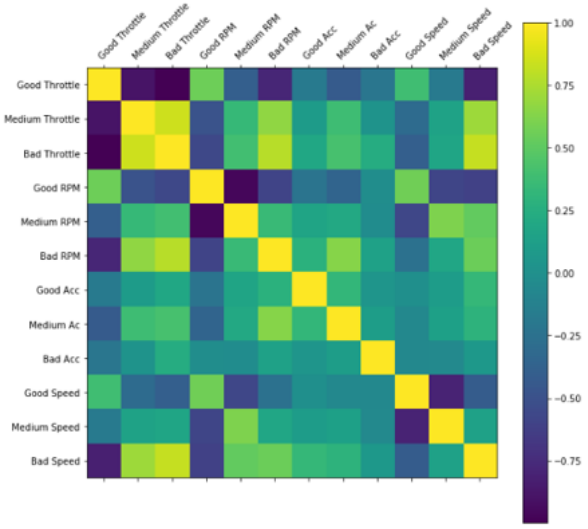


Figure 5.1. Good-Medium-Bad values' classification for Seoul dataset

The fact that, when the Bad Throttle values presence increase also the Bad Speed and Bad Rpm values presence increase, means that they are tightly bonded. Additionally, if we are in a particular situation like this one where we don't have GPS positions in which it is hard to determine the Speed limit, we can analyze the Throttle values to classify the behavior of a driver. Vice versa when we have high values of Speed, like in the highways, we can look at the Throttle correspondent value to understand if there is a reckless way of driving or not. Obviously when we have an urban section, where the speed is limited by traffic or semaphores, looking at the RPM peaks and the correspondent Throttle values allow us to understand if the various start and stop are done recklessly.

However, the matrix 5.1 underlines also some classification problems, for example, for urban traits the Acceleration thresholds taken into consideration are too high, because in presence of bad values in the other features, in the acceleration one we don't have any distinctive sign of recklessness.

A way to resolve the general problem of the Acceleration Threshold in the urban context could be to focus on the number of times in which the Vehicle Acceleration has the value greater than 0.981 (in particular Ax>0.981), this gives a hint in defining acceleration events, so instead of looking for the exact threshold of the Acceleration, we can count the acceleration events, to understand an average value of them that can be used for determining when there is gentile use of the pedal and when not.

In this way we obtain two interesting results, the nearness with the RPM values and Speed values is increased, and most of all, this nearness is higher for the good values section of those features rather than the bad ones.

Now, is the moment to apply also the scores that we have mentioned before, starting from the RPM Eco-Score (5.2).

The first element that we can notice is that this score is strictly connected to the "Tri-Relationship" bad values formed by Bad RPM values, Bad Speed values, and Bad Throttle values. It confirms our hypothesis that this score and these features will be very useful for driving behavior identification. There is also good correspondence

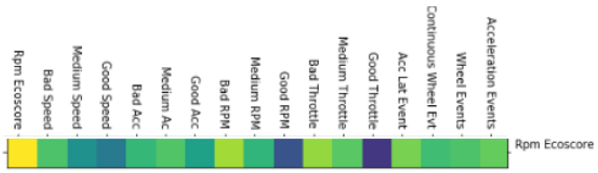


Figure 5.2. Correlation between RPM Eco-Score and other features

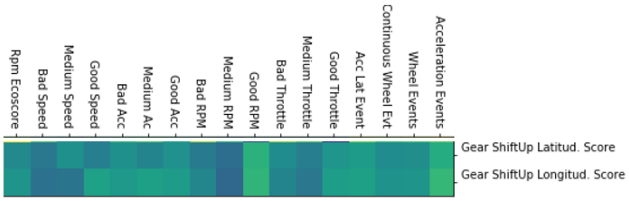


Figure 5.3. Correlation between Gear ShiftUp Score and other features

with the acceleration events and the events related to the steering wheel, which will be very important for the next steps. Let's switch to the Gear ShiftUp Score (5.3)

In this situation we have no strong correlations between the Gear ShiftUp Score, neither applied on Latitudinal Acceleration nor Longitudinal Acceleration, with the other features. The only feature that presents some similarities is, obviously, the one related to Acceleration Events, on which the score is based on.

Finally, we look to results coming from the Aggressiveness score's application (5.4)

Here, the aggressiveness score presents a high correlation with the majority of the bad sets of the Tri-Relationship. Another interesting interconnection is the one with the RPM Eco-Score, this means that these two scores could be used together in the successive steps, for the driver behavior classification.

Let's, then, pass to another dataset, the **Cephas** one with the first section of 14 drivers and the second of 19 drivers. In the first section, the one formed by **14 drivers** we don't have a maximum speed limit of roads, roads identifiers, gear values, GPS locations, and acceleration values, so here we can apply only the subdivision of the values in the three sets introduced before, the Good, Medium and Bad ones. First, for each trip we look for engine RPM, vehicle speed and throttle position values, then we distribute them in these sets basing our division on the thresholds cited above. After that, we relate these results to understand if what found in the previous dataset is true even here. The result confirms our hypothesis.

Here we have less strong connections between the Throttle position bad values

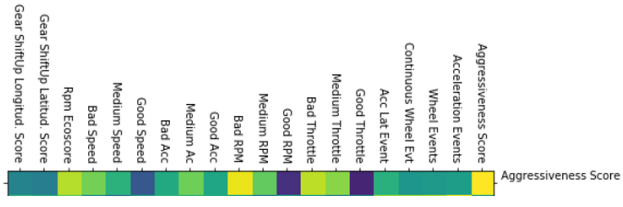


Figure 5.4. Correlation between Aggressiveness Score and other features

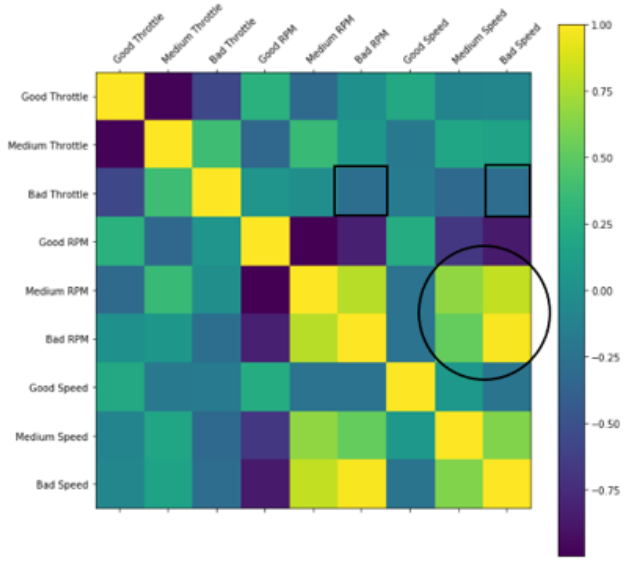


Figure 5.5. Good-Medium-Bad values' classification for Cephas14 dataset

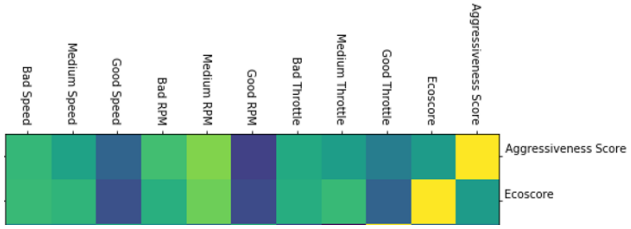


Figure 5.6. Correlation between Aggressiveness Score, Eco-Score and the other features

with the other features' bad values, but the interconnection of the Speed bad values and Engine RPM bad values is astonishing. This difference could be influenced by the type of road on which these trips were done.

Also in this scenario, we can apply some of the scores cited before, in particular the RPM Eco-Score and the Aggressiveness Score. Unluckily we managed to use only these two elements because of the lack of the Gears information that does not allow us to look to the ShiftUp events.

In the other section, the one based on **19 drivers**, we have some more information, like the GPS location made up by latitude and longitude coordinates. The problem is still that we don't have gear values so we can't apply the shiftUP event calculation or the Eco-Score event calculation. We still don't have acceleration values so some of the previous thresholds cannot be used to help us in this study. Another problem is that we don't have the Maximum Speed Limit for each section of the road that is traversed during a trip. However, having the GPS location, means that it can be used to identify the Maximum Speed Limit of the road that we are traversing through the use of two different services: the Nominatim OpenStreetMap and the Overpass API.

The first step that we have to apply is the format of the different GPS locations (for example in this dataset's trips there was corrupted information so we have reformatted them). After reformatting, the next step is the reverse geocoding, in few words for each GPS location we send to the Nominatim service an Http request in which we ask for the nearest road to this position if this information is in the database the service respond us giving the Bounding Box, a set of 4 values indicating North, South, East, and West, that are the exact coordinates of the road section that the driver is traversing at that moment. After this, we pass, through another Http request, this Bounding Box to the Overpass API, that will ask the OpenStreet Browser service the Maximum Speed Limit of the road for that specific box. These are two free services, so it is obvious that there were many situations in which the bounding boxes were not related to any speed limit. So, the solution of the Max Speed Limit problem is solved in this way: if there are not the speed limits in the services for a specific path, its speed values will be categorized as before, basing on the 50 km/h and 80 km/h threshold values, otherways they will be categorized as Good Speed if value Speed is lower than the Max Speed Limit for that section, Medium Speed if the value is very near to the maximum, and Bad Speed if the value is greater than the Max Speed Limit.

Starting from the thresholds used for the section of 14 trips, we can add this new knowledge to understand more appropriately if the speed value is good or not instead basing only on the two rough values as thresholds. The result coming out from the application of these elements are the following.

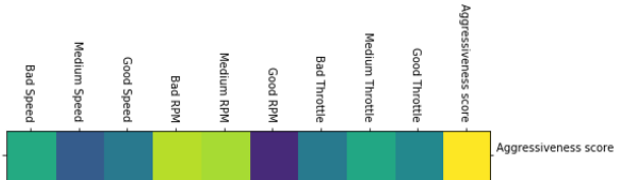


Figure 5.8. Correlation between Aggressiveness Score and the other features

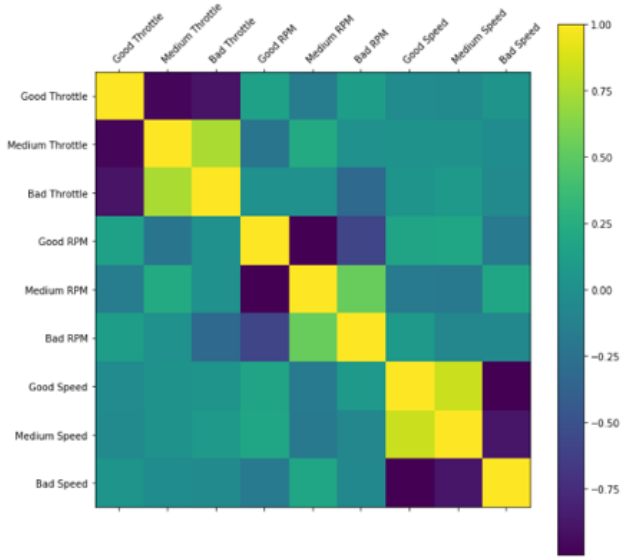


Figure 5.7. Good-Medium-Bad values' classification for Cephas19 dataset

Here we have some less confirmation, in the Tri-relationship of RPM-Speed-Throttle the bad values are still bounded each other but not strongly as before, instead, the medium values are still well connected. This means that, probably, the tri-relationship is a good classification tool, but in this case, we could have a set of calm drivers in which no bad behaviors are applied.

For what regards the appliable scores, we managed to apply the Aggressiveness Score, but we can't take in consideration neither the scores related to the ShiftUp due to the lack of Gear information nor the Eco-Score elements, due to a lack of timestamp information. The results that the Aggressiveness Score provides are represented in figure 5.8.

The situation described is quite interesting, the high correlation of the score with the bad values, especially the one related to the RPMs, confirms our hypothesis on the power of these factors related to an eventual classification of driving styles. The only thing that we have to notice is that there is also a good relationship between the aggressiveness score and the medium set of RPM, so we will need to be aware, again, to some false positive.

Finally, we switch to the **SparkWorks dataset**, in this dataset we have all the information needed to apply a Gear ShiftUp, and Eco-Score analysis, then we have also different acceleration features that could be useful to understand if the applied thresholds are effectively good or if they are too high to determine in a well-manner

the driving behaviors. We will also apply the comparison of the speed values with the Maximum Speed Limit for each road's segment that the user traverse (individuated through the above-mentioned approach), allowing us to use a better speed threshold ad hoc for each road's section.

Let's start employing the set of thresholds, as for the other scenarios. It shows the following interconnection among values:

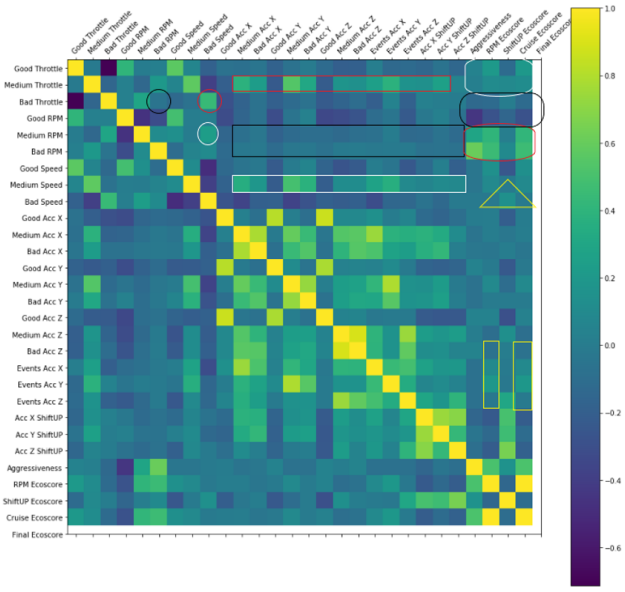


Figure 5.9. Good-Medium-Bad values' classification for SparkWorks dataset

First of all, as we suppose, the Gear ShiftUp is a strong factor for the classification of the driving style. It has, in fact, its most strong relationships with the bad section of the Tri-Relationship. Even the Eco-Score reveals to us that, as supposed, it can be very useful to classify the driving behavior, in particular, we can observe that it is related to all the medium-bad sections of the Tri-Relationship, especially for the RPM values. Other interesting elements can be found between the various acceleration categorization sets and the medium Throttle-Speed sections, and an interesting relationship between the Acceleration events and the Eco-Scores.

This means that even if the acceleration features are not considered main fields, they can be beneficial in confirming the behaviors suggested by the Tri-relationship elements.

So, let's recap what we have discovered until now. We have understood that, through the application of different studies and methods over the main features, they are truly connected. This is confirmed by the subdivision of their values in three categories, good-medium-bad. Near to this, firstly we have seen that the different combination of the Tri-Relationship features could be a good tool to classify the drivers' behaviors, and to understand the type of road that is traversed; secondly, we can affirm that both the Aggressiveness score, the Gear ShiftUp indicator, and the Eco-Score are strictly connected to the bad values of this Tri-relationship, so they are strong indicators of bad behaviors if used, near the above-mentioned elements, as an additional classifier tool.

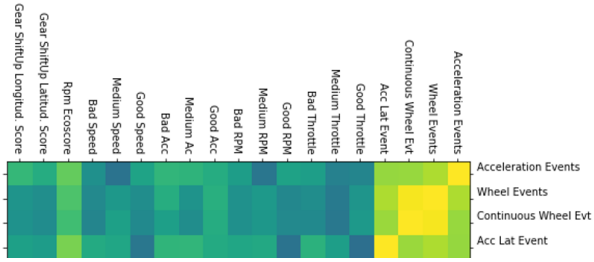


Figure 5.11. Correlation among Seoul's secondary features and general features

5.2.2 Secondary thresholds

The next step is to try to classify the secondary features, and to see if they can be useful as additional indicators for a driving style classification and identification method.

In the **Seoul dataset**, we managed to apply some studies on secondary features like steering wheel's rapid turning event at not so low speeds, the continuous rapid turning event (in particular 4 or more consecutive rapid turning events) and the lateral acceleration thresholds analysis. Let's look at the criteria for the analysis of the steering wheel events. Here the study is applied only on a set of filtered values, for example, the values where the driver only turns the steering wheel but has not yet engaged the throttle to complete the turning event is not considered as a turning event.

Turning events are defined as

$$\begin{aligned} \text{StartTurnEvent: } & |\text{SteeringAngle}| > \varepsilon_t \text{ AND VehicleSpeed} > 0 \\ \text{EndTurnEvent: } & |\text{SteeringAngle}| < \varepsilon_t \quad \text{where } \varepsilon_t \text{ is 30 degrees.} \end{aligned}$$

Abnormal driving will be detected if the driver turns left or right, for a single time or for four times continuously, with speed higher than 30 km/h. [24],[25]

For what concerns the lateral acceleration, instead, we have to notice that if the driver does not have the habit of slowing down in time when turning, the vehicle body can very often be unstable. This can be discerned from the data of Acceleration Meter X-Axis (1.58%/0.093%). In particular, we can individuate some bad turning events event if the Lateral Acceleration (acceleration on X-axis) is lower than -0.5 or greater than 0.5. [20]

Accelerometer -X-Axis(the left and right of the vehicle body)	$-0.5 \le a_x \le 0.5$ Unit: g

Figure 5.10. Lateral Acceleration threshold

Once that we have seen which are the analysis' criteria, the next step is to apply them, the result of their application can be seen in 5.11.

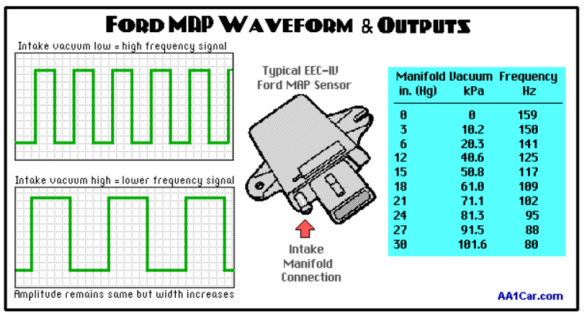
These new features seem to be useful for our study. They are strongly connected with the RPM Eco-Score, and quite related to some of the bad values of the Tri-Relationship. Looking closely we can see that the acceleration events are connected

to the bad acceleration values, but not so related to the bad throttle values and the bad RPM values.

The same interconnection among values can be seen if we look at the Acceleration Lateral event. For what regards the steering wheel events, except for the acceleration elements and the RPM Eco-score, they seem to be less related to all the bad values sections in which the bound are not so tight. Finally, against what we could suppose, the steering wheel elements are not so related to the Gear ShiftUp scores, while instead, the acceleration features are a little bit connected to them.

After that, we have focused ourselves on the Intake pressure values for a vehicle. But what is the Intake Air Pressure?

It regards the air that is sucked by the engine under some specific load situation, in particular, when the engine is working hard intake vacuum drops as the throttle opens wide and the engine sucks in more air. When the engine is not running, the pressure inside the intake manifold is the same as the outside barometric pressure. When the engine starts, a vacuum is created inside the manifold by the pumping action of the pistons and the restriction created by the throttle plates. At full open throttle with the engine running, intake vacuum drops to almost zero and pressure inside the intake manifold once again nearly equals the outside barometric pressure. The vacuum inside an engine's intake manifold, by comparison, can range from zero up to 22 inches Hg or more depending on operating conditions. Vacuum at idle is always high and typically ranges from 16 to 20 inches Hg in most vehicles. The highest level of vacuum occurs when decelerating with the throttle closed. The pistons are trying to suck in air, but the closed throttle chokes off the air supply creating a high vacuum inside the intake manifold (typically four to five inches Hg higher than at idle). When the throttle is suddenly opened, as when accelerating hard, the engine sucks in a big gulp of air and vacuum plummets to zero. The vacuum then slowly climbs back up as the throttle closes.



Manifold Vacuum in. (Hg)	kPa	Frequency Hz
0	0	159
3	10.2	150
6	20.3	141
12	40.6	125
15	50.8	117
18	61.0	109
21	71.1	102
24	81.3	95
27	91.5	88
30	101.6	80

So, for a very small value of the pressure, we have a hard acceleration, while the high-pressure values are obtained during deceleration with the Throttle closed; this implies a harder deceleration for a harder braking event. The average pressure threshold can be found around 22 Hg, which is equivalent to a threshold of 75 kPa. However, as we can see in 5.12, except for the connection between RPM Eco-Score and the Intake air pressure standard, all the other Tri-relationship clusters with which the intake pressure is bounded, are the good section ones. This means that this feature of the intake pressure is not so useful to develop bad behavior profiling, so we will not place it side by side with the main features.

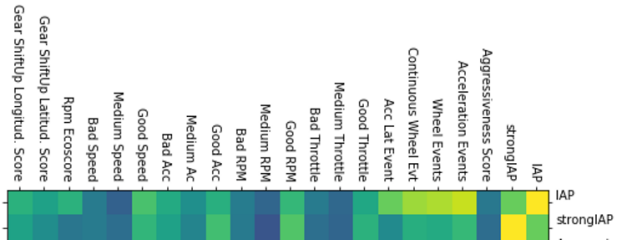


Figure 5.12. Correlation among Seoul's IAP features and general features

Now is the moment of the **Cephas dataset**, here we have no information related to acceleration features or steering wheel features, so, for this reason, we focus on the Intake Air pressure element, that in the section of 19 drivers is accompanied by the MAF feature. This new feature can be described and individuated as follow:

> *The MAF's PID value should read anywhere from 2 to 7 grams/second (g/s) at idle and rise to between 15 to 25 g/s at 2500 rpm, depending on engine size.*

Let's look at the results of these correlations.

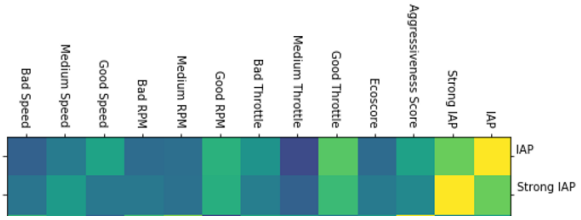


Figure 5.13. Correlation among Cephas 14 IAP features and general features

In the **Cephas 14 dataset** the bad speed element has a stronger connection with the IAP rather than before.

Nevertheless, the Intake Air Pressure cannot be used our study, firstly because its connection with the bad sections is, in general, not tight, and secondly because its connection with the other features changes a lot from dataset to dataset, making impossible to understand its behavior during the different guide style.

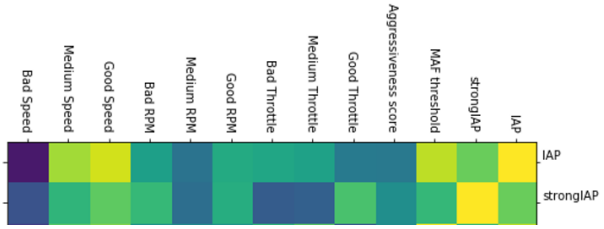


Figure 5.14. Correlation among Cephas19 IAP features and general features

In this case, the **Cephas 19 dataset**, instead, the situation is quite similar to the Seoul dataset. In fact, except for the MAF features and the bad Throttle

values relationship, the correlation with the other bad value sections are not so tight, especially the bound between the Intake Air Pressure and the Bad Speed that should be the strongest one.

From the notion 5.2.2 we have matched the 2500 RPM with the 25 g/s of the MAF as a limit threshold, then for the other values, we have applied the ratio basing on this limit developing a correlation with the other features.

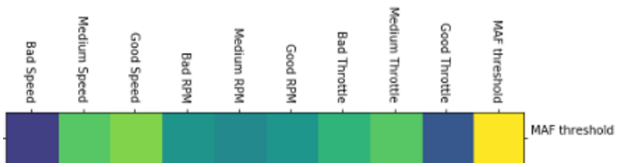


Figure 5.15. Correlation among Cephas19 MAF features and general features

However, as for the Intake Air Pressure, even the MAF is not a good secondary feature option for our analysis, for the same reasons said before.

This means that from the various attempts done to find features that could be useful if placed near the main feature, only some of them confirm us to be helpful for our aim, in particular, they are:

- Steering Wheel Events
- Acceleration Events
- Acceleration Lateral Events

The secondary features analysis was not applied to the first dataset, because of the lack of all the above mentioned specific information.

5.3 Driving characterization

In this section we will explain in detail the innovative mechanism used for the classification of drivers based on the concepts introduced so far, such as scores and thresholds, to then introduce the component of the peaks that will be addressed in the next section.

To do this, however, we must clarify how the information we have obtained through the various previous steps is structured. We aim to analyze how the main feature's values are distributed in a trip that could be considered as a good drive trip or bad drive trip, looking, as distinction value, the scores like the RPM Eco-Score, Aggressiveness score, and Bad Speed score (the first two scores are strictly related to the RPM of the car's engine, while the third is the percentage of speed values that are over the speed limit during the trip). Our initial step is to create a table for each set of trips where we insert, for each path, all the scores that we manage to apply, and all the value distribution percentages that we have introduced before. Here it is an example to understand better what we are talking about.

Filename	IAP	strongIAP	Aggressiveness Score	Acceleration Events	Wheel Events	Continuous Wheel Evt	Acc Lat Event
Driver_A_1_File_1.csv	176	0	0.7082605045021667	-28.875	34	4	11

Good Throttle	Medium Throttle	Bad Throttle	Good RPM	Medium RPM	Bad RPM	Good Acc	Medium Ac	Bad Acc	Good Speed
85.51	2.17	12.32	72.22	22.71	5.07	38.71	1.21	0.0	37.8

Medium Speed	Bad Speed	Rpm Ecoscore
31.16	6.94	25

When possible, or in other words when we have enough information, we insert in this table also other values like

Good Acc X	Medium Acc X	Bad Acc X	Good Acc Y	Medium Acc Y	Bad Acc Y	Good Acc Z	Medium Acc Z	Bad Acc Z	Events Acc X	Events Acc Y	Events Acc Z
10.13	0.32	0.04	13.82	0.94	0.09	16.92	2.22	0.09	##########	##########	##########

Acc X ShiftUP	Acc Y ShiftUP	Acc Z ShiftUP	Aggressive	RPM Ecoscore	ShiftUP Ecoscore	Cruise Ecoscore	Final Ecoscore
0.289879371	0.442529104(	0.446405704	0.2532282	35	6	0.507	42

Once the elements underlying the algorithm have been presented, it is necessary to show the actual method used in this study, introducing how the driver's behavioral classification is put into practice. Let's start by saying that a driver can be identified as an aggressive driver, a calm driver, or an intermediate driver. The differentiation is carried out initially by looking at the percentage of values of some features that have been considered negative. This division of values is carried out through the thresholds mentioned above, if one or more of these values exceeds a certain limit there is an initial insertion in the group of aggressive drivers.

After the creation of these groups, our algorithm takes into consideration the scores related to those paths characterized by a guide classified as unsuitable and those related to the paths that have instead been included in the group of calm guides.

Consequently, these values create new thresholds, this time related to the scores; in particular, the "Superior Limit Set" and the "Inferior Limit Set" are obtained. The first will contain the values resulting from the mean between the maximum and the mean values of the scores, the second instead will contain those values obtained through the mean between the minimum and the mean values of the scores.

Given A the set of trips defined as aggressive, and C the set of trips defined as calm, from the threshold classification approach, we have that

$$|A| = n$$

$$|C| = m$$

For each of the scores mentioned in 5.1 represented as **ScoreX** are here presented the *Superior Limit* and *Inferior Limit* related to it:

$$SuperiorLimit_{scoreX} = \frac{1}{2}(\max_{i=1,\dots,n}(ScoreX_i) + (\frac{1}{n}\sum_{i=1}^{n} ScoreX_i))$$

$$InferiorLimit_{scoreX} = \frac{1}{2}(\min_{i=1,\dots,m}(ScoreX_i) + (\frac{1}{m}\sum_{i=1}^{m} ScoreX_i))$$

Then we use these superior and inferior limits to create two subsets, one containing the paths having the scores over the superior limits that should be the bad behavior drivers' trips, and one containing the paths having the scores under the inferior limits that should be the good behavior drive trips. All the others that are not inserted in one of these two sets, instead, will simply be considered as drivers with an intermediate guide.

The confirmation of the effective validity of the two-phase classification will then be made by identifying the number of peaks above a certain peak value, to eliminate any classification errors previously made.

5.4 Analysis of peaks

Before introducing the topics that will be addressed in this section, it is necessary to define what we mean with a peak in our analysis, to make it easier to understand all those studies that we are going to explain.

> *Peak: a local maximum among the feature's values variation during the time*

In this section, we will introduce some studies regarding the peaks. The first of them will deal with the identification of the peak thresholds related to the primary features. This, to identify the number of peaks of the specific feature that exceeds this threshold. It will be done for each trip to verify the hypothesis according to which, a path classified as performed with an aggressive approach, presents a greater number of these peaks beyond the thresholds, and therefore to verify that the classifications made in the previous steps are effectively efficient.

Subsequently, a second study of the peaks will be introduced which will be based mainly on the analysis of their coincidence. In particular, after having identified all the peaks that stand out above a certain threshold, we start to enumerate the number of peaks that occurred simultaneously in the different features over a second, to try to identify the type of road faced in that specific period in that specific trip.

5.4.1 Main features' peaks relationship

After having formed these two sets, the next step is to see the number of peaks related to the main features that are present in these trips. However, we have to point out that we will not consider all the peaks in general, but only the ones related to an interval that goes from a quite high value to a very high value of the feature that we are analyzing.

To understand when a peak can be considered an indicator of bad behavior, we have focused ourselves on the identification of some limits, each of them ad hoc for a specific feature. The peaks' limits are created looking for the average peak values of the features, then, after having reached these averages from each trip, like the one that we present in figure 5.16, we've developed a mean of these results to obtain these limits. In particular, regarding the RPM was stated that the minimum threshold value to consider has to be 2000 RPM, for the Speed 40 km/h, and for the Throttle 25% of tilt.

```
speed peaks avg val:  39.78543307086614
rpm peaks avg val:  2067.016544117647
throttle peaks avg val:  33.23741007194245
```

Figure 5.16. Peaks' limits for Speed, RPM and Throttle features

Table 5.1. Trip clustering results among all dataset

DATASET	RPM Peaks Number	Speed Peaks Number	Throttle Peaks Number
Prof- set calm trips (47 trips)	2% 500<x<1500 2% 1500<x<2500 2% x>2500 For a total of 6% of trips over 500 RPM peaks	14% 150<x<600 0% x>600 For a total of 14% of trips having an average of 400 speed peaks	10% 500<x<1000 2% 1000<x<2000 2% x>2000 For a total of 15% of trips over 400 throttle peaks, and a total of 4% of trips over 1000 throttle peaks
Prof- set nervous trips (30 trips)	13% 500<x<1500 7% 1500<x<2500 7% x>2500 For a total of 27% of trips over 500 RPM peaks	17% 150<x<600 7% 600<x<800 7% x>800 For a total of 30% of trips over 150 speed peaks	27% 500<x<1000 10% 1000<x<2000 3.3% x>2000 For a total of 40% of trips over 400 throttle peaks, and a total of 14% of trips over 1000 throttle peaks
Cephas- 14 trips- calm (3 trips)	33% x>=35 So the max RPM peaks of this subset is 35	33% x>=35 So the max speed peaks of this subset is 35	66% 10<x<15 So the avg throttle peaks of this subset is 13
Cephas- 14 trips- nervous (4 trips)	50% 200<x<600 25% x>600 For a total of 75% of trips over 200 RPM peaks	25% 200<x<400 25% x>1000 For a total of 50% of trips over 200 speed peaks	50% 150<x<800 25% x>1000 For a total of 75% of trips having an average of 150 throttle peaks
Cephas- 19 trips- calm (5 trips)	40% 30<x<50 For a total of 40% of trips over 30 RPM peaks	60% 35<x<50 40% x>50 For a total of 100% of trips between 35 and 60 speed peaks	40% 100<x<110 40% x>110 For a total of 80% of trips between 100 and 115 throttle peaks
Cephas- 19 trips- nervous (8 trips)	25% 30<x<70 25% x>70 For a total of 50% of trips over 30 RPM peaks	38% 35<x<50 25% x>50 For a total of 63% of trips between 35 and 55 speed peaks	38% 60<x<100 25% x>100 For a total of 63% of trips between 60 and 120 throttle peaks
Seoul- set calm trips (8 trips)	87% 80<x<140 13% x>180 For a total of 50% of trips over 120 RPM peaks	25% x<50 50% 50<x<60 25% 80<x<90 For a total of 75% of trips over 50 speed peaks	38% 30<x<50 50% x>100 For a total of 66% of trips over 40 throttle peaks
Seoul- set nervous trips (3 trips)	66% 80<x<140 33% x>200 For a total of 66% of trips over 120 RPM peaks	66% 40<x<60 33% x>100 For a total of 66% of trips over 50 speed peaks	66% 40<x<100 33% x>160 For a total of 66% of trips over 80 throttle peaks

Now, with these thresholds well defined, our peaks analysis can finally be introduced. The first step is to understand if the following hypothesis is true: if a trip is clustered in the bad-behavior set of trips, it should have a higher number of peaks that overtake the margins derived from the means of the peaks' values. We can see the table of the results in figure 5.1.

Since we are using a set of thresholds based on the mean between the maximum (or minimum) and the means of the score's values, if we have a small set of trips in which there is no so much difference among the scores of the trips, the subdivision in two trip sets is done, but the results of the analysis on these two sets are quite similar even if we are talking about two different sets. For this reason, to have a sharp difference among the results, we have to use a set composed of a non-small number of trips, or, at least, a set of trips that present very different scores.

The example of what we are saying can be found in the sets of **Cephas19**, in which the values in the "good" trips set are higher than the ones in the "bad" trips set, this because the number of the trips is small and the scores are very similar

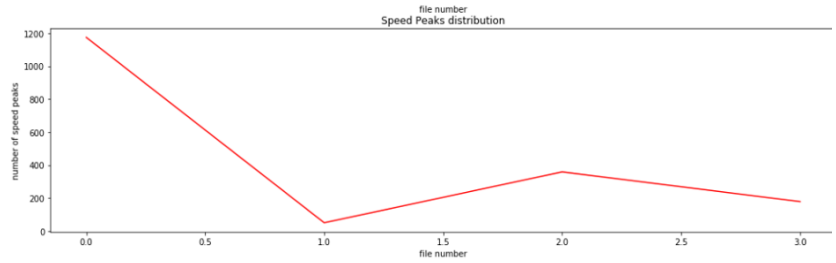


Figure 5.18. Couple IDTrip-Speed Peaks number for Cephas14's classified trips

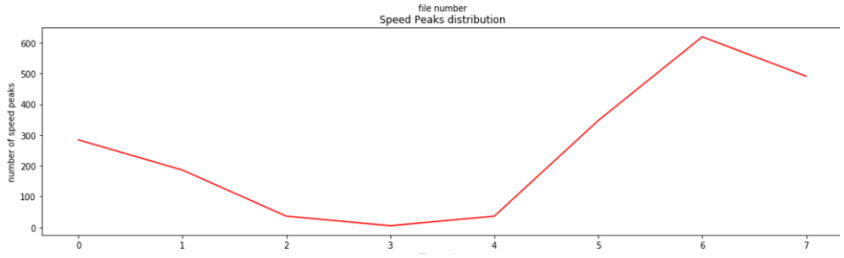


Figure 5.19. Couple IDTrip-Speed Peaks number for Cephas14's non-classified trips

each other.

A different scenery, instead, is proposed from the group of **Cephas14** in which, despite the few trips, the score values difference is very distinct and this leads to having the "bad" set of trips contains peaks way higher than the "good" one.

Aggressiveness Score	Ecoscore
0.43216785062612595	102
0.5508245371347755	1871
0.38545139711013277	58
0.7139779963171424	199
0.25388069708255756	8
0.19531864894742407	0
0.2525814609026823	0
0.34056191543313374	149
0.32736018433401454	227
0.40443317595559386	5472
0.7138204202202187	177
0.37484448026661493	64

Figure 5.17. Scores Trends

As stated, in 5.17 the Eco-Score values are way diverse between the calm trips and the nervous trips, while the aggressiveness score presents some differences that are not so much noticeable.

This allows to develop on the Cephas14 a quite good classification even with a smaller number of scores rather than the other dataset. In particular, we present the peaks comparison of the classified and non-classified trips, for both Speed (5.18)(5.19), Rpm (5.20)(5.21) and Throttle factors (5.22)(5.23).

However, it is not a perfect classification, in fact, the lack of some scores and the small trips number, leave unclassified some trips having a high number of peaks (in both RPM, Speed and Throttle scenarios)

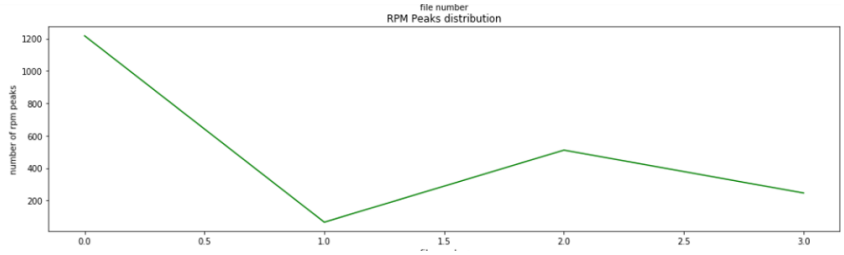


Figure 5.20. Couple IDTrip-RPM Peaks number for Cephas14's classified trips

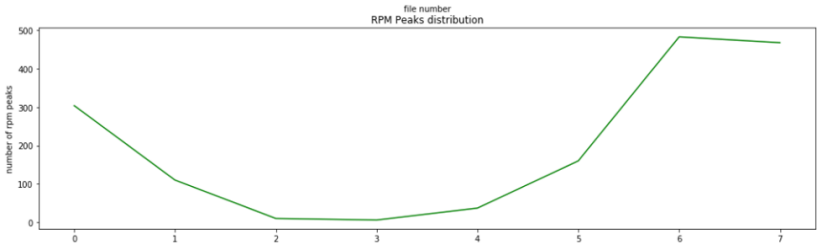


Figure 5.21. Couple IDTrip-RPM Peaks number for Cephas14's non-classified trips

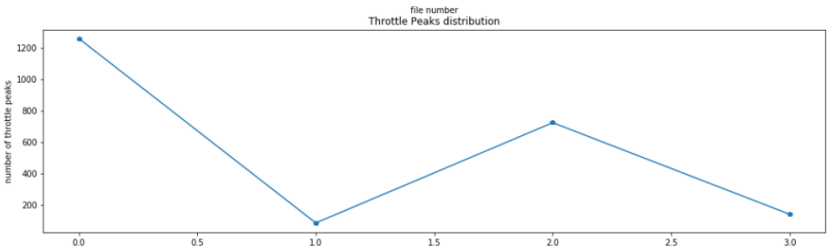


Figure 5.22. Couple IDTrip-Throttle Peaks number for Cephas14's classified trips

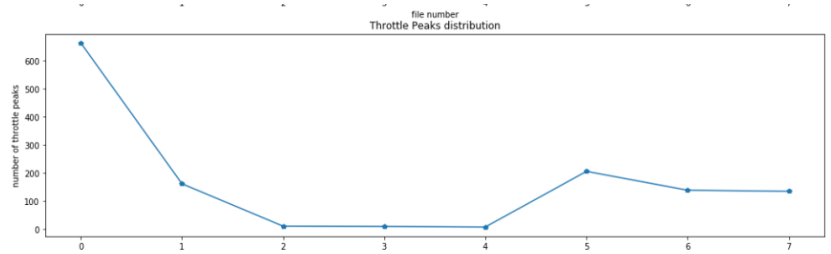


Figure 5.23. Couple IDTrip-Throttle Peaks number for Cephas14's non-classified trips

Way different, instead, is the situation in the **SparkWorks dataset**. Here, the possibility of having a high number of trips, based on different paths and different driving styles, confirms our hypothesis about the drivers' distinction. The first thing to notice is that, even if we have a calm trip set that is slightly larger than the nervous one, the result does not change. Looking at the percentage, we can notice that the nervous trips set presents a higher number of trips with more than 500 RPM peaks (5.26) with respect to the calm trips (5.27), and this is repeated for all the other features. The speed peaks number, in fact, for the calm trips are about 400 (5.25), while in the nervous trips (5.24) reach values near one thousand, and even in the throttle feature, we notice the distinction peaks number, where we have the 4% of trips over 1000 throttle peaks of the calm set (5.29) against the 14% of trips over 1000 throttle peaks of the nervous set (5.28).

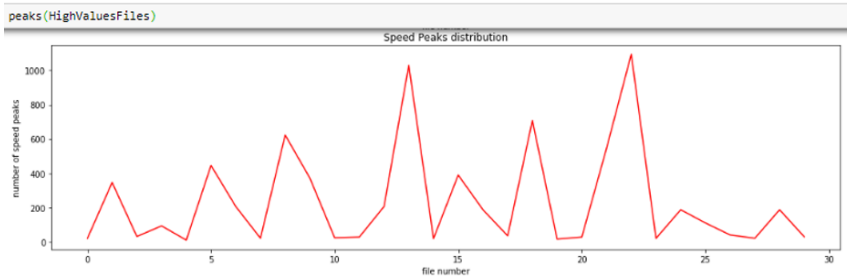


Figure 5.24. Speed peaks' number for each trip classified as bad behavioral in SparkWorks dataset

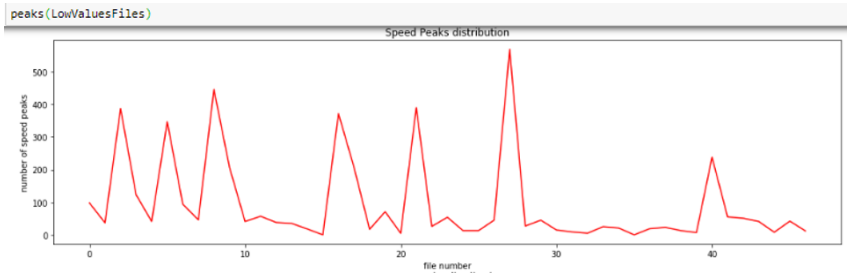


Figure 5.25. Speed peaks' number for each trip classified as good behavioral in SparkWorks dataset

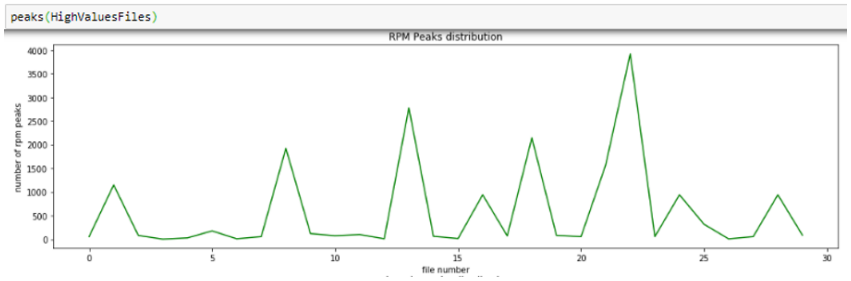


Figure 5.26. RPM peaks' number for each trip classified as bad behavioral in SparkWorks dataset

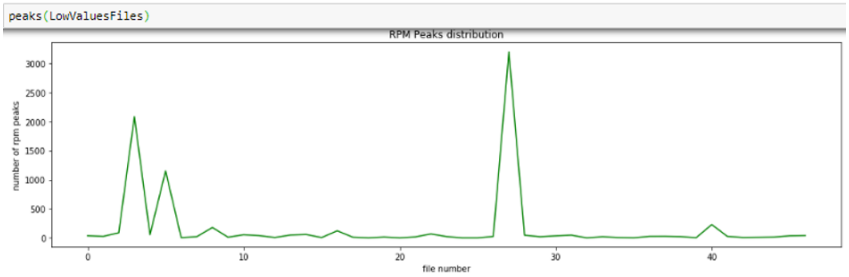


Figure 5.27. RPM peaks' number for each trip classified as good behavioral in SparkWorks dataset

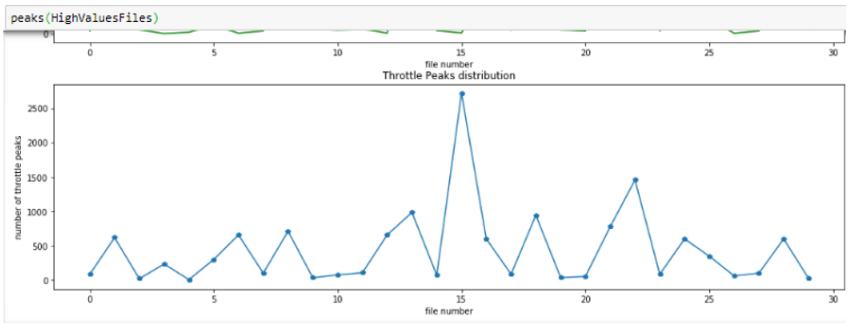


Figure 5.28. Throttle peaks' number for each trip classified as bad behavioral in SparkWorks dataset

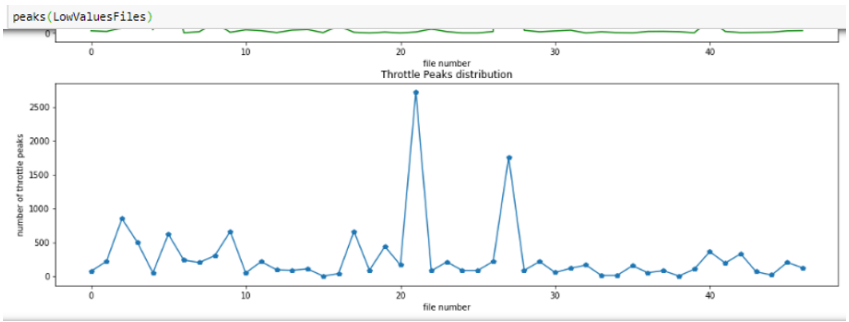


Figure 5.29. Throttle peaks' number for each trip classified as good behavioral in SparkWorks dataset

We have understood that under precise conditions, like enough trips number and different score values among these trips, is possible to classify the driver behavior basing on the usage of the values' thresholds (5.2.1), score values (5.1), and the existence of peaks that go over some specific floor. This intuition is now strengthened from the idea of looking for coincident peaks among many features, in this case, the main ones. This peaks coincidence tells us that, even if the classification split the trips in calm and nervous driving, and even if there are more peaks coincidence in the nervous trip rather than calm ones, there are some exceptions in both the scenarios. The first is that some of the nervous drive trips does not present peaks coincidence, and the second is that few trips that were classified as calm, present a

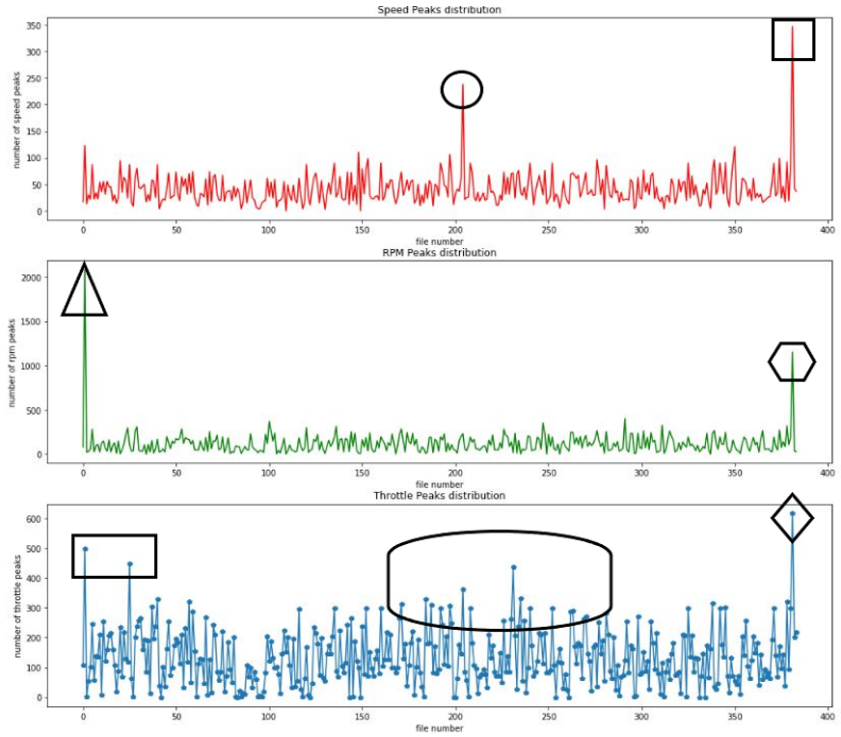


Figure 5.30. Representation of all the SparkWorks dataset's trips with their number of peaks

set of peaks coincidence. One may think that an element influencing these errors could be the different trip times of the paths (as for the SparkWorks dataset) but this is dispelled by the plot of the Seoul dataset. In that scenario the trip times were all the same and these problems are still there, so, the fact that the trip times are different, has to be taken into consideration but we cannot reconduct our situation only to this factor.

This dispel, is confirmed through the subdivision of the SparkWorks dataset's trips in more sets basing on the trip duration. The trips with a higher number of peaks, after the creation of subset based on time similarity, are not individuated in their totality. It happens both if we consider the general thresholds obtained from the set in its entirety (Case A) or if we look up for thresholds created ad hoc for each subset (Case B) .

To present the proof of what we are affirming, we firstly introduce all the trips peaks' number, where the underlined ones are the paths having a higher number of peaks (5.30).

Consequently, we analyze the *Case A*, in which we have the creation of a certain number of subset based on the trips' length

- Greater or equal than 2 hours of trip time
- Between 2 hours and 1 hour of trip time
- Between 1 hour and half an hour of trip time

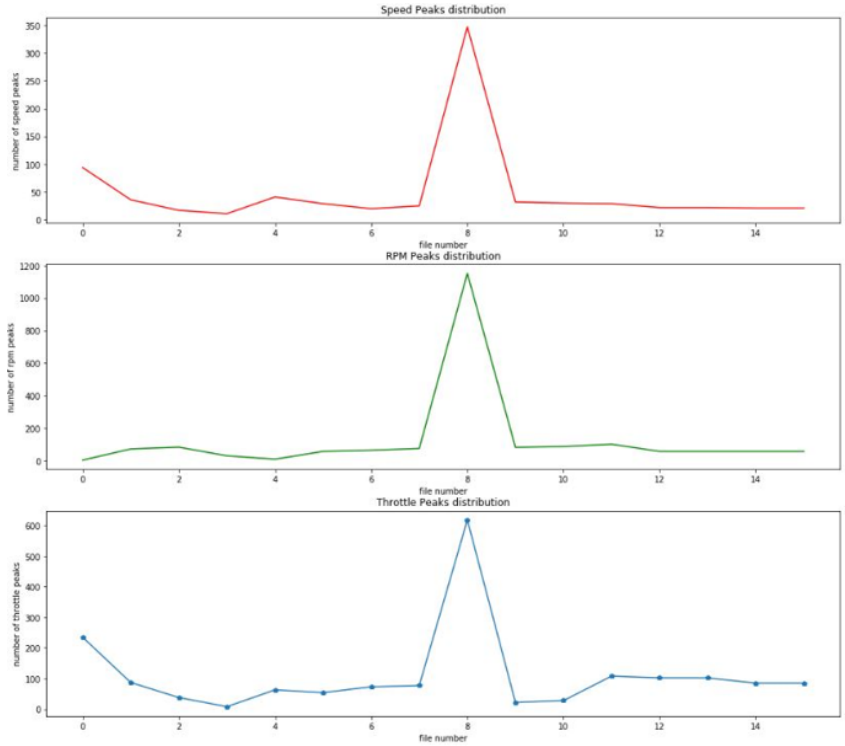


Figure 5.31. Trips' peaks classified as nervuos driving style (Case A)

- Less than half an hour of trip time

and a classification between calm behavior and nervous behavior made up through thresholds obtained from all the trips' value averages.

We notice that even doing the above-mentioned operations, the situation is not solved, firstly because in the plot 5.31 a lot of trips having the higher peaks number are not individuated. Secondly because, in the plot 5.32 where are classified the trips labeled as good behavior, we find some elements that should not be present to validate our classification technique.

Then, we consider the *Case B*, where the subset composition is the same as the previous case, but the thresholds are created ad hoc from each of these subsets instead of coming from the entirety of the trips. We can say that, even in this situation, the result is not different.

Even if we don't have good behavioral trips (with small peaks number) wrongly classified as bad behavioral (5.33) we notice that, comparing the result with the plot presenting all the trips' peaks of that specific subset (5.34), a group of them having a high number of peaks was not classified as nervous behavior.

Another probable element that could be the cause of these misclassifications, is the presence of some score or features that we have not considered until now, and that could complete the classification approach, through the merging between the set of nervous driving style and the wrong-classified trips that present high peaks coincidence.

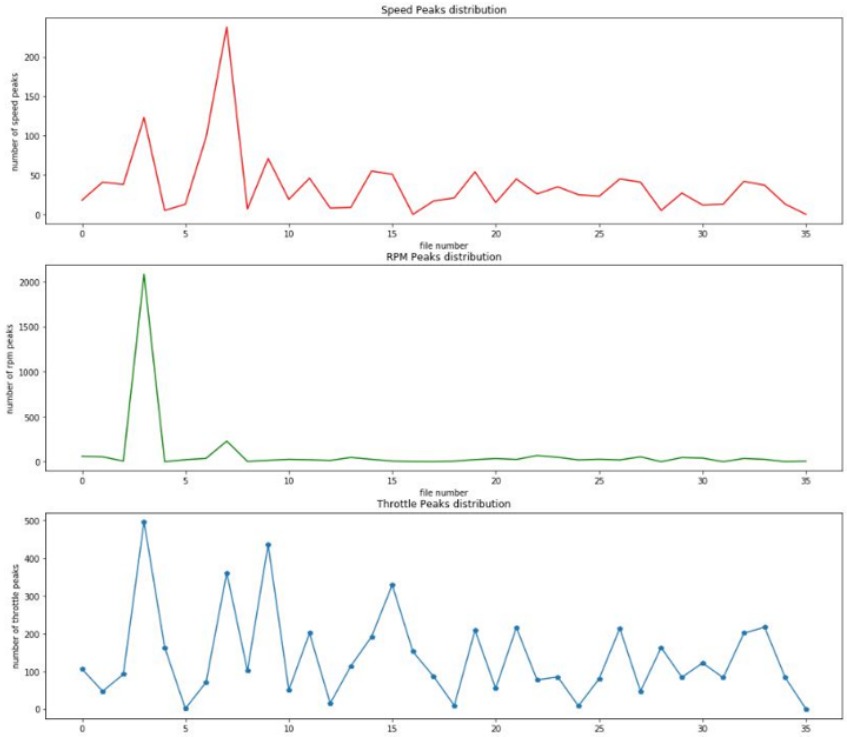


Figure 5.32. Trips' peaks classified as calm driving style (Case A)

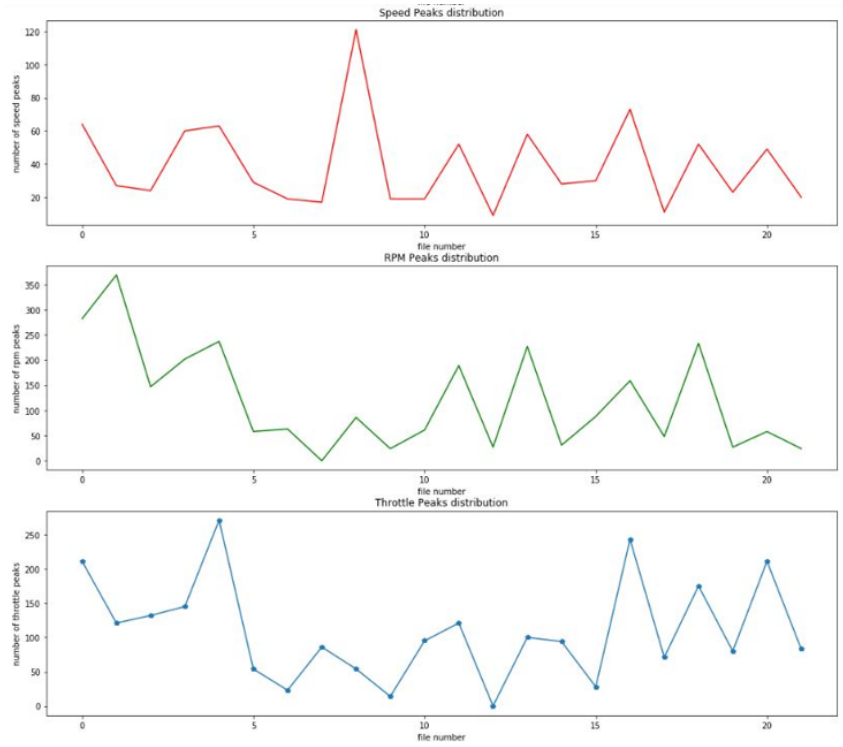


Figure 5.33. Trips' peaks classified as nervuos driving style (Case B)

Figure 5.34. Trips' peaks of the Case B subset

5.4.2 Scores and secondary features' classification support

Our next step is to use the secondary features, secondary thresholds, and secondary scores for developing a more complete and correct classification of trips, to use their peaks coincidence to confirm our hypothesis of the clustering approach. Taking care of the different trip time of these trips.

Let's take a short recap of which are the secondary features that we have noticed until now for each dataset, to explain which of them are usable in this context.

For what regards the SparkWorks dataset, we have managed to find scores like Aggressiveness Score and RPM Eco-Score. Near to them, we have also found some other elements like the Gear ShiftUp, that can be calculated on each of the three Acceleration axis returning us the Acceleration X ShiftUp, Acceleration Y ShiftUp, and Acceleration Z ShiftUp, the Acceleration Events coming from the analysis of these three axes, and the ShiftUP Eco-Score, obtained from [10] and [11]

Trips time distinction for SparkWorks dataset

The first thing to do is to integrate these secondary factors in the mechanism of paths splitting, to understand which and how they influence the trip classification.

We have two situations, the first in which we consider the overall set of trips not splitted, and the second in which we create a group of sub-set containing the trips divided per trip duration (5.4.1), in which there were some clustering errors.

After doing this splitting, and calculating the thresholds of each specific subset, we obtain this classification of the trips, where "High" stands for the trips classified

as nervous driving style, and "Low" for the trips classified as calm driving style, for the various subset with ad hoc thresholds,

```
Two Hours Ad Hoc, High:  0 on 10  Low:  0  on  10
One Hours Ad Hoc, High:  0 on 11  Low:  0  on  11
Half Hour Sup Ad Hoc, High:  1 on 83  Low:  1  on  83
Half Hour Inf Ad Hoc, High:  22 on 384  Low:  25  on  384
```

Figure 5.35. Trips' classification through ad hoc thresholds

and for the split sets, with general thresholds.

```
Two Hours, High:  2 on 10  Low:  0  on  10
One Hours, High:  5 on 11  Low:  0  on  11
Half Hour Sup, High:  7 on 83  Low:  3  on  83
Half Hour Inf, High:  16 on 384  Low:  36  on  384
```

Figure 5.36. Trips' classification through general thresholds

Now, we can switch to the insertion of secondary features to optimize the path distinction.

We have applied, with many experiments, all the above-mentioned features to understand which subset of them manage to improve our trip clustering. The result is that, for the general situation only the *Acceleration Event on the X axis*, and the *Gear ShiftUp Score related to the Acceleration on the X axis* brought some additional trips (5.37).

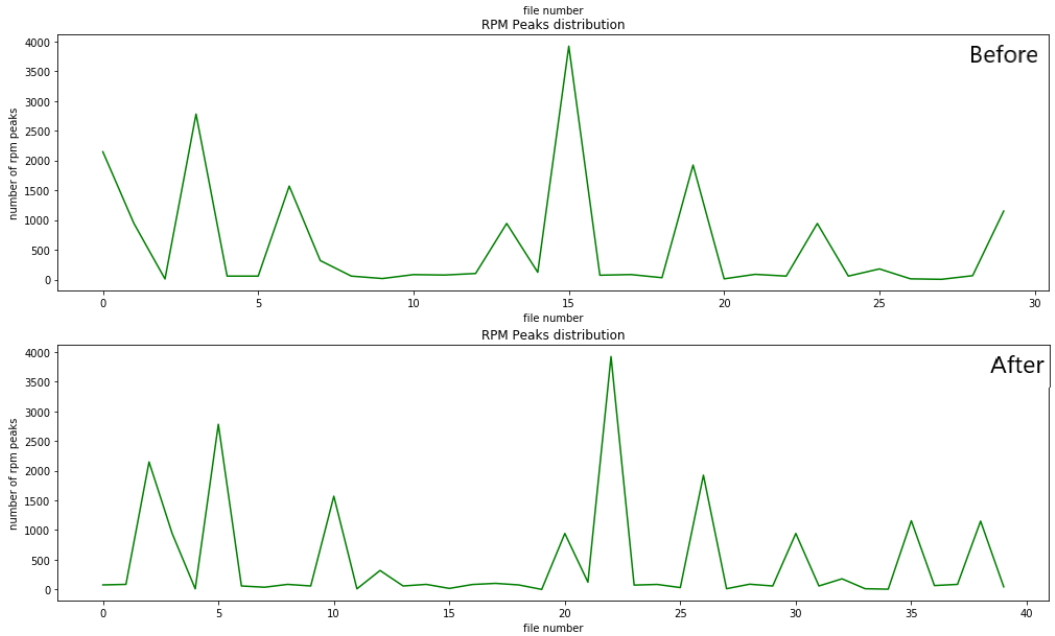


Figure 5.37. RPM peaks number' comparison between two classification approaches for bad behavior set

In any case, the difference is not so evident, and most of all, looking to the

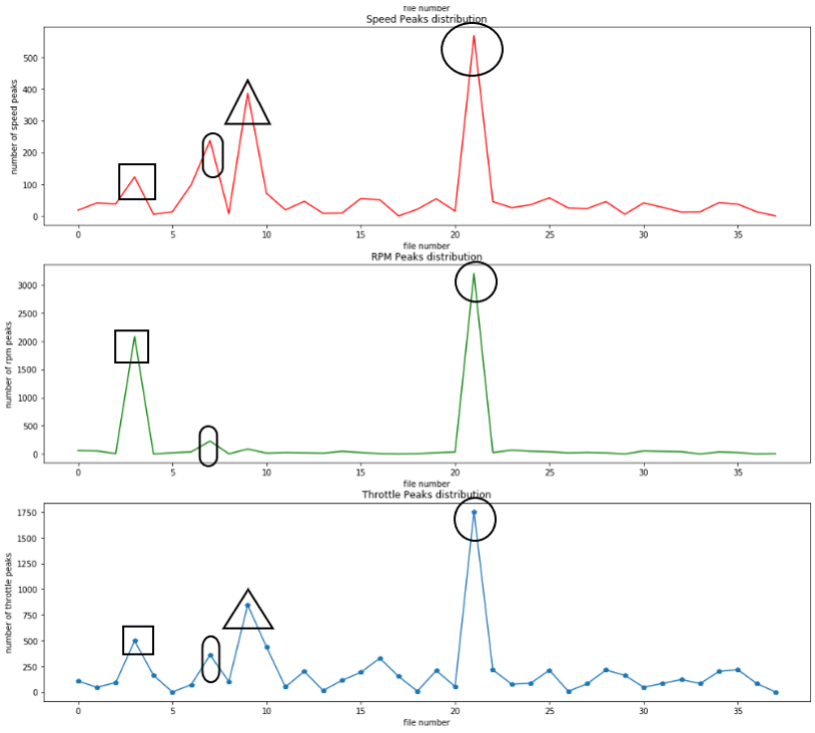


Figure 5.38. Not classified trips

nonclassified trips (5.38) we notice that some of them should be inserted on the bad behavior but they are not taken in consideration.

Another attempt was done with the usage of ShiftUp Ecoscore, which is based on the mechanism explained in [10]. Also here the result was not so good. It adds some trips to the set of nervous driving style but their majority presents a low number of peaks, and this is not correct.

After this, we switched in the other scenario of the sub-set subdivision (Case B), but, even here, the result wasn't better, in fact, neither the Acceleration Event on the X-axis, nor the Gear ShiftUp Score related to the Acceleration on the X-axis brought good news. There is, however, a thing that must be said; the ShiftUp Ecoscore in this situation was the best factor among all. It adds a big number of trips having some high peaks, but again, the problems introduced are two. The first is that it also adds trips with a small number of peaks to the list of the high numbered peaks trips, and the second is that all these additional trips are related only to the sub-set of trips with short trip duration (as we can see in 5.35). Surprisingly, in both scenarios, not even the HeartRate average value of the trips helps us in solving these problems.

The study presents again the problem introduced previously (5.35), where, for case B, the clustering was done only for the sub-set trips with short trips. For case A, instead, there are some interesting secondary features not as strong as the one that we will see in the beneath scenarios, but still useful for the behavior splitting. However, even in case A, there is the problem of the wrong classification due to the insertion of some false positives; let's try to understand why this happens.

Looking carefully the trips causing us troubles, we have discovered that in most

of them the errors are caused by the lack of Gear identifier, very small sampling ratio of Acceleration factors, or other values' problems. This explains why the peaks are individuated but the scores do not improve the classification, because the scores are created on missing or corrupted information, even after the dataset cleaning, and this leads to erroneous values, lower than they should be. Anyway, this situation gives us intuition about these peaks. Even if the scores are not useful in some cases, like where the values are negatively influenced by the lack of some information, there is the possibility that looking at the coincidence of peaks among the main features we can still understand if a trip can be identified as bad or good. In addition to this, there is the possibility that the common peaks could help us in understanding which type of road is traversed. This will be covered in section 5.4.3, in which we will look for the number of coincidence of these peaks to understand road type and, maybe, in improving the driving style's classification.

Unified trip times

Different is the situation that we face in the other dataset. In these scenarios, we don't have trips' duration difference, so it is useless to apply a case subdivision as the one done in the 5.4.1 and 5.4.2. However, we must improve the remaining dataset' classification through the insertion of scores, and secondary features.

Starting with the **Cephas dataset**, we immediately notice that, in both the section of 14 and 19 drivers, due to a lack of some information, the only thing that we could use as secondary features would be the MAF and the IAP (Intake Air Pressure). Nevertheless, as we have seen in section 5.2.2, they are not useful for our analysis because they are not bounded with any of the other main features in a strong way.

For what concerns the **Seoul dataset**, we can't find out all the axis sections of the acceleration, but we still can take into consideration the Gear ShiftUP related to Acceleration lateral and Acceleration longitudinal. Near to this we also want to look at the counters related to the Steering Wheel Events (singular or continuous) and Lateral Acceleration Events.

The main point to focus on is which of these features we should use. The best result among the various tests is the one in which we consider the Events of the Acceleration Latitudinal and the Steering Wheel Event number. These two elements allow us to distribute in a better way the trips through a more precise categorization in the two sets of trips. We can see in detail how the usage of these values changes the distribution of the sample paths in the 5.39 and 5.42.

In the images 5.39, 5.40 is clear how the use of these two additional elements change the sets, in particular, the number of selected trips increase from two to eight, but most of all, we notice that the bigger part of the additional trips are coherent with the subdivision that we aim to reach. In other words, the trips selected presents a high number of thresholds peaks, and this is what we were looking for. There is one additional doubt, if the selected trips are the ones with higher values in the overall set or if there are other paths with a high number of peaks that were not considered.

To answer these questions, the only solution is to propose the distribution of the peaks of all paths.

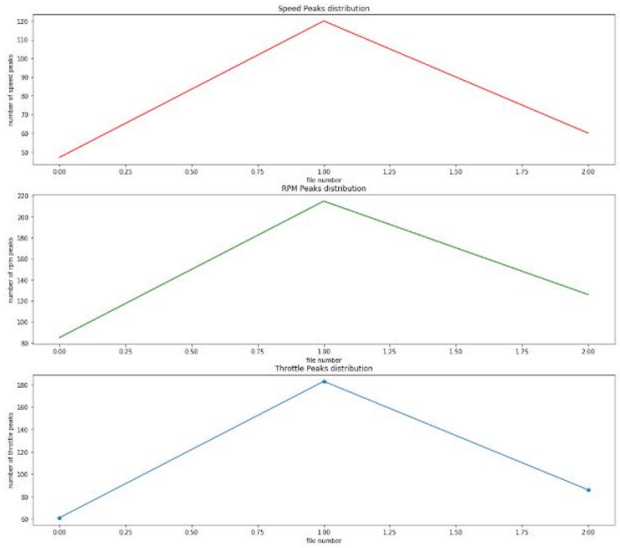


Figure 5.39. Peaks of bad classified trips before the secondary features' usage

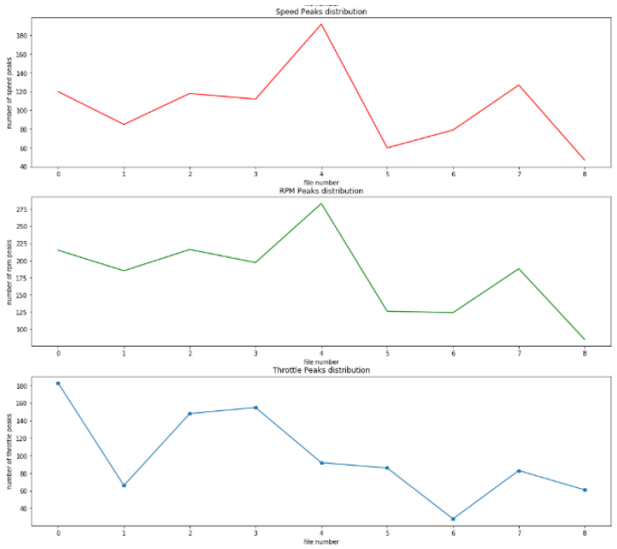


Figure 5.40. Peaks of bad classified trips after the secondary features' usage

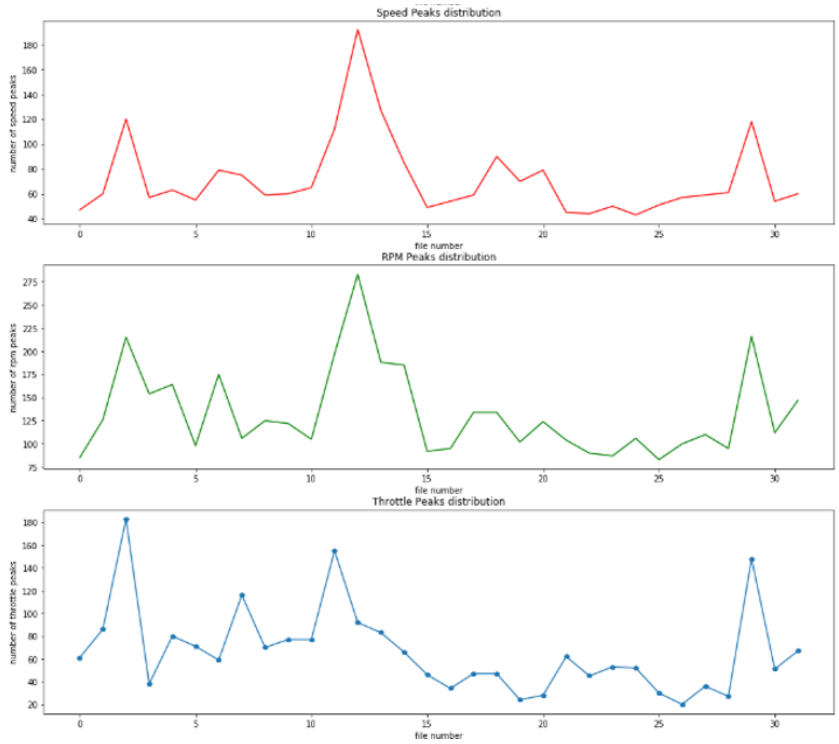


Figure 5.41. Trips' peaks number for all the Seoul dataset' trips

As we can see, the three trips with higher peaks of the Speed are present even in our selection. The same is for the five trips having several RPM peaks over 175, and for the four trips having many Throttle peaks over 100.

The differences can be noticed also through this table.

DATASET	RPM Peaks Number	Speed Peaks Number	Throttle Peaks Number
Seoul- set old nervous trips (3 trips)	66% 80<x<140 33% x>200 For a total of 66% of trips over 120 RPM peaks	66% 40<x<60 33% x>100 For a total of 66% of trips over 50 speed peaks	66% 40<x<100 33% x>160 For a total of 66% of trips over 80 throttle peaks
Seoul- set new nervous trips (9 trips)	33% 80<x<140 33% x>200 For a total of 88% of trips over 120 RPM peaks	11% 40<x<60 55% x>100 For a total of 90% of trips over 50 speed peaks	55% 40<x<100 11% x>160 For a total of 66% of trips over 80 throttle peaks

If we look at the trips inserted in the "calm" set, we have another proof. We notice how the elements with very high peaks number are not anymore present, or, anyway, their number is reduced leaving only very few exceptions, with a peaks number not so high.

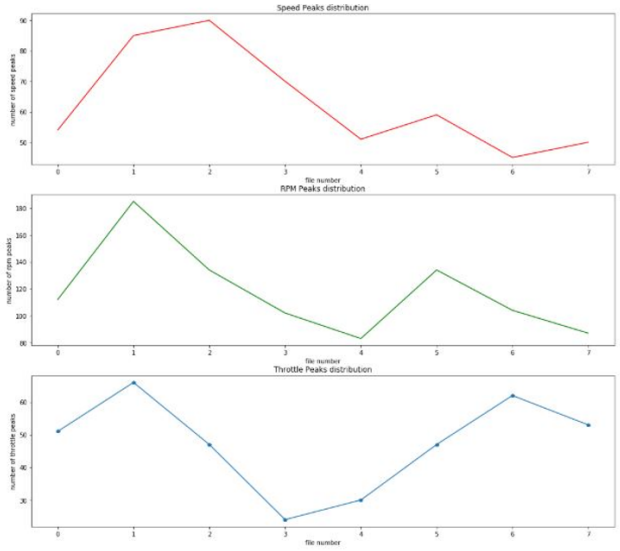


Figure 5.42. Peaks of calm classified trips before the secondary features' usage

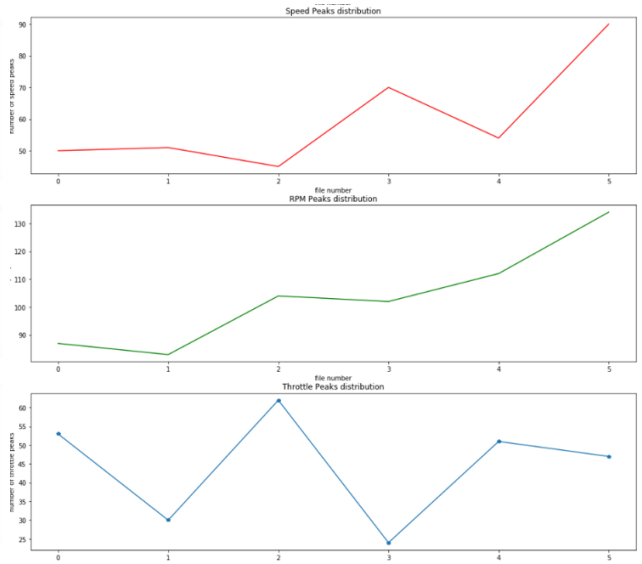


Figure 5.43. Peaks of calm classified trips after the secondary features' usage

This cancels all the previous doubt, and allow us to connect these two new features, the *Events of the Acceleration Latitudinal*, and the *Steering Wheel Event number*, with the RPM Ecoscore and the Aggressiveness score, to define a nice method for classify the driving trips in two sets, the one based on trips characterized by smooth and calm driving style, and the one based on trips characterized nervous and reckless driving style. Remember, anyhow, that this approach is shaped up on this specific dataset, in the sense that, it can also be used in other situations, but needs to have samples regarding certain features, without which it would be useless to exploit it. In fact, if we apply this approach to the dataset of Cephas Barreto, since those trips contain other features sampling, it will not give the expected result.

5.4.3 Peaks coincidence support

We start to connect the results came out from the analysis of the peak of the main feature scores and the secondary scores, with the idea of using the coincidence of

these peaks to support the scores where their results are corrupted by erroneous values, and to use them to distinguish which type of road was traversed. In particular, we have focused on the interconnection of this high number of peaks for each feature, with the possibility that some of them could present themselves in the same period of other peaks of the other features.

The starting point is to look up for coincident peaks among the main features, and the hypothesis is that the coincidence of peaks will be higher for the nervous driving trips rather than the calm ones. Let's see if something is interesting and if our supposition can be confirmed from the beneath plots.

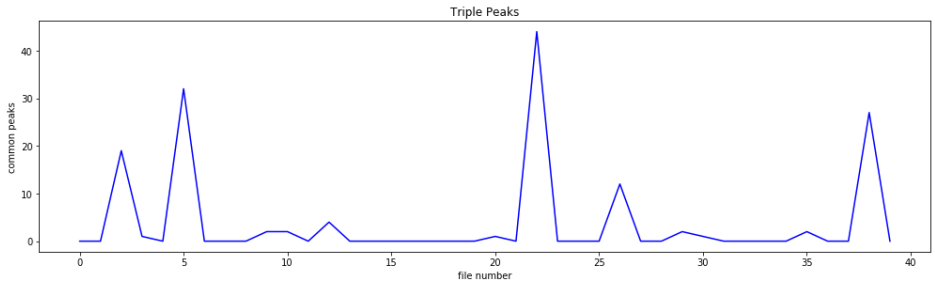


Figure 5.44. Triple peaks coincidence for bad behavior trips classification

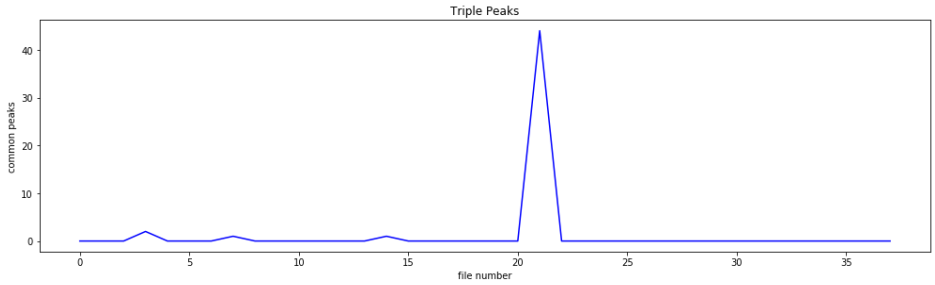


Figure 5.45. Triple peaks coincidence for good behavior trips classification

In this scenario, we can see if three features peaks appear at the same time and, most of all, how many times this happens in each trip. As we could imagine there are more coincidences in the nervous driving trips rather than in the calm ones, even if there are some exceptions that must be isolated among the calm paths, that we have discovered to be trips with corrupted values in which the scores have a powerless classification.

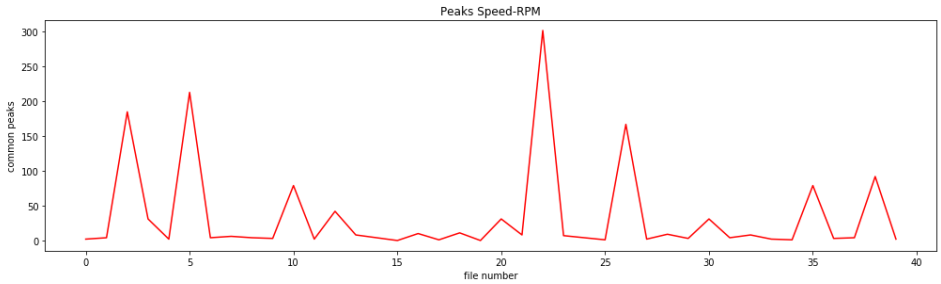


Figure 5.46. Speed-RPM peaks coincidence for bad behavior trips classification

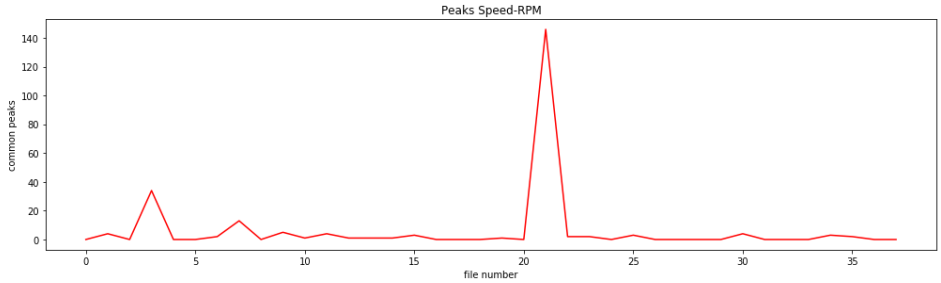


Figure 5.47. Speed-RPM peaks coincidence for good behavior trips classification

Also in the Speed-RPM coincidence the hypothesis is confirmed, there are way more peaks coincidence between speed and RPM features in the nervous drive trips rather than the calm ones, even if there are still some exceptions that must be excluded. But there is another fact, also the number of the couple coincidences is way higher in the trips that are classified as nervous driving, in fact, they reach even 300 couples (trip 22 of this set).

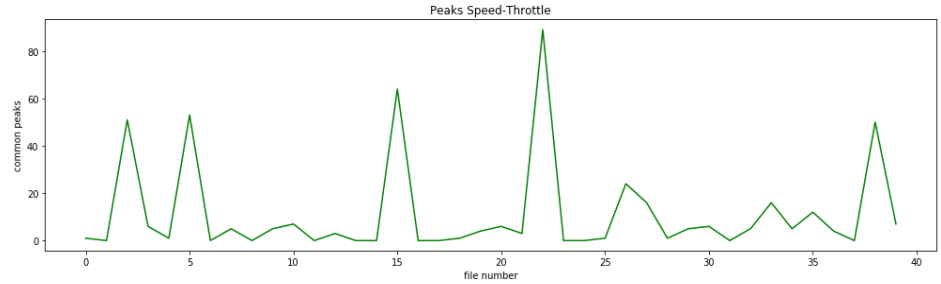


Figure 5.48. Speed-Throttle peaks coincidence for bad behavior trips classification

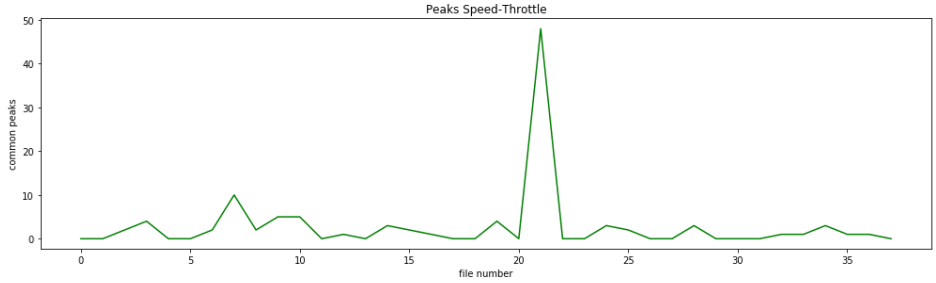


Figure 5.49. Speed-Throttle peaks coincidence for good behavior trips classification

Unlike the other peaks sceneries, in the plot regarding Speed and Throttle peaks coincidence, it is shown a weaker difference among the nervous-classified trips and the calm-classified ones, even if some "bad" paths present a bigger amount of coincidence.

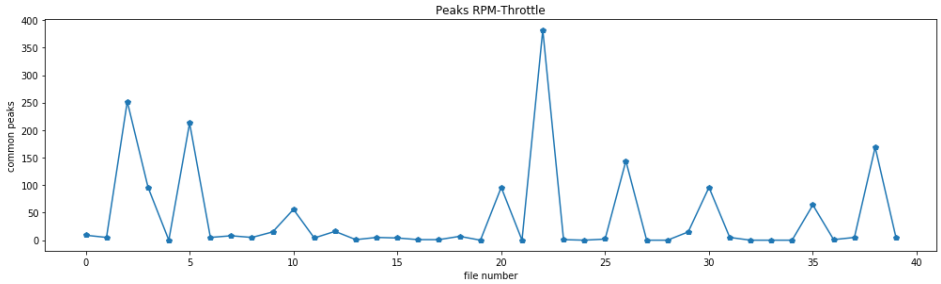


Figure 5.50. RPM-Throttle peaks coincidence for bad behavior trips classification

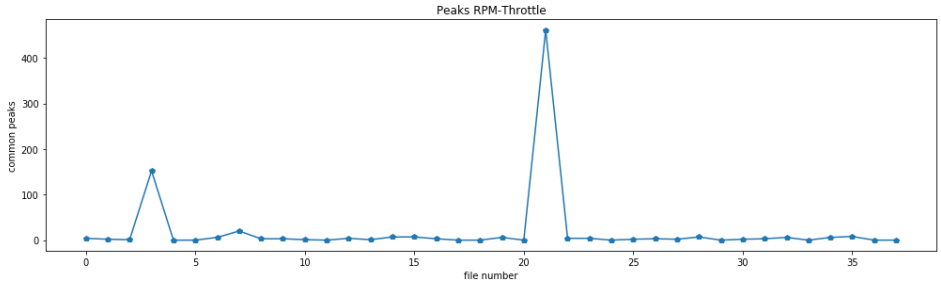


Figure 5.51. RPM-Throttle peaks coincidence for good behavior trips classification

Finally, these are the coincidences between the RPM peaks and the Throttle peaks. The bigger difference is in the number of trips having a couple of coincident peaks, ten against three. Again, these coincidences in the calm paths could be eliminated if during the selection we managed to find another score that helps us to classify in a more precise way the trips in the different sets.

Common peaks analysis and trip type individuation

In this section, we want to amplify the analysis regarding the trips' peaks number and trips' peaks coincidences, both for searching additional useful elements that we could use in our algorithm, and see if their coincidence could help identify the type of road traversed.

Our first step is to obtain the list of paths that presents some coincidence of peaks, both for all the three main features, both for the different couples of main features. This is done for each of the two datasets in which we have managed to develop the previous study of common peaks, the Seoul one and the SparkWorks one.

Starting with the **Seoul dataset**, once that we have obtained the names of the trips that present these peaks coincidences, we start to look for some features that could help in classifying the type of path in which these couple of peaks are discovered. The problem, again, is the lack of information, here the only feature that could help us classify the trip type among urban-miscellaneous-highway is the Gear array. The idea is to obtain the average of these Gear values of each trip and see if there are hints for this path classification, or even if these Gear average can be used to classify the driver behavior when the scores are useless.

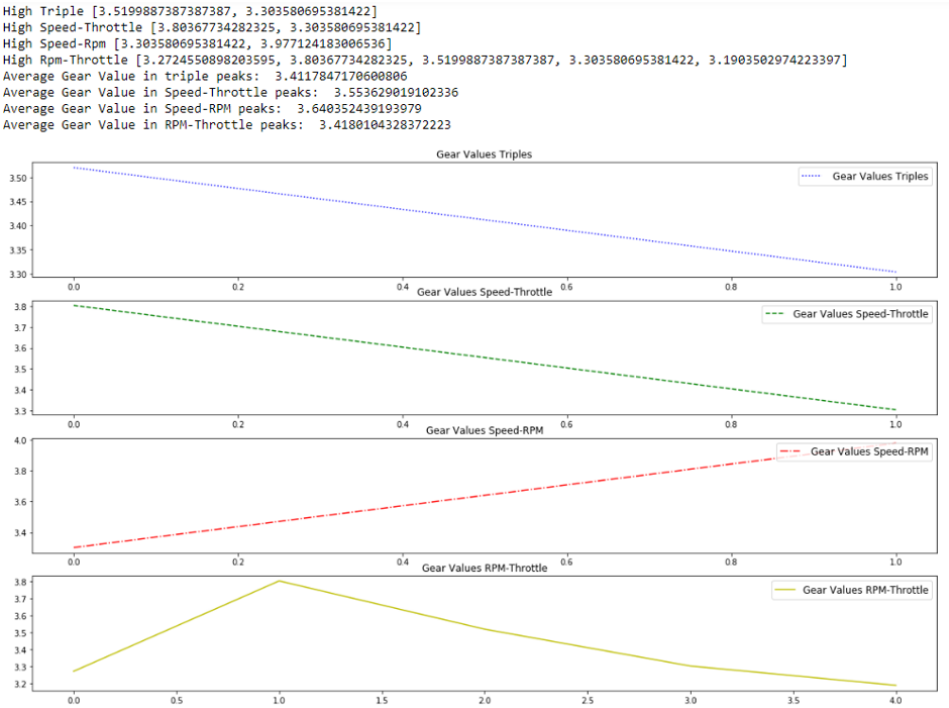


Figure 5.52. Gear Analysis for Seoul dataset

What we obtained from this analysis is a too small subset of trips that present peaks coincidence, due also to the small number of trips present in the starting set. So, near to the problem of the lack of useful features, we have also the small set's size that not give us the expected results.

Let's then switch to the **SparkWorks dataset**, after having obtained the list of trip names having common peaks, the first step is to individuate the helpful features for classifying the path and for clustering the driver behavior. The results are better than the previous scenario where we have again the Gear values array, but also the different speed limits of the traversed roads. These features were not available in the Seoul dataset because, as explained in the reverse geocoding section, that dataset does not present any GPS position (Latitude-Longitude couple) in the paths.

The first element that we will focus on is the one related to the Gear values (5.53).

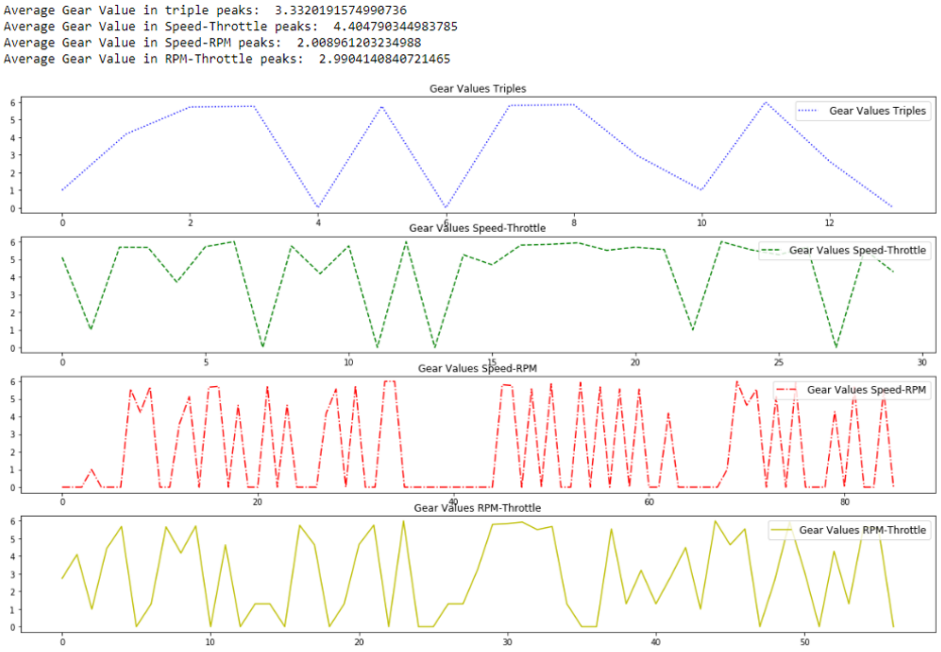


Figure 5.53. Gear Analysis for SparkWorks dataset

If we look deeply the image 5.53, we can see how much difference there is concerning the same analysis done on the other dataset. With only a higher number of trips, we can obtain way more information. The highest average Gear clutch value, is the one related to the common peaks couple of Speed and Throttle, but why this result? This is exactly what we were expecting for, if we think about of the different path types, and we take into account the ones related to the highway we could imagine that in the highway, except for high traffic condition, the gear value distribution is near the high values, because the vehicle is already running and we don't have many start-and-stop actions. Near to this if we consider the general case of a car on a highway, we can imagine that the higher number of peaks will be the ones related to throttle pedal and speed, this because the RPM values, even if the engine could constantly be under hard work, are not changing repeatedly in time, so they did not develop peaks of high intensity. Different is the situation for the speed peaks and, most of all, the throttle ones, because in a scenario based on high velocity, the pedal pressure is the element that changes most. This hypothesis is confirmed by the fact that most of the trips presenting that Speed-Throttle peaks coincidence also presents a high average value for the inserted gear.

Another result is the one related to the Gear average coming from the paths having Speed and RPM coincident peaks. Here the average value is very low, so, a reason for this could be that when we develop many starts and stops, we apply the following steps:

– Insert the first gear, that, with a small variation of the accelerator pedal develops a peak on the velocity, and due to the short gear, it develops also an

RPM peak

- Insert the other gears gradually repeating the situation explained in the previous step
- Insert a lower gear to use the engine brake, that will slow the car and will raise the RPM
- Eventually stop the car and put the gear selector in a neutral position (so a 0 value)
- Repeat the steps during the time

This means that the Gear average will be lower due to the neutral position factors and the low values factors. This also means that the trips presenting these types of peaks coincidence have a higher probability to be trip in an urban context or a traffic situation.

A way to understand if our reasoning is correct is to look to the Traffic factor of these trips.

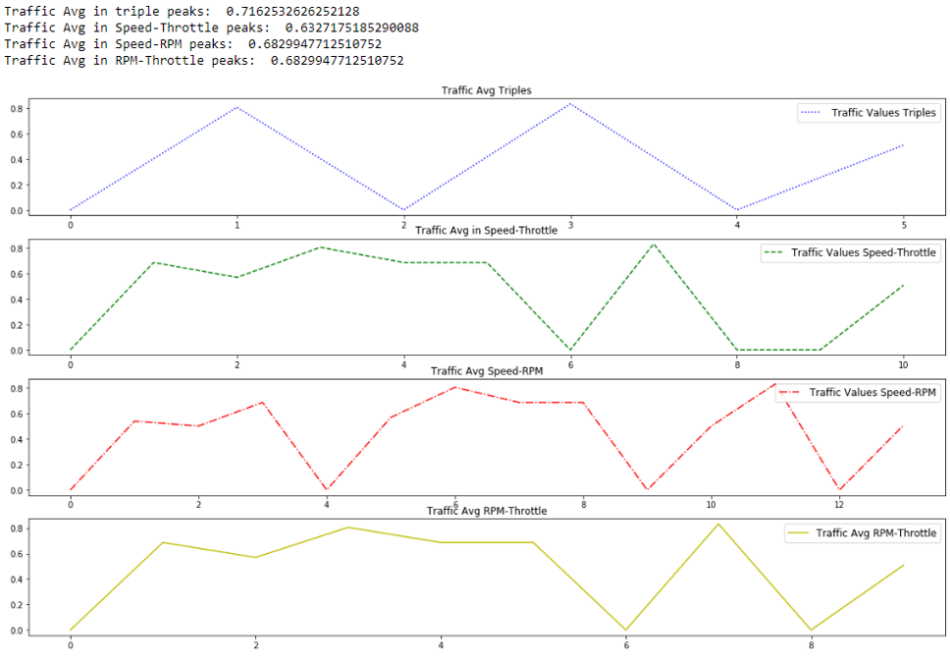


Figure 5.54. Traffic confidence analysis for SparkWorks dataset

As we suppose, the average value of the Traffic Confidence is higher in the trips presenting the coincidence peaks of RPM and Throttle rather than the coincidence peaks of Speed and Throttle. However, we have to notice that this difference is not so sharp, so we cannot assume this information as actual proof of our thought but only as a hint of good intuition.

The last coincidence of peaks, the one related to the RPM and Throttle can be useful for understanding if the driver has a bad driving style. Taking into account the two situations above, because the Gear average value in this scenery is around

three, suggest that the trip type is an urban or a miscellaneous one, in which the driver makes a strong pressure on the accelerator pedal developing high Throttle peaks. This, in union with the RPM peaks, suggests that the driving style is a reckless one because in an urban or miscellaneous trip there is no necessity of high pedal pressure. This idea is strengthened by the Traffic Confidence average value, underlined above, that is the same as the Speed-RPM coincidence, or in other words, related to an urban context.

Then we have the Triple coincidence that presents a higher Gear average value with respect the other situations, and, most of all, that presents the higher Traffic Value, that could be an additional hint to understand that we are not talking about a highway trip type (due to the quite low Gear average value). This last value can be also seen as factors of driver's nervousness, that apply a reckless driving style in an urban-miscellaneous scenario (suggested both from the quite low Gear value and from the RPM, Speed and Throttle peaks).

Let's then switch to the other factor to confirm our hypothesis, the Speed Limit average value of these trips.

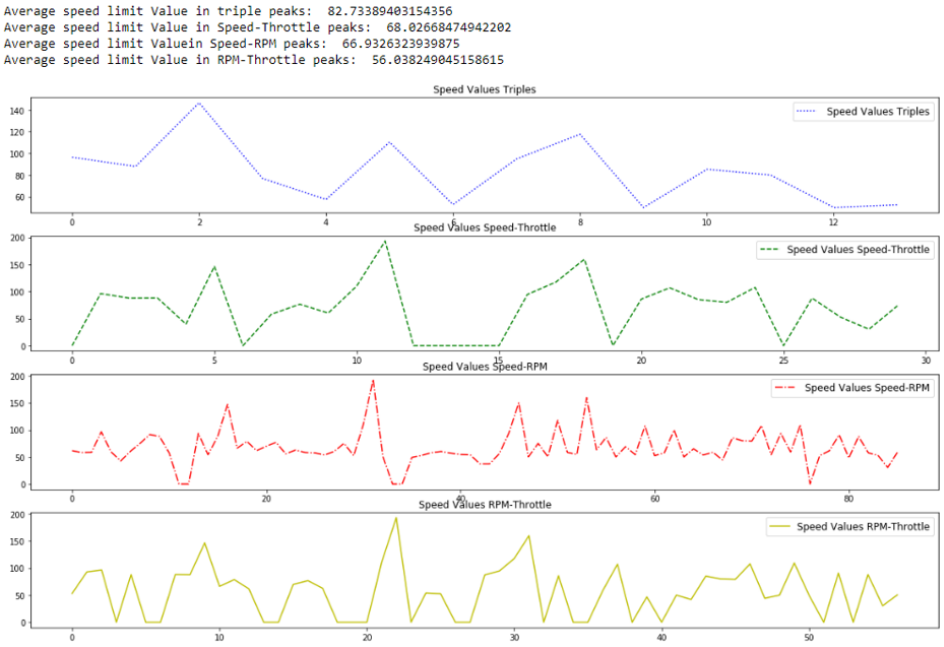


Figure 5.55. Speed Limits analysis for SparkWorks dataset

From these plots and the average values of the speed limit, we can confirm some of the ideas that we have exposed before, in particular the one related to the coincidence of Speed and Throttle peaks regarding the urban-miscellaneous trip type, and the one regarding the coincidence of RPM and Throttle that we have supposed to be an indicator of bad driving style on an urban context.

In the Speed and Throttle peaks coincidence, we can observe how the speed limit average is around 68km/h, it suggests a miscellaneous trip type, in which the pedal is used in a stronger way rather than the highways. At the same time, the fact that the average Speed limit in the trips having high RPM and Throttle peaks

coincidence is about 56 km/h, that is the lowest value among the averages, tells that the drivers have applied pressure on the accelerator pedal or have brought the RPMs to high values in sceneries where the speed limit is not high and there is no need of this behavior.

For what concerns, instead, the Speed Limit average related to the trips having Speed and RPM peaks coincidence, the result does not confirm our supposition of being connected to highway trips because this value is lower than the one of the Speed and Throttle interconnection (66Km/h against 68km/h) that we have assumed to be related to a miscellaneous trip type. However, it neither disproves it, due to a medium-high speed limit.

Finally, in the Triple peaks scenario, we have a real disproof of what we thought. Here the speed limit average is very high, this means that there are many high values (the ones from 90 km/h to 130 km/h) that influence the final average, and this suggests us that the triple peaks can occur in the highway rather than in the urban context, letting us understand that this factor cannot be used as an indicator of bad behavior driving style, due to the possibility of being part of both the highway trips or the bad urban trips.

Before leaving this paragraph there is another element that we should look at since in other studies it can be an additional feature for driver behavior. We are talking about the Heart-Rate factor, the one that could suggest if a driver is upset or if it is calm, or even if it has a high heartbeat due to the high speed. For this reason, the idea is to apply the same concept as before, defining the average heart rate for those trips that were underlined as a probable bad example of driving, and see if these averages change among the different scenarios (5.56).

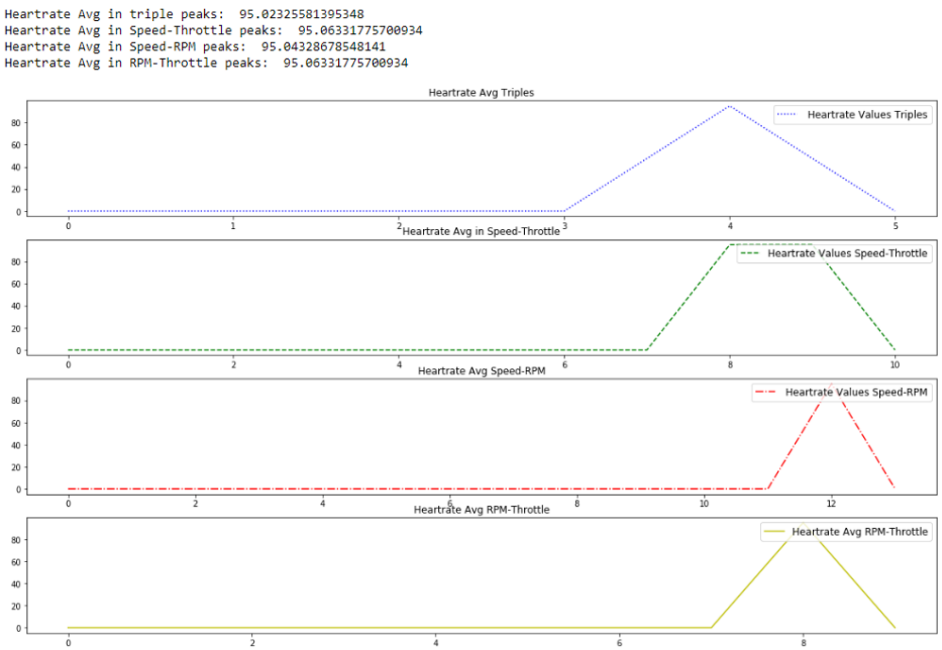


Figure 5.56. Speed Limits analysis for SparkWorks dataset

Unfortunately from both the single trip averages and the plots, that most of

these trips do not present any Heart-Rate information. This means that we don't have any mathematical ground knowledge to say if it could be useful or not, but this idea remains a good starting point for the studies that present a high number of starting trips with homogeneous information.

Chapter 6

Unsupervised approach

The following idea is to try to identify some algorithms that, with all the information that we have, manage to emulate our classification approach. The first thing to focus on is the typology of the approach that we could use. For the general mechanisms that are suitable for our situation, we distinguish two types of approaches, Supervised Learning, and Unsupervised Learning.

In Supervised learning, we train the classification process using data that is already classified. It means that some data is already tagged with the correct answer. So our task is to act like a teacher that gives a knowledge base to the student, that, when will receive another input, on the base of the knowledge that we have brought to him, will produce the correct answer, or in this case, a correct label for the new trip. So, a supervised learning algorithm learns from labeled training data, helping in predicting outcomes for unforeseen data.

Completely different is the Unsupervised learning approach, it is a machine learning technique, in which we don't need to supervise the model. It means that we don't have to bring any classification knowledge to the mechanism, except for the raw data that we would like to label or classify, leaving the model to work on its own to discover information. This means that it mainly deals with unlabeled data.

Once that we have understood the main differences between the two approaches, we have to choose which use in our situation. The answer is quite simple, we don't need a mechanism that already knows which are the bad behavior trips and the good behavior ones, but a mechanism that, given all the scores, number of peaks, number of events, and percentage of good/medium/bad values, manages to classify autonomously the driving style. For this reason, we will search for an appropriate unsupervised algorithm.

The most useful approaches that could be enough strong to emulate our classification process, are the K-Means and the PCA [7]

> "The k-means algorithm defines the centroid of a cluster as the mean value of the points within the cluster. It proceeds as follows. First, it randomly selects k of the objects in D, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the Euclidean distance between the object and the cluster mean.

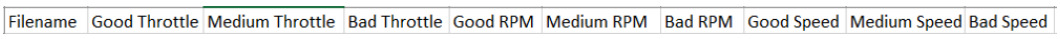


Figure 6.1. Seoul dataset's values used in K-Means (1)

Figure 6.2. Seoul dataset's values used in K-Means (2)

> The k-means algorithm then iteratively improves the within-cluster variation. For each cluster, it computes the new mean using the objects assigned to the cluster in the previous iteration. All the objects are then reassigned using the updated means as the new cluster centers. The iterations continue until the assignment is stable, that is, the clusters formed in the current round are the same as those formed in the previous round."

[16]

> "Principal components analysis (PCA) searches for k n-dimensional orthogonal vectors that can best be used to represent the data, where $k \leq n$. The original data are thus projected onto a much smaller space, resulting in dimensionality reduction. Unlike attribute subset selection, which reduces the attribute set size by retaining a subset of the initial set of attributes, PCA "combines" the essence of attributes by creating an alternative, smaller set of variables. The initial data can then be projected onto this smaller set. PCA often reveals relationships that were not previously suspected and thereby allows interpretations that would not ordinarily result"

[16]

Once that we have explained the mechanisms that we are going to use, we switch to the description of their usage on our information. The first thing was to take all the scores, events numbers, and percentages of good/medium/bad features values of each trip, listed in the 5.4.1. Then, to this information, we have added the set of discoveries related to the analysis of the peak, completing the general overview of the single trip. After this, each trip information was grouped with the ones of the other trips coming from the same dataset, forming a unique data frame for each dataset, composed by a large number of features. We have then applied the k-means mechanism to these dataframes, trying a different number of centroids, to see if the classifications were similar to the one developed in the previous points. The further stage was to use the PCA for reducing the features number to represent the clustering result of the k-means in a two-axis plot.

Let's see in detail the single results for each dataset.

The first dataset that we look at is the **Seoul** one, in this situation each trip presents the sets of values listed in 6.1,6.2,6.3, and6.4 that are used in the k-means classifier.

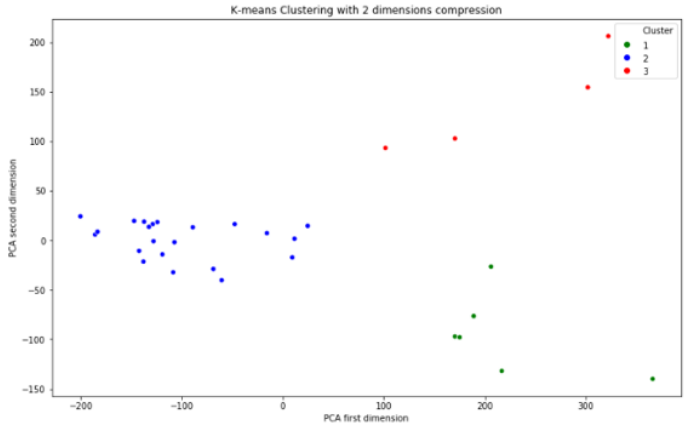
We then start with the selection of three centroids, and this is what we have obtained from the k-means application and the PCA representation with k=3.

Acc X ShiftUP	Acc Y ShiftUP	Acc Z ShiftUP	Aggressiveness	Heartrate average	RPM Ecoscore	ShiftUP Ecoscore	Cruise Ecoscore	Final Ecoscore

Figure 6.3. Seoul dataset's values used in K-Means (3)

# High Speed Peaks	# High RPM Peaks	# High Throttle	# High Triple Peaks	# High Speed-Rpm Peaks	# High Speed-Throttle Peaks	# High Rpm-Throttle Peaks

Figure 6.4. Seoul dataset's values used in K-Means (4)




```
Cluster 1 bad behaviors classified:  66.66666666666666  false positive bad behavior classified: 0
Cluster 2 bad behaviors classified:  22.22222222222222  false positive bad behavior classified: 20
Cluster 3 bad behaviors classified:  11.11111111111111  false positive bad behavior classified: 3
```

Figure 6.5. K-Means classification of Seoul's trips with 3 centroids

The results are not so bad, in fact, if we compare the trips in cluster 1 with the ones selected with our algorithm, we see that the differences are not so heavy (6.7), (6.9).

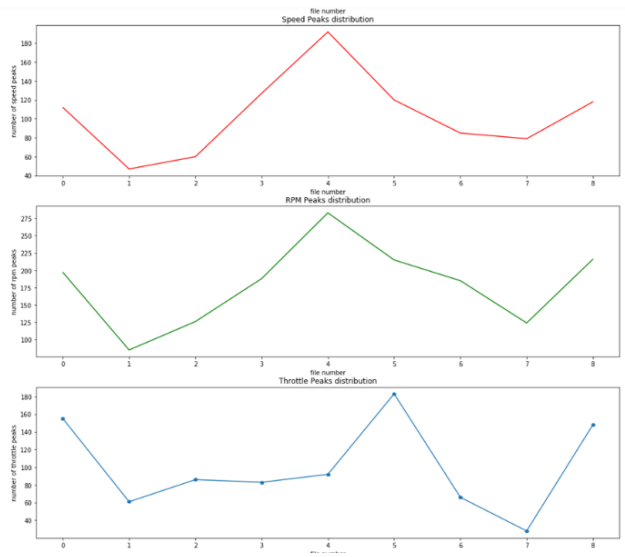


Figure 6.6. Peaks of bad classified trips through general method

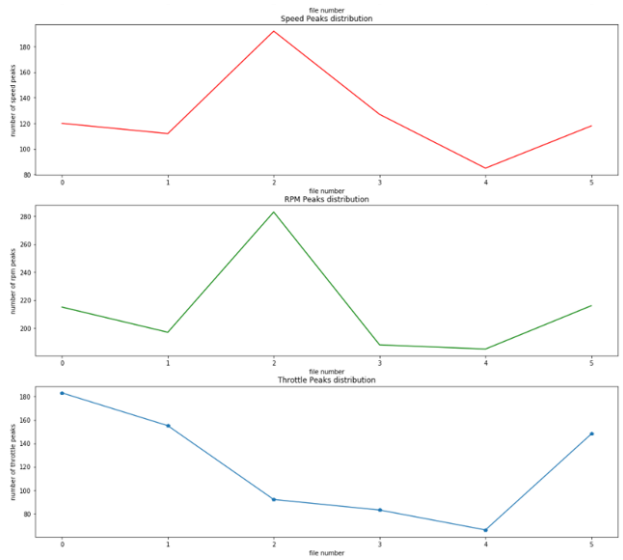


Figure 6.7. Peaks of bad classified trips inserted in cluster 1

The trips that were not inserted in the cluster 1, belonging to the set of bad behavior trips, are most of all trips having not so high number of peaks; this allows us to affirm that the above approach filters some of the false-positive classified from our mechanism.

Switching to the k-means based on two centroids we obtain the following

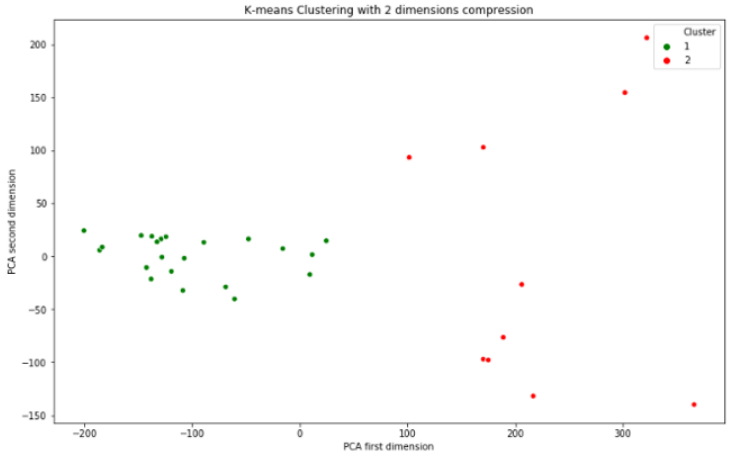


Figure 6.8. K-Means classification of Seoul's trips with 2 centroids

```
Cluster 1 bad behaviors classified:  22.22222222222222 , False positive bad behavior classified: 20
Cluster 2 bad behaviors classified:  77.77777777777779  False positive bad behavior classified: 3
```

Even here, the results could have been worse. Comparing the peaks of the trips collected in cluster 2 with the trips classified with the general approach is clear that the selected trips are very similar to the one of the general approach, even if with the insertion of some errors.

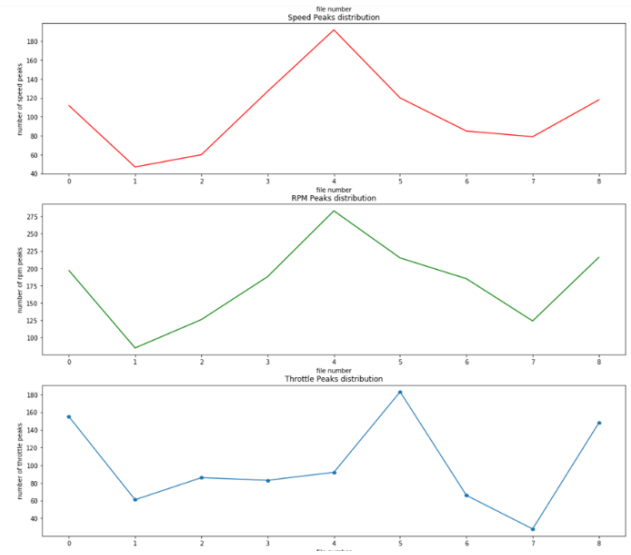


Figure 6.9. Peaks of bad classified trips through general method

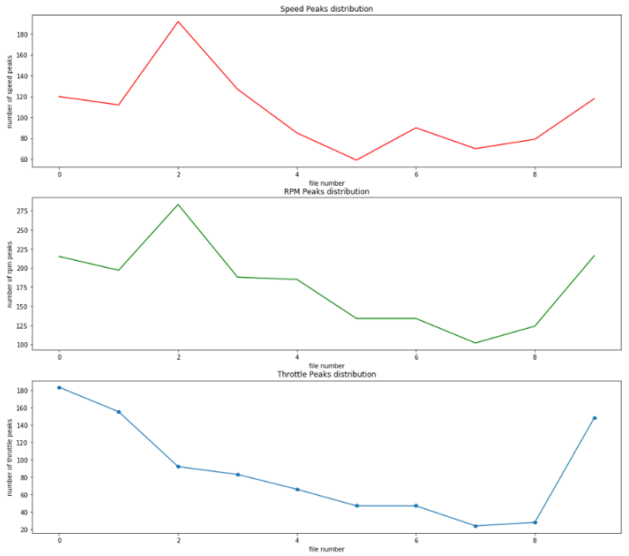
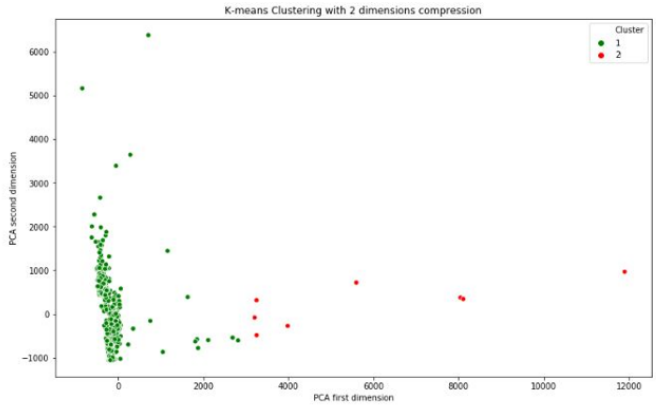


Figure 6.10. Peaks of bad classified trips inserted in cluster 2

So, for this dataset, we can say that the k-means with 3 centroids, based on all the scores and information collected by us, is a good alternative to our algorithm because the missing trips were not the strongly bad ones.

Let's now switch to the **SparkWorks dataset** where we have applied k-means and the PCA, basing on the same set of variables listed before, with a different number of centroids, from 2 to 8. The results, nevertheless, are not so satisfying because, due to the high number of trips, when a cluster contains a high number of bad behavior trips (following the classification done with our algorithm) it also contains a high number of wrong classified trips.

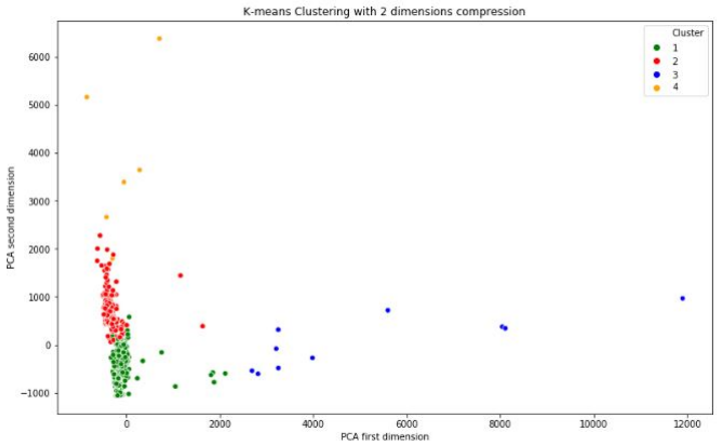
- 2 Clusters




```
Cluster 1 bad behaviors classified:  97.67441860465115 , False positive bad behavior classified: 574
Cluster 2 bad behaviors classified:  2.3255813953488373  False positive bad behavior classified: 17
```

Figure 6.11. K-Means classification of SparkWorks dataset's trips with 2 centroids

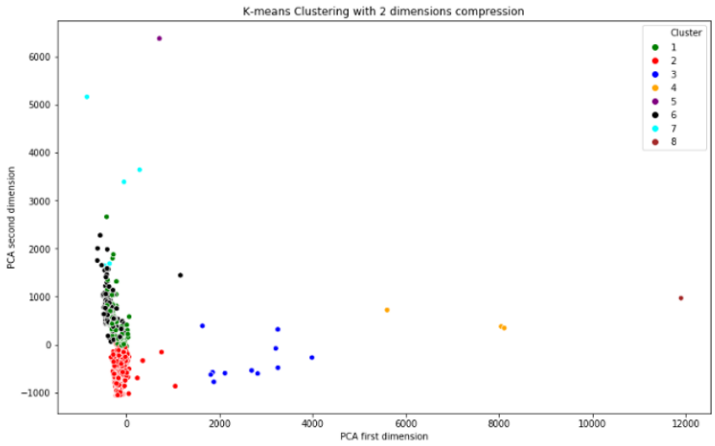
- 4 Clusters



```
Cluster 1 bad behaviors classified:  58.139534883720934 , False positive bad behavior classified: 319
Cluster 2 bad behaviors classified:  30.23255813953488  False positive bad behavior classified: 251
Cluster 3 bad behaviors classified:  4.651162790697675  False positive bad behavior classified: 18
Cluster 4 bad behaviors classified:  6.976744186046512  False positive bad behavior classified: 3
```

Figure 6.12. K-Means classification of SparkWorks dataset's trips with 4 centroids

- 8 Clusters




```
Cluster 1 bad behaviors classified:  18.6046511627907 , False positive bad behavior classified: 155
Cluster 2 bad behaviors classified:  48.837209302325576  False positive bad behavior classified: 293
Cluster 3 bad behaviors classified:  9.30232558139535  False positive bad behavior classified: 14
Cluster 4 bad behaviors classified:  0.0  False positive bad behavior classified: 12
Cluster 5 bad behaviors classified:  2.3255813953488373  False positive bad behavior classified: 0
Cluster 6 bad behaviors classified:  13.953488372093023  False positive bad behavior classified: 114
Cluster 7 bad behaviors classified:  6.976744186046512  False positive bad behavior classified: 2
Cluster 8 bad behaviors classified:  0.0  False positive bad behavior classified: 1
```

Figure 6.13. K-Means classification of SparkWorks dataset's trips with 8 centroids

And this leads to a non-similar classification approach from the k-means one and general one. An example of it is shown from the number of speed peaks of the two classifications (6.14),(6.15).

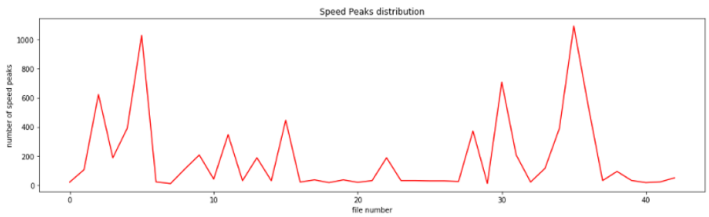


Figure 6.14. Speed peaks number of trips classified as nervous from general method

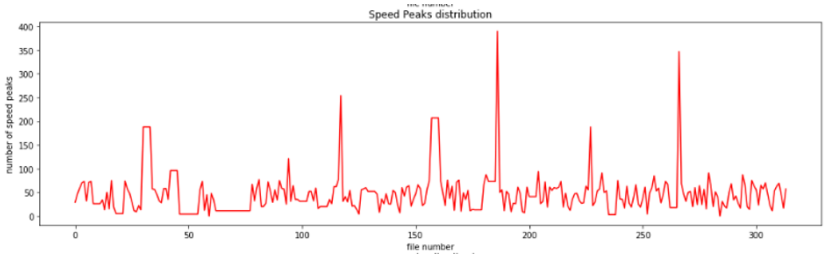
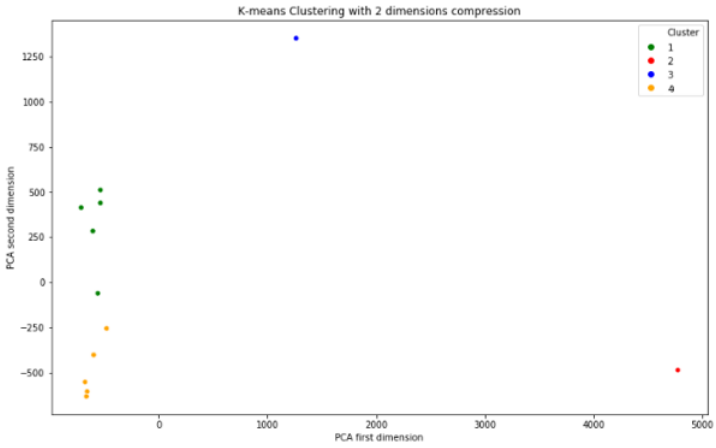


Figure 6.15. Speed peaks number of trips classified as nervous from K-Means with k=8 (Cluster 2)

This means that the couple k-means/PCA is not so effective for the SparkWork dataset, proposing a completely different result from the one seen in the Seoul scenery.

Finally, we move to the **Cephas dataset**, both the sections related to 14 drivers and 19 drivers. The first one that we take into account is the **14 drivers** subset, however, also here the classification does not bring the desired results. Both using three, four, five or eight clusters the percentage of trips correctly classified is low, with a high number of false positives.

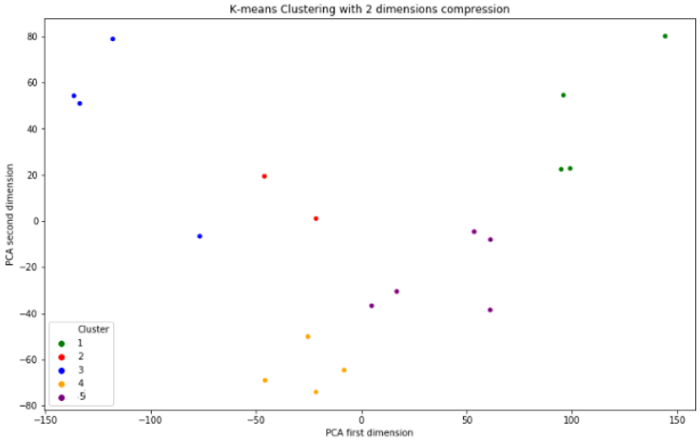



```
Cluster 1 bad behaviors classified:  25.0 , False positive bad behavior classified: 4
Cluster 2 bad behaviors classified:  25.0  False positive bad behavior classified: 0
Cluster 3 bad behaviors classified:  25.0  False positive bad behavior classified: 0
Cluster 4 bad behaviors classified:  25.0  False positive bad behavior classified: 4
```

Figure 6.16. K-Means classification of Cephas14 dataset's trips with 5 centroids

As we were saying the clusterization achieved is not satisfying, because the bad trips are split into many clusters, and the cluster themselves contains also other trips that are false positives for the bad behavior classification.

For what concerns the section of the **19 drivers**, instead, we have a partially good classification of half of the bad behavior trips with no other false positives, but the other bad behavior trips are clustered in a wrong way.




```
Cluster 1 bad behaviors classified:  12.5 , False positive bad behavior classified: 3
Cluster 2 bad behaviors classified:  0.0  False positive bad behavior classified: 2
Cluster 3 bad behaviors classified:  50.0  False positive bad behavior classified: 0
Cluster 4 bad behaviors classified:  12.5  False positive bad behavior classified: 3
Cluster 5 bad behaviors classified:  25.0  False positive bad behavior classified: 3
```

Figure 6.17. K-Means classification of Cephas19 dataset's trips with 5 centroids

Concluding this section, we can say that the k-means with the PCA are a good alternative only for some situations, but not in general. This underlines the strength of our classification mechanism for driver behaviors.

Chapter 7

Classification through features' subset

Until now we have managed to individuate the main features and the secondary features that help us to classify the drivers' behavior during their trips. This was possible also through the use of some thresholds, scores, and thanks to the analysis of the peak of these features for the trips that have been considered. Now, our work is to split up the results obtained to understand with which approximation precision we could still classify the drivers' behavior, but using only the elements coming from a smartphone or obtainable from an OBD II (On-Board Diagnostic) device.

7.1 Smartphone Features

First of all, we have to list the set of smartphone tools that could be useful for our aim. If we consider a quite recent smartphone, we can say that in it are certainly present some tools and sensors like:

- **Gyroscope**

  A three-axis gyroscope determines if your device is twisted in any direction. Using rotational force, it measures angular velocity around three axes. The absolute orientation of your phone, represented as the angles yaw, pitch, and roll, is detected by a combination of the accelerometer, compass, and gyroscope.

- **Thermostat**

- **GPS Location**

  GPS is another type of sensor in your mobile device. It relies on satellites to determine location. Originally developed for the military, GPS was made available for everyone in the 1980s

- **Air Humidity Sensor**

- **Heart Rate Sensor**

- **Barometer**

More advanced smartphones have a chip that can detect atmospheric pressure. But to use it, the phone needs to pull down local weather data for a baseline figure on barometric pressure. What's more, conditions inside a building, such as heating or air-conditioning flows, can affect the sensor's accuracy.

– **Magnetometer / Compass**

It detects the direction of magnetic north and, in conjunction with GPS, determines the user's location

– **Accelerometer**

A three-axis accelerometer in your smartphone reports on how fast your phone is moving in any given linear direction. The accelerometer can detect gravity as a static acceleration as well as dynamic acceleration applied to the phone. There are various types of MEMS accelerometer hardware available, such as microscopic piezoelectric crystals that change voltage under stress when vibrations occur, or differential capacitance caused by the movement of a silicon structure. The magnetometer, GPS, gyroscope, and accelerometer on your phone all work together to create the perfect navigation system.

– **Ambient Light Sensor**

It detects data about lighting levels in the environment to adapt the display accordingly.

And many others that are not useful for our purpose. Starting from this selection of tools and sensors we want to look up which of these can be used for obtaining the dataset features that we have used in our various steps; to do this, we have to consider dataset per dataset.

For what concern the **Seoul dataset** we briefly recall the features that we have focused on during the study, putting beside them the features that are usable in a context based only on smartphone's elements.

Feature in the Dataset	Used until now	Usable working only with smartphone
Fuel_consumption	✓	✗
Accelerator_Pedal_value	✓	✗
Throttle_position_signal	✓	✗
Intake_air_pressure	✓	✗
Absolute_throttle_position	✓	✗
Engine_speed	✓	✗
Engine_coolant_temperature	✗	✗
Current_Gear	✓	✗
Single_Wheel_velocity	✓	✗
Vehicle_speed	✓	✓
Acceleration_speed_-_Longitudinal	✓	✓
Indication_of_brake_switch_ON/OFF	✗	✗
Acceleration_speed_-_Lateral	✓	✓
Steering_wheel_speed	✓	✓
Steering_wheel_angle	✓	✓
Time(s)	✓	✓

These, instead, are the scores that were useful for the trip classification.

Scores and Thresholds	Used until now	Usable working only with smartphone
Aggressiveness Score	✓	✗
Eco-Score (RPM Eco-Score)	✓	✗
Eco-Score (Cruising Eco-Score)	✗	✗
Eco-Score (ShiftUp Eco-Score)	✗	✗
Gear ShiftUp Score	✓	✗
Speed thresholds	✓	✓
RPM Thresholds	✓	✗
Throttle Thresholds	✓	✗
Acceleration (Event) Thresholds	✓	✓
Steering Wheel Events	✓	✓

For this reason, we have to understand that there is a high possibility to develop a worst classification with respect to the one done in the general context, but to defeat any doubt and to have an idea of the classification level we have to analyze this classification based only on this subset of factors/scores.

The results, in this case, are very interesting.

```
Number of files classified with general features and score
 Good Behavior:  15      Bad Behavior:  9


Number of files classified with OBD features and score
 Good Behavior:  7       Bad Behavior:  12
Common classified files with General Method
 Good Behavior:  46.666666666666664 %    Bad Behavior:  88.88888888888889 %
```

Figure 7.1. Comparison of Seoul trip classification's quality between the general features and the Smartphone's ones

Basing only on these factors and scores, we see that the number of trips that have been classified as good behavior trips is not so precise (less than 50%), while for the ones classified as bad behavior driving style is very precise, around the 90% selected by the general approach.

Another interesting thing is that the missing trips of the bad behavior classification are the ones that are borderline between the good and the bad classification. This means that this alternative does not introduce a hard error of classification

even if the classifiers' number is lower. We then demonstrate graphically the results, proposing the remaining trips of the general set in the two situations (7.2,7.3,7.4).

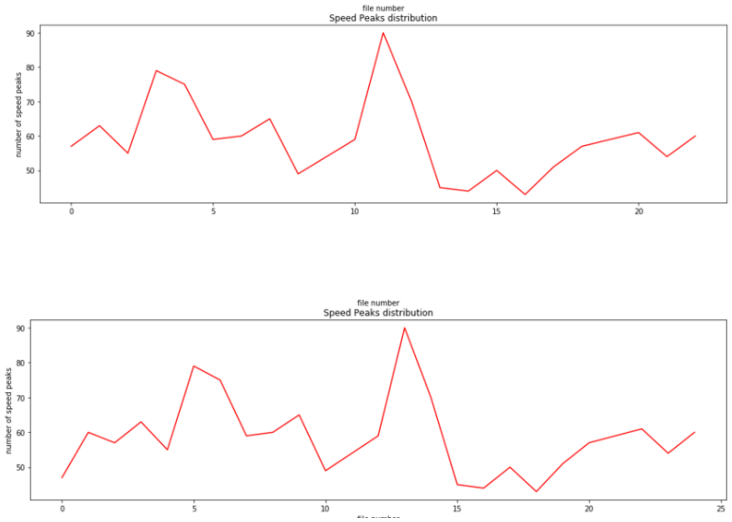


Figure 7.2. Speed peaks number of bad classified trips through general method (higher image) versus smartphone method (lower image)

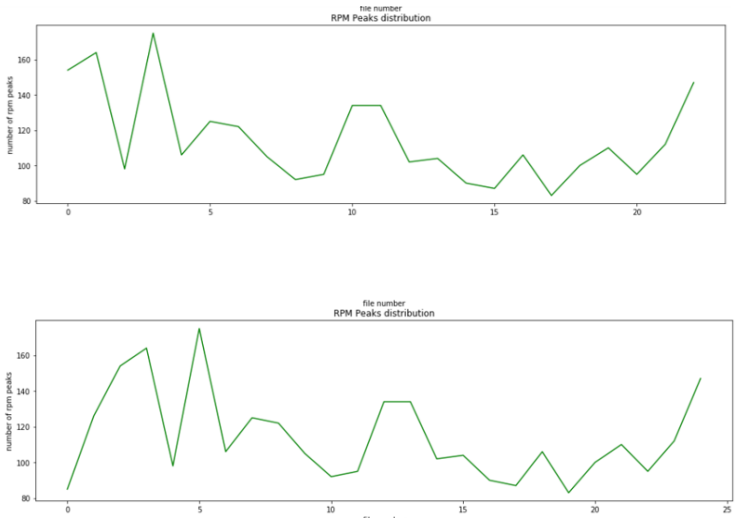


Figure 7.3. RPM peaks number of bad classified trips through general method (higher image) versus smartphone method (lower image)

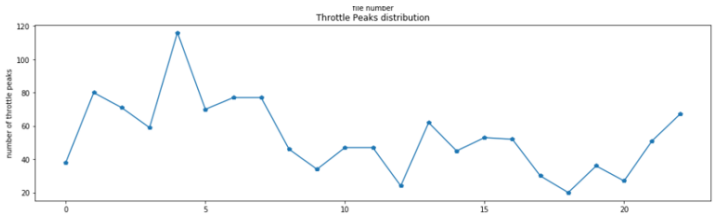
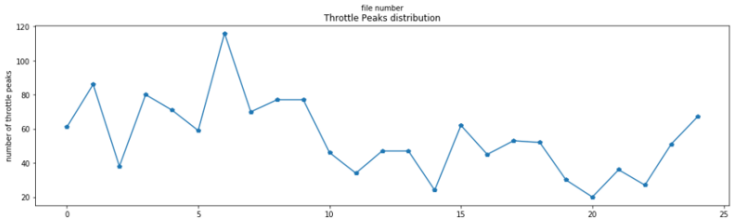


Figure 7.4. Throttle peaks number of bad classified trips through general method (higher image) versus smartphone method (lower image)



As we can see in all the images, the additional trips do not have a high number of high-value peaks, but in general, we can affirm that this classification based strongly on acceleration events, steering wheel events and vehicle speed, is a good one. We must say that if we have the lack of the other factor, like Throttle or RPM, the verification is not appliable because we won't have all their peaks information that we need to understand if the classified trip is a false positive or a right classified one.

For what concerns the dataset of Cephas Barreto, in particular the **Cephas19**, even the general classification on this data was not powerful, due to the lack of some important scores and secondary features.

However, if we reduce the classification of the trips basing only on the features related to the smartphone (in this case only on Speed Score) we obtain that the classification of the Bad behavior trips has an accuracy of the 50%, while the accuracy of Good behavior trips is 0%.

```
Number of files classified with general features and score
 Good Behavior:  5       Bad Behavior:  8

Number of files classified with Smartphone features and score
 Good Behavior:  0       Bad Behavior:  4
Common classified files with General Method
 Good Behavior:  0.0 %   Bad Behavior:  50.0
```

Figure 7.5. Comparison of Cephas19 trip classification's quality between the general features and the Smartphone's ones

For a better understanding of this section see the 7.3 in which are presented bar plots and Venn diagram of the various classification sets.

Regarding instead the **Cephas14** the situation of the general classification is a little bit different (see the section 5.4.1), but the results of the "smartphone classification" are the same, because we don't have information regarding the steering wheel angle, the steering wheel speed, or the acceleration. This means that only basing on the vehicle speed we cannot classify the behaviors, even more, because we don't have any data related to the traversed road, and so, related to the traversed road speed limit.

Let's then look at the scenario of the **SparkWorks dataset**. As before we have to recap the features and the scores used in the general classification to develop a drivers' trip clustering, comparing them with the elements that are still appliable in the smartphone one.

Feature in the Dataset	**Used until now**	**Usable working only with smartphone**
Fuel Consumption Rate	✗	✗
Accelerator X	✓	✓
Accelerator Y	✓	✓
Accelerator Z	✓	✓
Throttle position	✓	✗
Heart-Rate	✓	✓
Road Type	✓	✗
Traffic Confidence	✓	✓
GPS Position (Latitude and Longitude)	✓	✓
Gear	✓	✗
Engine RPM	✓	✗
Vehicle Speed	✓	✓
Air Fuel Ratio	✓	✗
Air intake temperature	✓	✗
Humidity	✗	✓
Pressure	✗	✓
Weather	✗	✓
Timestamp	✓	✓

For what concerns the scores, here are the differences in the approaches.

Scores and Thresholds	**Used until now**	**Usable working only with smartphone**
Aggressiveness Score	✓	✗
Eco-Score (RPM Eco-Score)	✓	✗
Eco-Score (Cruising Eco-Score)	✓	✗
Eco-Score (ShiftUp Eco-Score)	✓	✗
Gear ShiftUp Score (Acceleration X)	✓	✗
Gear ShiftUp Score (Acceleration Y)	✓	✗
Gear ShiftUp Score (Acceleration Z)	✓	✗
Speed thresholds	✓	✓
RPM Thresholds	✓	✗
Throttle Thresholds	✓	✗
Acceleration X (Event) Thresholds	✓	✓
Acceleration Y(Event) Thresholds	✓	✓
Acceleration Z(Event) Thresholds	✓	✓

So, the situation is that we have a lot of features that are usable with only the smartphone (even more with respect to the previous approaches), but, at the same time, many of them are not reachable only with the phone's sensors that are fundamental for calculating a lot of scores.

For what concerns of the classification comparison the results are explained in 7.6.

```
Number of files classified with general features and score
 Good Behavior:  176     Bad Behavior:  40

Number of files classified with Smartphone features and score
 Good Behavior:  359     Bad Behavior:  22
Common classified files with General Method
 Good Behavior:  74.43181818181817 %     Bad Behavior:  47.5 %
```

Figure 7.6. Comparison of SparkWorks dataset trip classification's quality between the general features and the smartphone's ones

In particular, we can see that the percentage of common classified trips is not so low (around 50%), the real problem is the number of missing trips that were not classified as bad.

The lack of some scores does not allow us to distinguish a quite big group of bad behavior trips properly.

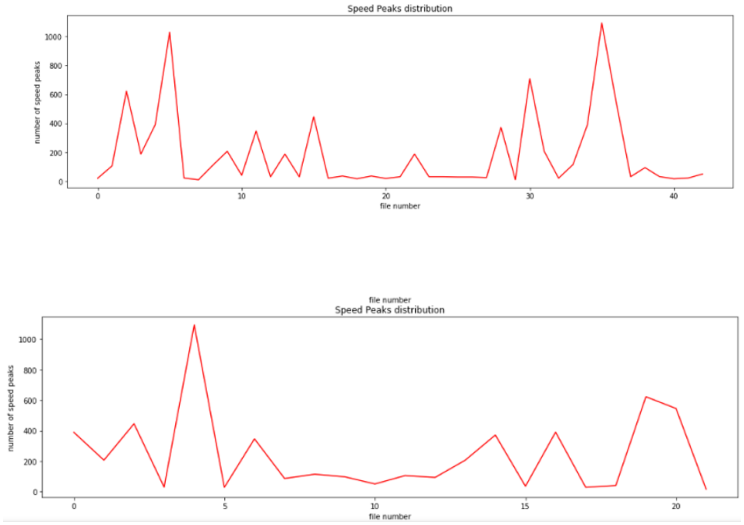
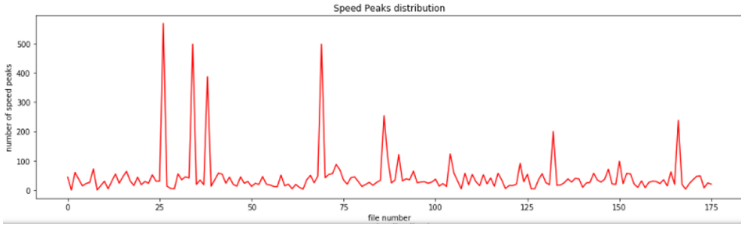


Figure 7.7. Speed peaks number of bad classified trips through general method (higher image) versus smartphone method (lower image) for SparkWorks dataset

We have also the problem of the bigger number of good behavior trips classified. These paths that were categorized as calm trips are instead false positive. They should be inserted in the set of bad behavior trips, and this can be seen with plots 7.8, where it is obvious that, with respect to the general approach, the trips with way higher peaks are inserted in the wrong set.



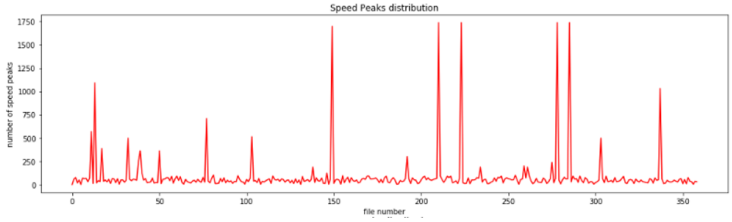


Figure 7.8. Speed peaks number of good classified trips through general method (higher image) versus smartphone method (lower image) for SparkWorks dataset

The output of this analysis leaves interesting food for thought. Firstly, we have seen how some factors and some of the scores were not so essential as we thought, even among the main ones. The deletion of them if accompanied with other features, produce (in particular in the example of the Seoul dataset) a good classification rate, leaving aside only some false positive identified through the general method. Secondly, we have seen that the most serious error in this classification approach is the one related to the false positives.

In the SparkWorks dataset many of the main scores and factors were discarded leaving the place to a small set of classifiers, this leads to many errors, so the main concept coming out from this study tells us that is possible to classify a driver without some important features, but they have to be replaced with other strong elements, like for example the steering wheel information and the accelerations events.

7.2 OBD Features

For what regards the opposite case, where we want to classify a trip only basing ourselves on the signal obtained from the car that can be read through the OBD II, we have a very different situation. The On-Board Diagnostic (OBD), in fact, allow us to obtain several information like:

FUEL_STATUS	Fuel System Status	**THROTTLE_POS**	Throttle Position
ENGINE_LOAD	Calculated Engine Load	**COOLANT_TEMP**	Engine Coolant Temperature
SHORT_FUEL_TRIM_1	Short Term Fuel Trim - Bank 1	**FUEL_PRESSURE**	Fuel Pressure
LONG_FUEL_TRIM_1	Long Term Fuel Trim - Bank 1	**RPM**	Engine RPM
SHORT_FUEL_TRIM_2	Short Term Fuel Trim - Bank 2	**SPEED**	Vehicle Speed
LONG_FUEL_TRIM_2	Long Term Fuel Trim - Bank 2	**TIMING_ADVANCE**	Timing Advance
INTAKE_PRESSURE	Intake Manifold Pressure	**INTAKE_TEMP**	Intake Air Temp
MAF	Air Flow Rate (MAF)	**AIR_STATUS**	Secondary Air Status
DISTANCE_W_MIL	Distance Traveled with MIL on	**FUEL_RAIL_PRESSURE_VAC**	Fuel Rail Pressure (relative to vacuum)
FUEL_RAIL_PRESSURE_DIRECT	Fuel Rail Pressure (direct inject)	**COMMANDED_EGR**	Commanded EGR
EGR_ERROR	EGR Error	**EVAPORATIVE_PURGE**	Commanded Evaporative Purge
WARMUPS_SINCE_DTC_CLEAR	Number of warm-ups since codes cleared	**DISTANCE_SINCE_DTC_CLEAR**	Distance traveled since codes cleared
EVAP_VAPOR_PRESSURE	Evaporative system vapor pressure	**BAROMETRIC_PRESSURE**	Barometric Pressure
COMMANDED_EQUIV_RATIO	Commanded equivalence ratio	**RELATIVE_THROTTLE_POS**	Relative throttle position
AMBIANT_AIR_TEMP	Ambient air temperature	**ACCELERATOR_POS_D**	Accelerator pedal position D
THROTTLE_ACTUATOR	Commanded throttle actuator	**TIME_SINCE_DTC_CLEARED**	Time since trouble codes cleared
FUEL_LEVEL	Fuel Level Input	**ABSOLUTE_LOAD**	Absolute load value
MAX_MAF	Maximum value for mass air flow sensor	**FUEL_TYPE**	Fuel Type
ETHANOL_PERCENT	Ethanol Fuel Percent	**EVAP_VAPOR_PRESSURE_ABS**	Absolute Evap system Vapor Pressure
EVAP_VAPOR_PRESSURE_ALT	Evap system vapor pressure	**FUEL_RAIL_PRESSURE_ABS**	Fuel rail pressure (absolute)
RELATIVE_ACCEL_POS	Relative accelerator pedal position	**OIL_TEMP**	Engine oil temperature
RUN_TIME	FUEL_RATE	**FUEL_INJECT_TIMING**	Fuel injection timing

This means that among all the datasets, and the features related to them, only

```
Number of files classified with general features and score
 Good Behavior:  15      Bad Behavior:  9

Number of files classified with OBD features and score
 Good Behavior:  7       Bad Behavior:  12
Common classified files with General Method
 Good Behavior:  46.666666666666664 %    Bad Behavior:  88.88888888888889 %
```

Figure 7.9. Comparison of Seoul trip classification's quality between the general features and the OBD's ones

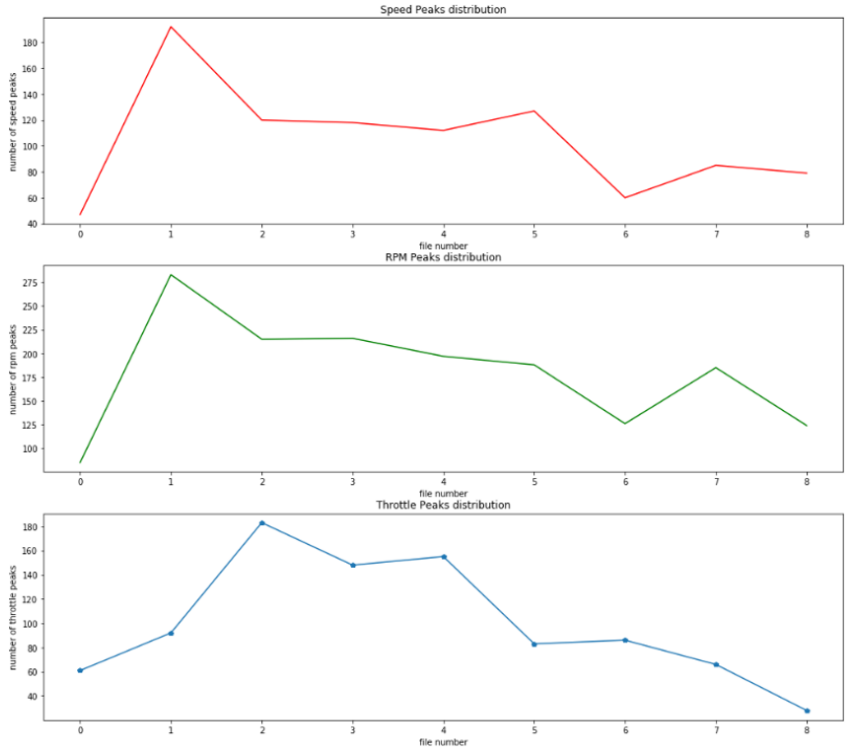


Figure 7.10. Number of peaks for Seoul dataset's trips classified as bad from the general method

a small subset is obtainable only from the smartphone. They are in particular the information regarding the angulation and speed of the steering wheel, the GPS Location, the Speed Limits of a precise road, and the weather.

So, let's see how the classification changes deleting the above-mentioned elements.

Starting from the **Seoul dataset** we can understand, from the plots 7.9 results, that the majority of the trips classified as bad, only basing on OBD features, are very similar to the general one.

The only thing that we have to say is that in this situation are introduced some trips that era labeled as erroneous, but their peaks are not so high, so we can affirm that these additional classifications of trip introduce false positives with a quite good margin of error.

Another proof of what we are saying can be noticed in the 7.10 and 7.11 plots, where the addition of these trips does not bring elements having a high number of

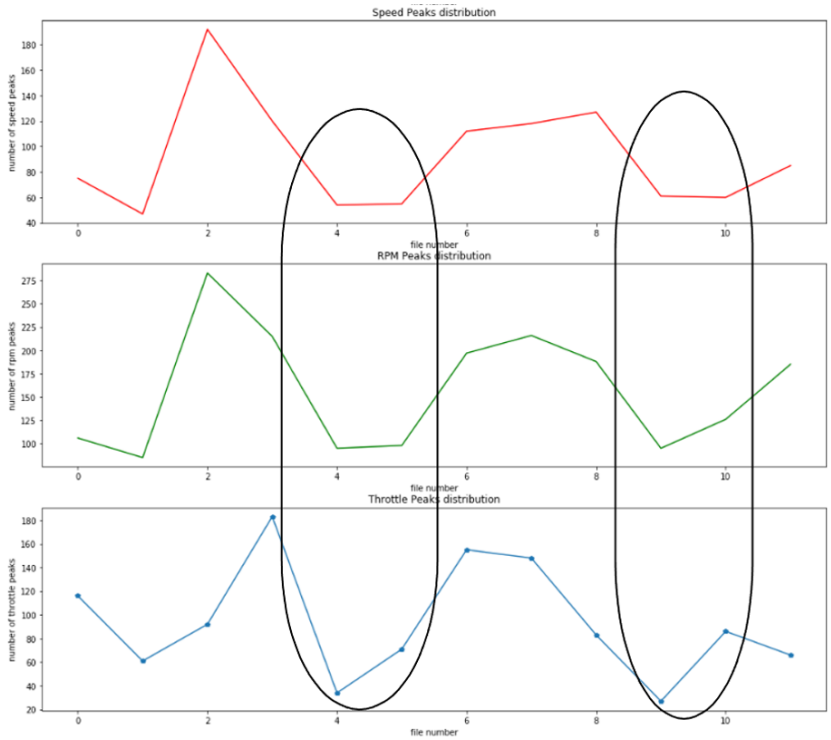


Figure 7.11. Number of peaks for Seoul dataset's trips classified as bad from the OBD method

peaks. This means that the usage of smartphone tools near the OBD one is not useless, they help to avoid e delete the false positive bad behavior classification based only on the car's Central Unit queried through the OBD commands.

Now is the turn of the **SparkWorks dataset**.

```
Number of files classified with general features and score
 Good Behavior:  176     Bad Behavior:  40


Number of files classified with OBD features and score
 Good Behavior:  186     Bad Behavior:  24
Common classified files with General Method
 Good Behavior:  100.0 %         Bad Behavior:  60.0 %
```

Figure 7.12. Comparison of SparkWorks dataset trip classification's quality between the general features and the OBD's ones

Looking at the result coming out from the "OBD approach" immediately raises a question regarding the bad behavior cluster; if the trips that were not classified as bad behavior driving are truly not bad. To answer this let's see the differences through the plots 7.13 and 7.14.

As we can notice, the missing trips that were not classified in the second scenario were truly bad behavior trips. So, this approach, in which we have not the comparison

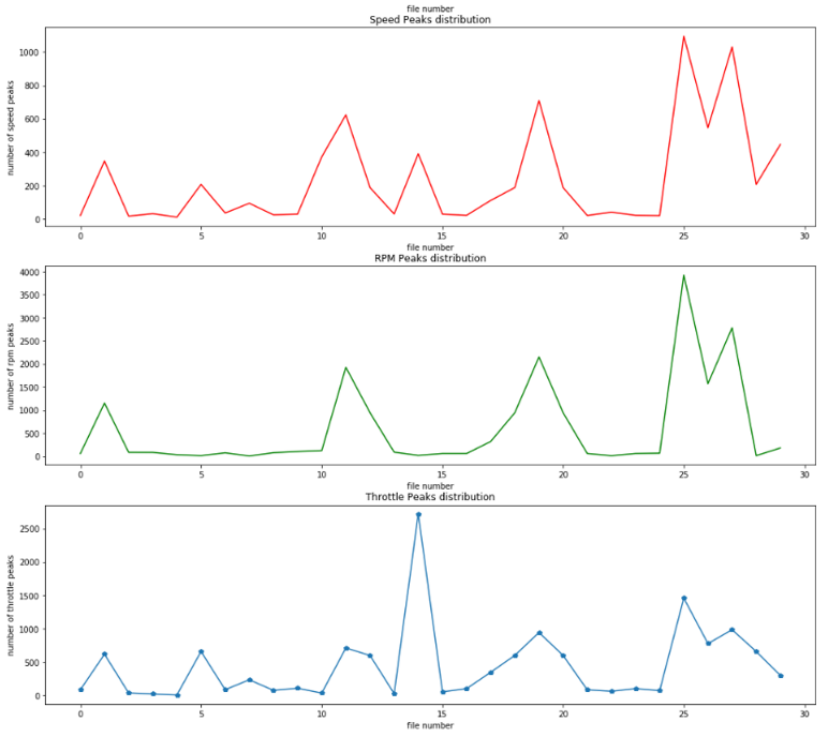


Figure 7.13. Number of peaks for SparkWorks dataset's trips classified as bad from the general method

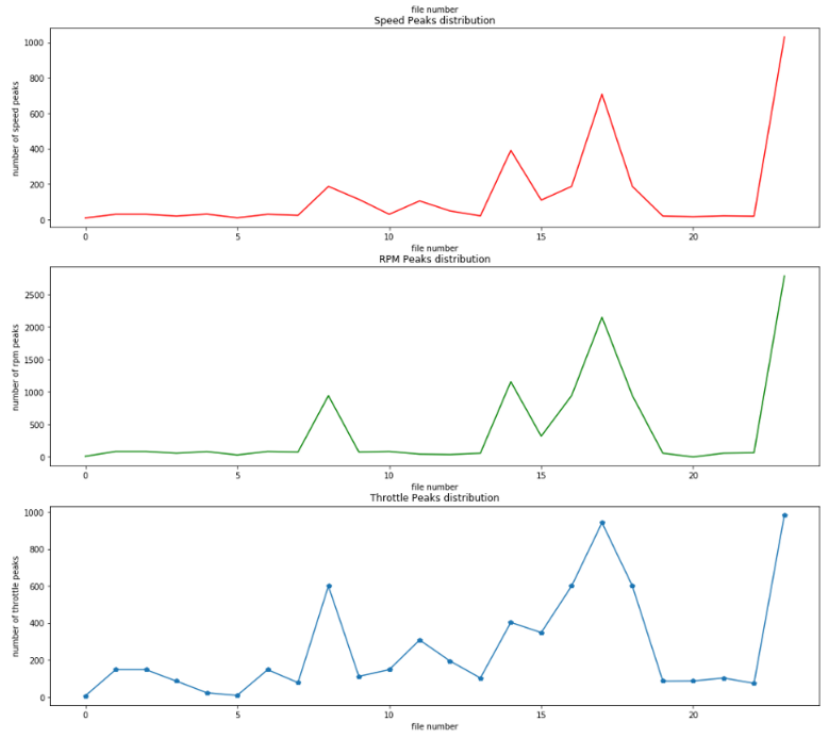


Figure 7.14. Number of peaks for SparkWorks dataset's trips classified as bad from the OBD method

of instant speed with the road speed limit, is worse than we thought, and it is not recommended.

Finally, we put under examination the Cephas dataset, in particular the **Cephas19**, in which we have a perfect classification of the "Good behavior" trips and a quite good classification of the "Bad Trips" with some missing element 7.15.

```
Number of files classified with general features and score
 Good Behavior:  5       Bad Behavior:  8

Number of files classified with OBD features and score
 Good Behavior:  5       Bad Behavior:  5
Common classified files with General Method
 Good Behavior:  100.0 %         Bad Behavior:  62.5
```

Figure 7.15. Comparison of Cephas19 dataset trip classification's quality between the general features and the OBD's ones

This means that a classification made up only through OBD features proposes a good alternative to the general method, but obviously, it is not perfect, due to the insertion of a small number of erroneous trip in the wrong set, and due to the lack of some important trips in the classification sets (like for the SparkWorks dataset's bad behavior trips).

7.3 Comparison with entire features' set

In this final section, we introduce the different graphical representations of the classification results obtained from the general method, the method based on the smartphone features, and the subset obtained from the OBD features.

For each dataset, we will firstly recap the elements taken into account for the various approaches, and then two types of plots.

- **Bar plot**

  In which we can see the number of trips that are classified as "Bad behavior trip" and "Good behavior trip"

- **Venn diagram**

  To understand the number of trips that are commonly selected by the different approaches

The first dataset that we recap is the **SparkWorks one**, here we have based the general trips' classification on the following features:

- Speed Score
- Aggressiveness Score
- Final Eco-score
- Acceleration Events on X-axis

- Gear ShiftUp related to Acceleration on X-axis

For the subset of Smartphone features, we have reduced them to

- Speed Score
- Acceleration Events

And for the OBD's features subset we have selected

- Aggressiveness Score
- Final Eco-score
- Acceleration Events on X-axis
- Gear ShiftUp related to Acceleration on X-axis

The results are the following; for the "Bad behavior classification", we can affirm that most of the smartphone classification bad trips are correctly classified, as the OBD classification one.

It can be noticed from the orange bar, that represents the number of a commonly classified bad trip between the general approach and the smartphone one, and from the red bar that represents the number of commonly classified bad trips between the general approach and the OBD one.

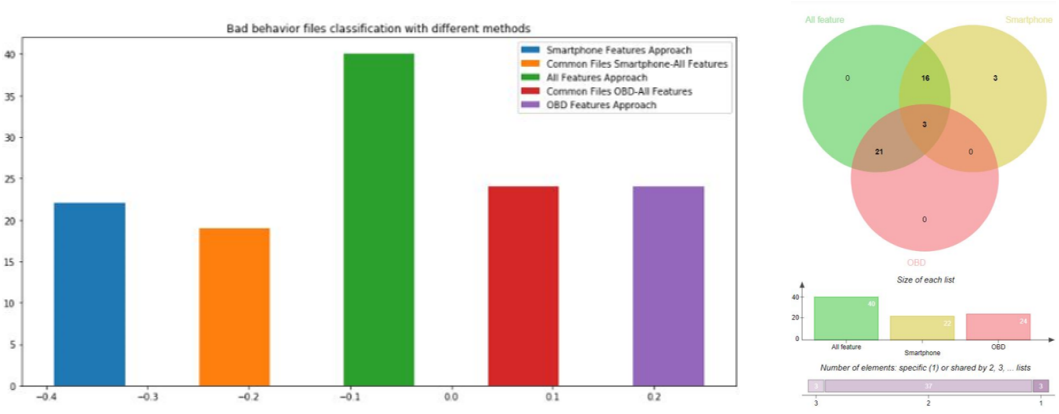


Figure 7.16. Bars plot and Venn plot of the different SparkWorks dataset's trips classification methods - **Nervous driving group**

For the "Good behavior classification", instead, there is a huge number of trips that are wrong classified as good from the smartphone approach, while the result coming from the OBD approach is quite the same as the general one.

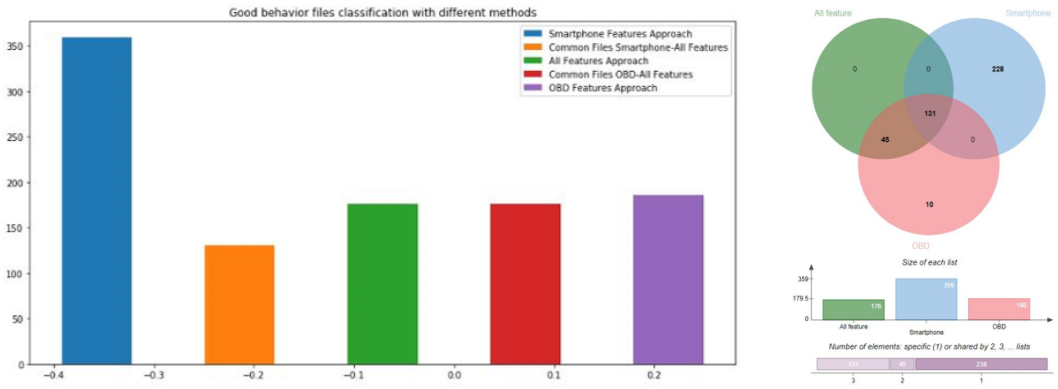


Figure 7.17. Bars plot and Venn plot of the different SparkWorks dataset's trips classification methods - **Calm driving group**

Let's now switch to the Cephas dataset. Here we won't look at the different results from the Cephas14 due to the impossibility of application of the smartphone approach. Instead, for what concerns **Cephas19** some things that are good to be observed.

In it, the general classification approach is based on

- Aggressiveness Score
- Speed Score

This means that for the Smartphone approach we base ourselves only on

- Speed Score

And for the OBD classification approach on

- Aggressiveness Score

What came out from the "Bad behavior trip" classification, as we could think, since we use only two factors, one for each type of classification, is that the general result is perfectly made up from the sum of the two partial classifications.

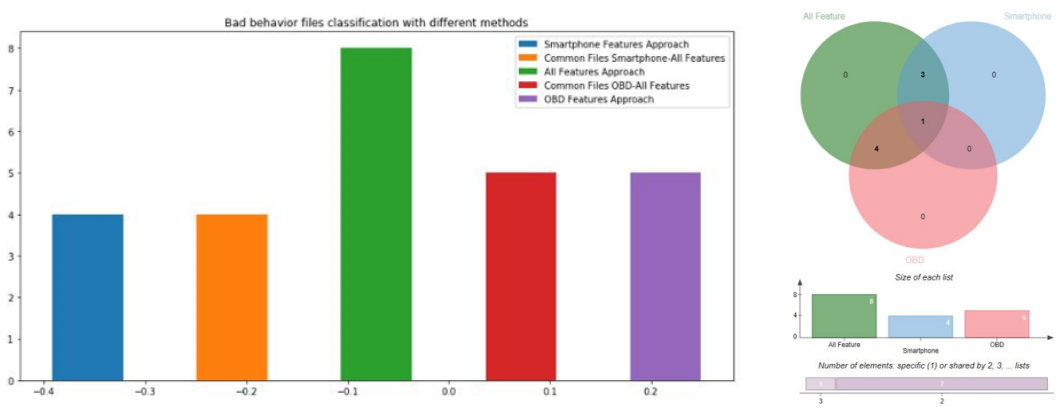


Figure 7.18. Bars plot and Venn plot of the different Cephas19 dataset's trips classification methods - **Nervous driving group**

Different, instead, is the situation in the "Good behavior trip" classification's result. Here, the Smartphone approach is completely useless because it doesn't find any trip of the final set represented in the general approach.

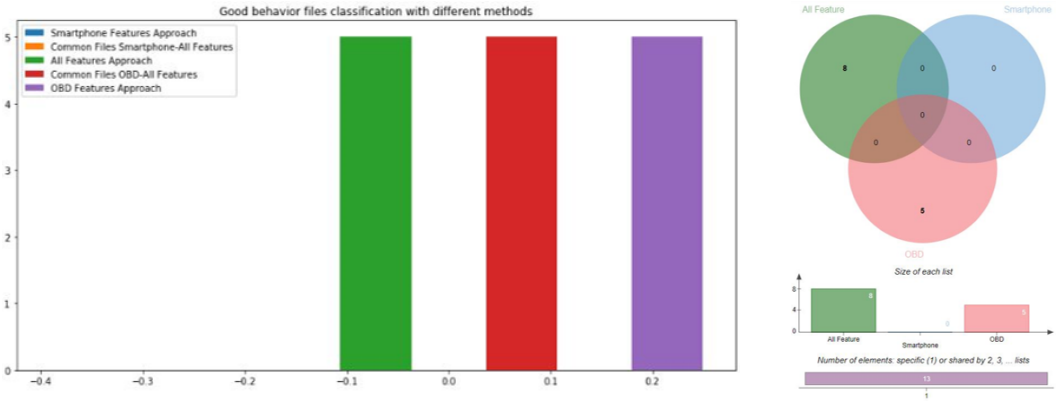


Figure 7.19. Bars plot and Venn plot of the different Cephas19 dataset's trips classification methods - **Calm driving group**

The last dataset is the **Seoul** one, in which the general approach works on

- Eco-score
- Aggressiveness Score
- Steering Wheel Events
- Lateral Acceleration Events

Where the first two are used in the OBD classification approach, and the others in the Smartphone approach. However, in this scenario, we have applied another distinction. In the smartphone's set, we divide its overall features result from the

one in which the Steering Wheel event are not considered, this to understand which of the two smartphone's features has more classification strength against the other.

Firstly, we look at the results where the smartphone approach utilizes all its features. In the "Bad behavior trips" we notice a good Smartphone classification's demeanor with no false positive, but a quite bad demeanor for the OBD classification with some false positives.

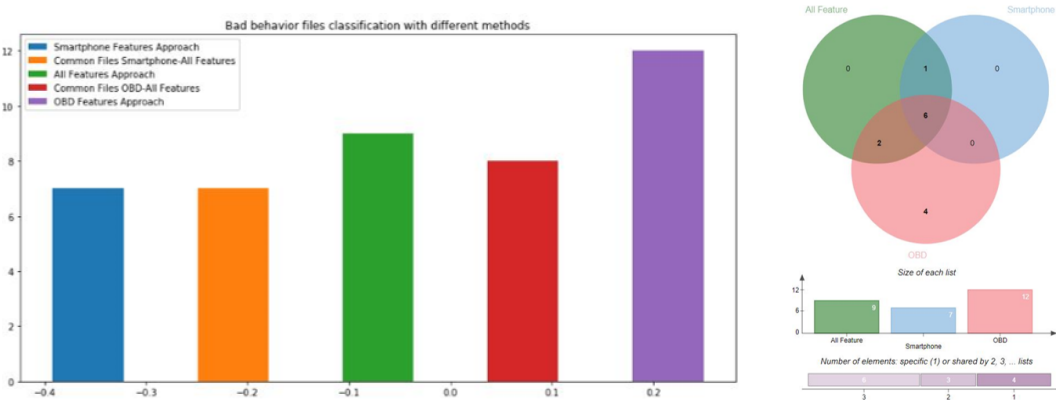


Figure 7.20. Bars plot and Venn plot of the different Seoul dataset's trips classification methods - **Nervous driving group**

Regarding the "Good behavior trips" classification, we have the specular scenario, in which the Smartphone approach introduces some false positives, while the OBD returns a good demeanor.

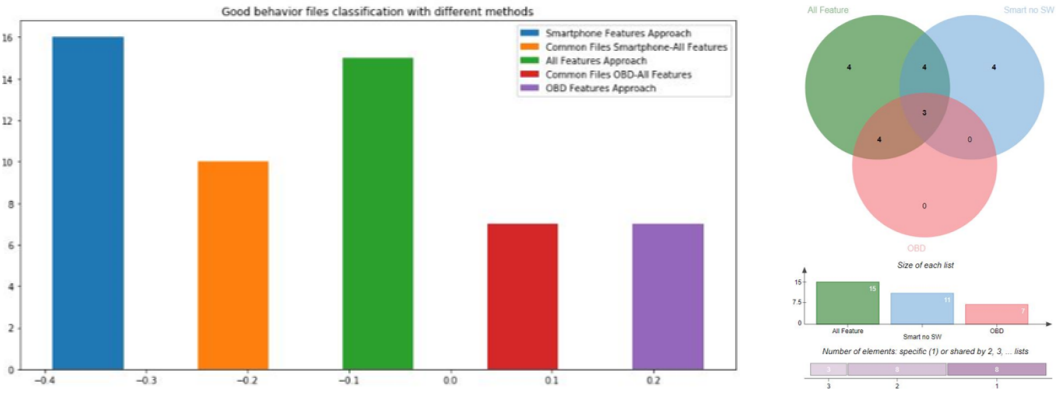


Figure 7.21. Bars plot and Venn plot of the different Seoul dataset's trips classification methods - **Calm driving group**

Secondly, the classification results in which there are no steering wheel events.

In particular, the "Bad behavior trip" results introduce a quite high classification error in the OBD approach.

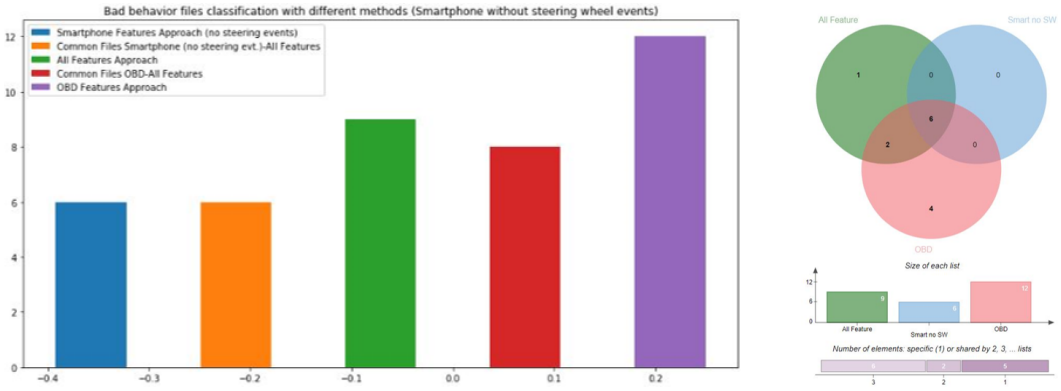


Figure 7.22. Bars plot and Venn plot of the different Seoul dataset's trips classification methods (no steering events in Smartphone) - **Nervous driving group**

While the "Good behavior trip" results introduce a quite high Smartphone approach classification error.

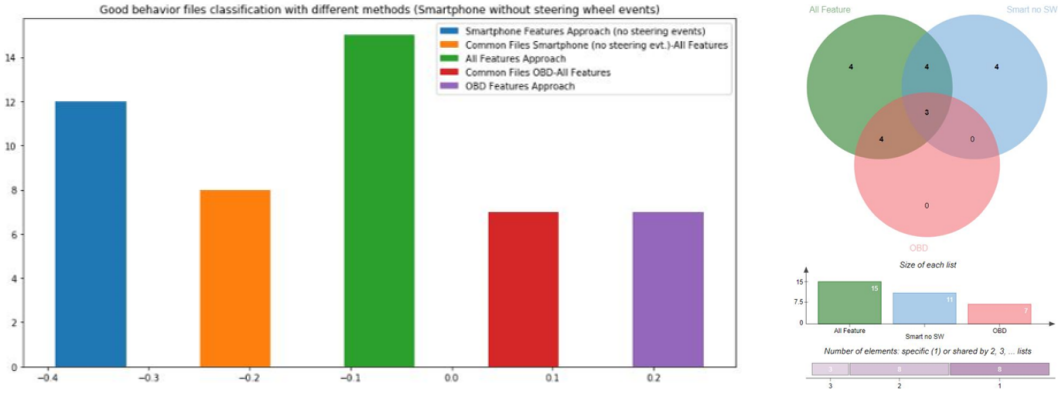


Figure 7.23. Bars plot and Venn plot of the different Seoul dataset's trips classification methods (no steering events in Smartphone) - **Calm driving group**

In the end, we present the Venn representation with both the analysis done with and without the Steering wheel events.

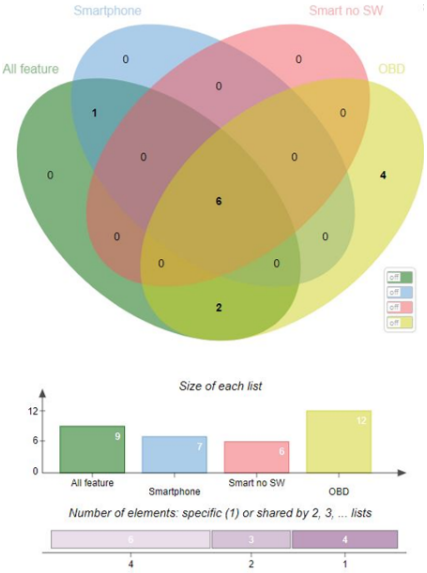


Figure 7.24. Venn plot of the different Seoul dataset's trips classification methods - **Nervous driving group**

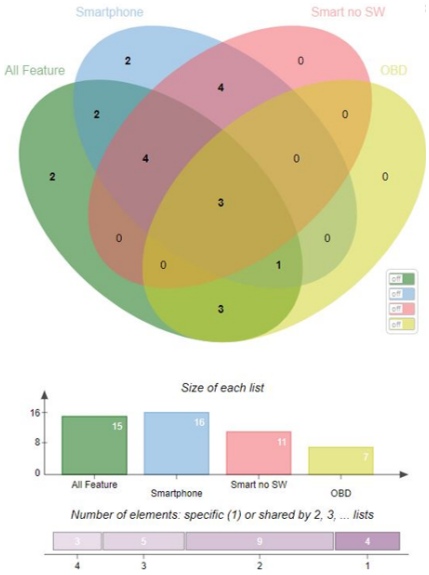


Figure 7.25. Venn plot of the different Seoul dataset's trips classification methods - **Calm driving group**

Chapter 8

Conclusions and future work

It is clear that, with the identification of the features that most characterize the road driving methods, it is possible to outline what the drivers' behaviors are when they are, behind the steering wheel of their vehicles. The result developed gradually, finding initially the most influencing elements of a given driving style through the correlations based on averages, variances and standard deviations of the values and on their entire vectors.

The same approach was then carried out concerning what could have been the supporting elements of the primary features. This, to switch in the use of thresholds, and scores both introduced in different studies, and developed ad hoc for the different datasets. The evolution of this algorithm took a substantial form when the peaks component related to the primary features was introduced. The study takes into account the relative maximums of the various routes, concerning certain thresholds, developed by the drivers. It has allowed a great leap forward in the efficiency of our approach because it was possible to have a graphical confirmation of the classifications previously made, which were different from dataset to dataset due to the diversity of the information contained in them. It manages to confirm the success of this clustering, also indicating those that could be the false positives or false negatives introduced by this classification.

This analysis of the peaks was deeply addressed, as there was the assumption that the results could change in particular situations, such as the insertion of paths much longer than the general average of all the samples. This has been dispelled through various simulations and subdivisions of trips into ad hoc created subsets. It has been discovered that the study of the coincidence peaks' number related to the main features gives us an idea of the type of route that the driver has made, allowing us to distinguish the stretch of the road traveled between urban, highway or miscellaneous. Later we test a method widely used for clustering in an unsupervised environment such as the K-means combined with the PCA, to understand if our algorithm could be emulated by it. The result, however, states that the clustering carried out by the K-means offers partially interesting results only under some tight conditions and datasets. Finally, an attempt was made to understand which of the components strictly coming from smartphones and those related to the control unit of the vehicles had more strength in identifying a specific driving behavior.

All this work has brought great results in the distinction between calm driv-

ing behaviors and more impetuous behaviors, allowing us to understand how the mechanisms inserted in modern vehicles can still be helped by external elements for understanding when the different human moods changes in agitated states.

In any case, this work can be easily improved, going to provide more complete sets of paths and data, with all those characteristics that in our case were divided into the different compartments and therefore they could not be used in common as an overall dataset.

In addition to this, a next job could be the combination with a smartphone application interacting with the vehicle's control unit via Bluetooth communication through devices like the OBD, creating special structures inside the vehicle to make the most from all the various smartphone's sensor. An example of this could be a belt with smartphone housing positioned on the steering wheel of the vehicle that can detect the severity of the steering turn and the steering's angles with high precision, or heart sensors placed on the steering wheel that can send these signals to the smartphone, already connected through this belt on the steering wheel.

Bibliography

[1] Dimitrios Amaxilatis, Christos Tselios, Orestis Akrivopoulos, and Ioannis Chatzigiannakis. On the design of a fog computing-based, driving behaviour monitoring framework. In *2019 IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pages 1–6. IEEE, 2019.

[2] C Antoniou, V Papathanasopoulou, V Gikas, C Danezis, and H Perakis. Classification of driving characteristics using smartphone sensor data. In *Proc. Symp. European Association for Research in Transport*, 2014.

[3] Cephas Alves da Silveira Barreto. Uso de técnicas de aprendizado de máquina para identificação de perfis de uso de automóveis baseado em dados automotivos. Master's thesis, Brasil, 2018.

[4] Juan Carmona, Fernando García, David Martín, Arturo de la Escalera, and José María Armingol. Data fusion for driver behaviour analysis. *Sensors*, 15(10):25968–25991, 2015.

[5] German Castignani, Raphaël Frank, and Thomas Engel. An evaluation study of driver profiling fuzzy algorithms using smartphones. In *2013 21st IEEE International Conference on Network Protocols (ICNP)*, pages 1–6. IEEE, 2013.

[6] Oussama Derbel et al. Driving style assessment based on the gps data and fuzzy inference systems. In *2015 IEEE 12th International Multi-Conference on Systems, Signals & Devices (SSD15)*, pages 1–8. IEEE, 2015.

[7] Umberto Fugiglando, Emanuele Massaro, Paolo Santi, Sebastiano Milardo, Kacem Abida, Rainer Stahlmann, Florian Netter, and Carlo Ratti. Driving behavior analysis through can bus data in an uncontrolled environment. *IEEE Transactions on Intelligent Transportation Systems*, 20(2):737–748, 2018.

[8] Narelle Haworth and Mark Symmons. The relationship between fuel economy and safety outcomes. 2001.

[9] V Heijne, N Ligterink, and U. Stelwagen. Potential of eco-driving. udrive deliverable 45.1. *EU FP7 Project UDRIVE Consortium*, 2017.

[10] Eleonora Bressan (KITE), Gabrio Mauri (KITE), Aristotelis-Spathis Papadiotis (UPAT), Stavros Nousias (UPAT), Dimitris Kalabalikis (UPAT), and Konstantinos Moustakas (UPAT). D4.3 – online analysis of data. Technical report, H2020 GamECAR Grant 732068, 2018.

[11] Eleonora Bressan (KITE), Gabrio Mauri (KITE), Aristotelis-Spathis Papadiotis (UPAT), Stavros Nousias (UPAT), Dimitris Kalabalikis (UPAT), and Konstantinos Moustakas (UPAT). D4.4 – gamecar decision support system. Technical report, H2020 GamECAR Grant 732068, 2018.

[12] Eleonora Bressan (KITE), Gabrio Mauri (KITE), Aristotelis-Spathis Papadiotis (UPAT), Stavros Nousias (UPAT), and Dimitris Kalabalikis (UPAT)and Konstantinos Moustakas (UPAT). D7.3 – field trials report and socio economic guidelines. Technical report, H2020 GamECAR Grant 732068, 2018.

[13] Fabio Martinelli, Francesco Mercaldo, Albina Orlando, Vittoria Nardone, Antonella Santone, and Arun Kumar Sangaiah. Human behavior characterization for driving style recognition in vehicle system. *Computers & Electrical Engineering*, 2018.

[14] Rana Massoud, Francesco Bellotti, Riccardo Berta, Alessandro De Gloria, and Stefan Poslad. Eco-driving profiling and behavioral shifts using iot vehicular sensors combined with serious games. In *2019 IEEE Conference on Games (CoG)*, pages 1–8. IEEE, 2019.

[15] G Meers and M Roth. Road safety and ecological sustainability working together. In *AUSTRALASIAN TRANSPORT RESEARCH FORUM (ATRF), 24TH, 2001, HOBART, TASMANIA, AUSTRALIA*, 2001.

[16] What Is Data Mining. Data mining: Concepts and techniques. *Morgan Kaufinann*, 10:559–569, 2006.

[17] Stavros Nousias, Christos Tselios, Dimitris Bitzas, Dimitrios Amaxilatis, Javier Montesa, Aris S Lalos, Konstantinos Moustakas, and Ioannis Chatzigiannakis. Exploiting gamification to improve eco-driving behaviour: The gamecar approach. *Electr. Notes Theor. Comput. Sci.*, 343:103–116, 2019.

[18] Stavros Nousias, Christos Tselios, Dimitris Bitzas, Aris S Lalos, Konstantinos Moustakas, and Ioannis Chatzigiannakis. Uncertainty management for wearable iot wristband sensors using laplacian-based matrix completion. In *2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pages 1–6. IEEE, 2018.

[19] Stavros Nousiasl, Christos Tseliosl, Olivier Orfila, Samantha Jamson, Pablo Mejuto, Dimitrios Amaxilatis, Orestis Akrivopoulos, Ioannis Chatzigiannakis, Aris S Lalosl, Konstantinos Moustakasl, et al. Managing nonuniformities and uncertainties in vehicle-oriented sensor data over next generation networks. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 272–277. IEEE, 2018.

[20] Yen-Jen Chen Shu-LinHwang. Combining obd technology with acceleration sensor to analyze aggressive driving behavior. *American Journal of Engineering Research (AJER)*, 7:139–144, 2018.

[21] John Thomas, Shean Huff, Brian West, and Paul Chambon. Fuel consumption sensitivity of conventional and hybrid electric light-duty gasoline vehicles to driving style. *SAE International Journal of Fuels and Lubricants*, 10(3):672–689, 2017.

[22] Christos Tselios, Stavros Nousias, Dimitris Bitzas, Dimitrios Amaxilatis, Orestis Akrivopoulos, Aris S Lalos, Konstantinos Moustakas, and Ioannis Chatzigiannakis. Enhancing an eco-driving gamification platform through wearable and vehicle sensor data integration. In *European Conference on Ambient Intelligence*, pages 344–349. Springer, 2019.

[23] E Tzirakis, F Zannikos, and S Stournas. Impact of driving style on fuel consumption and exhaust emissions: defensive and aggressive driving style. In *Proceedings of the 10th International Conference on Environmental Science and Technology*, pages 1497–1504. Global Network for Environmental Science and Technology (Global-NEST), 2007.

[24] Minh Van Ly, Sujitha Martin, and Mohan M Trivedi. Driver classification and driving style recognition using inertial sensors. In *2013 IEEE Intelligent Vehicles Symposium (IV)*, pages 1040–1045. IEEE, 2013.

[25] Zhuowen Wang, Fuqiang Liu, Xinhong Wang, and Yuyan Du. Driver modeling based on vehicular sensing data. In *2018 International Conference on Advanced Control, Automation and Artificial Intelligence (ACAAI 2018)*. Atlantis Press, 2018.

[26] Wikipedia contributors. Kendall rank correlation coefficient — Wikipedia, the free encyclopedia, 2020.

[27] Wikipedia contributors. Pearson correlation coefficient — Wikipedia, the free encyclopedia, 2020.

[28] Wikipedia contributors. Spearman's rank correlation coefficient — Wikipedia, the free encyclopedia, 2020.