The use of both regression tree and mixed ANOVA techniques in the interpretation of repeated speed measures belonging to an-on road experiment

Pasquale Colonna, Paolo Intini, Nicola Berloco, Filomena Mauriello, Antonio Perruccio and Vittorio Ranieri

Abstract— A data set composed by repeated speed measurements belonging to an on-road experiment was inquired into this study by using the regression tree and the ANOVA mixed model techniques. The on-road experiment consisted in the repetition of a driving test six times in six different days for a sample of 20 drivers on a rural road in the province of Bari, Italy. In order to explain the changes in the response variable speed, the drivers' behavioral differences, the road geometry (in terms of sight distance) and the time (in terms of the six days of testing) were used as variables. All the three chosen variables are responsible for improvements in the tree growing and they have significant effects on speed according to the mixed one-way ANOVA. Some considerations about the results of the two analyses were given in this paper, together with an analysis of the interactions between the variables based on them. In more detail, some interesting differences between the behavior of aggressive drivers and the behavior of prudent drivers were highlighted.

Keywords— Driving Behavior, Mixed one-way ANOVA, On-road Experiment, Regression Trees, Route Familiarity, Sight Distance, Speed Measures.

I. INTRODUCTION

Driving behavior is universally accepted as a potential factor able to influence the occurrence of road accidents.

However, driving behavior is not characterized by a universally accepted theory, because of the various factors involved in the process [1]. For example, the zero-risk model [2], the risk homeostasis theory [3], the rule-based model [4], the risk allostasis theory [1] and/or the risk monitor model [5] could be taken into account.

Speed choice is one of the main indicators of driver behavior and it is influenced in turn by many factors, among which risk perception is crucial [6]. The way in which users perceive accident risk while they are driving is a topic currently studied, a perplexing topic due to the lack of consensus about measuring risk and users' risk misperceptions [7].

One influential feature in drivers' behavior can be identified in their familiarity with a route, determined by their habit of driving on it. In fact, there is some research about the relationships between route familiarity and driving performance. Yanko and Spalek [8] e.g. carried out an experiment involving 20 drivers and a driving simulator. They found that route familiar users (users who had driven on the experimental route four times before the test) needed greater reaction times than route unfamiliar users (users who drove on the experimental route for the first time during the test) in order to respond to unexpected external stimuli simulated in the presented scenarios. The results obtained from the presented experiment are similar to what Martens and Fox [9] suggest about route familiarity: it can lead to a greater distraction while driving, probably because familiarity could increase the effect of "mind wandering". Mind wandering occurs when the mind is occupied by thoughts not concerning the task being undertaken and so, responses to external stimuli are potentially slowed down. This interpretation is coherent with the MART theory presented by Young and Stanton [10], which assumes that driving performance varies as a function of mental workload and that in low demand conditions (normal driving tasks) attention capacity is reduced. However, those studies are based on driving simulator experiences. Instead, in this study, route familiarity was inquired by considering data belonging to an on-road study. The investigation of speed behavior based on a real world setting has the advantage of producing data with the greatest validity in comparison with those obtained in a simulated scenario (see e.g. [11]).

Pasquale Colonna is with the Department of Civil, Environmental, Building Engineering and Chemistry, Technical University of Bari, Bari, 70127, Italy (corresponding author, phone: +39337832927; fax: +390805963329; e-mail: pasquale.colonna@poliba.it).

Paolo Intini is with the Department of Civil, Environmental, Building Engineering and Chemistry, Technical University of Bari, Bari, 70127, Italy (corresponding author, phone: +393466978279; fax: +390805963329; e-mail: paolo.intini@poliba.it).

Nicola Berloco is with the Department of Civil, Environmental, Building Engineering and Chemistry, Technical University of Bari, Bari, 70127, Italy (e-mail: nicola.berloco@poliba.it).

Filomena Mauriello is with the Department of Civil, Architectural and Environmental Engineering, University of Naples "Federico II", Naples, 80125, Italy (e-mail: filomena.mauriello@unina.it).

Antonio Perruccio is with the Department of Civil, Environmental, Building Engineering and Chemistry, Technical University of Bari, Bari, 70127, Italy (e-mail: antonio.perruccio@poliba.it).

Vittorio Ranieri is with the Department of Civil, Environmental, Building Engineering and Chemistry, Technical University of Bari, Bari, 70127, Italy (e-mail: vittorio.ranieri@poliba.it).

In more detail, the interpretation of speed data was based on the use of the technique of classification and regression trees (CART) and on the use of the mixed one-way ANOVA. The influence of behavioral differences, road geometry, familiarity and the interactions between those factors on speed were inquired.

The interpretation of speed data will be made by considering the objective of explaining speed variations over time due to the acquired road familiarity. Moreover, the influence of road geometry and human factors (behavioral differences between drivers) on this process will be shown.

Moreover, a comparison between the results obtained by employing the two different techniques was performed, in order to assess the main differences between them in the interpretation of speed data.

The remainder of the paper summarizes the methods employed for the on-road experiment, the data obtained, the methods employed for data analysis (section II – Materials and Methods) and the presentation and discussion of results (section III – Results and Discussion).

In more detail, some important features such as the variations in speed over the days of testing will be related to route familiarity and controlled for other variables as a result of the analysis.

II. Materials and Methods

A. The on-road study

An on-road study was planned and realized by the Technical University of Bari with the aim of understanding how memory of drivers can influence speed choice [12], [13], [14]. To inquire into this relationship, a sample of drivers were recruited among students of the university by using advertisements.

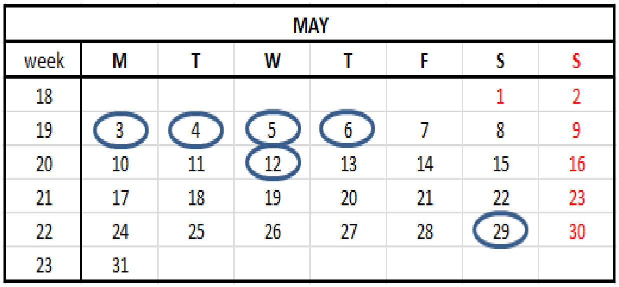
Drivers with less than 22 years, with less than 3 years of driving license, who declared to experience a mileage of less than 10 km per week on rural roads, who have not an available car and who already known the test route were discharged from the sample.

Therefore, the final test sample was composed by 20 drivers, characterized by the following features; age: 24.45 ± 1.10 years old, 16 males and 4 females, years licensed: 5.75 ± 1.25 years.

They travelled on a given route, on which they never travelled before, six times in six different days, according to the following fixed chronological schedule.

The first four tests were scheduled in four consecutive days (1st, 2nd, 3rd and 4th days of testing). The other two tests were fixed in the ninth day after the first test (5th day of testing) and in the twenty-sixth day after the first test (6th day of testing).

Two stretches of two-lane two-way rural roads (SP31 and SP18, situated in the municipality of Cassano delle Murge, district of Bari, Italy) were chosen as driving test routes. Those stretches were selected in order to ensure free flow traffic conditions during all tests.



MAY							
week	M	T	W	T	F	S	S
18						1	2
19	[✓ 3]	[✓ 4]	[✓ 5]	[✓ 6]	7	8	9
20	10	11	[✓ 12]	13	14	15	16
21	17	18	19	20	21	22	23
22	24	25	26	27	28	[✓ 29]	30
23	31						

Fig. 1 example of a driving test schedule for all users belonging to the sample

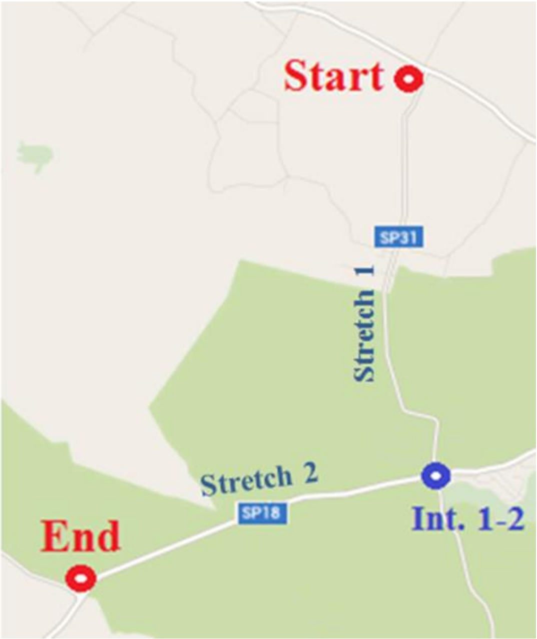


Fig. 2 layout of the test driving routes

All users selected for the driving test used the car that they were driving usually. In fact, drivers, even if experienced, could modify their driving behavior by using another car for the first time.

Speed data were collected by using the Differential Positioning GPS technology (Dynamic Method). They were acquired by using a frequency of a measurement per second. Speed profiles were drawn for each driver by putting distances on the x-axis and measured speeds on the y-axis.

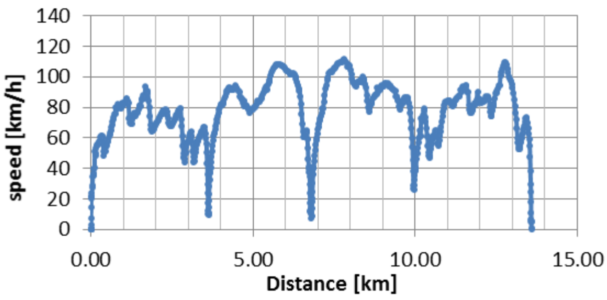


Fig. 3 an example of obtained speed profile

After, cross-sections were positioned along the driving routes each 25 meters. Cross sections were not placed in segments of the stretches near to intersections or significant driveways. In fact, those areas could have an unverifiable influence on the speed of each driver.

Hence, 61 road cross-sections on the stretch 1 and 76 road cross-sections on the stretch 2 were identified along the driving routes.

Speed data were assigned to each road cross-section so defined by connecting the value of distance corresponding to each cross-section to the respective value of speed in the speed profile.

B. Predictor variables of speed

Drivers' speed can be influenced by several factors depending on users' characteristics, road geometric layout, traffic conditions, environment, vehicle features.

The variables considered in the experiment are three: the time, the road geometry and the behavioral differences. The first two variables were known a-priori, since the chronological schedule was fixed and the road geometry of the two stretches was reconstructed. The third variable was defined a-posteriori based on observed speed.

The variable "time" is based on the temporal structure of the driving tests. In fact, it varies from 1 to 6, according to the six days of testing:

- Test 1, the first day of testing (1st day)
- Test 2, the second day of testing (2nd day);
- Test 3, the third day of testing (3rd day);
- Test 4; the fourth day of testing (4th day);
- Test 5; the fifth day of testing (10th day);
- Test 6; the sixth day of testing (27th day).

The repetition of tests over days was planned in order to test if the route familiarity acquired over time can lead to changes in speed behavior.

The chosen schedule is similar to the experimental plan used by Martens and Fox [9]. However, as in this study also the possible presence of a long term memory effect after some interruptions in administering stimuli (represented by driving tests) was inquired, the fifth test was postponed and a sixth test more distant in time was introduced.

The variable "road geometry" is based on the geometric layout of the chosen test routes. As a synthetic variable representing road geometric characteristics, the available sight distance was employed. The sight distance is the unhindered length of road section that the driver can see ahead without considering the influence of traffic, weather and lighting. Sight distance takes into account both the horizontal and vertical alignments and can be computed for each direction of travel. The sight distance profile, for each stretch of road and for both directions of travel, was obtained by using the method of the Italian Road Design Standard [15] and video recordings of the paths in both directions. In this way, a value of the sight distance was assigned to each road cross-section for both directions of travel.

Cross-sections were clustered into four classes in respect to their computed value of sight distance. The four visibility classes were so defined:

- Low visibility: cross-sections with low sight distance (0-100 m);
- Medium-low visibility: cross-sections with medium-low sight distance (100-200 m);
- Medium visibility: cross-sections with medium sight distance (200-400 m);
- High visibility: cross-sections with high sight distance (400-600 m).

The low visibility interval was chosen considering that sight distances of about 100 m are indicated as critical sight distances, that is, the accident rate increases rapidly for smaller sight distances [16]. Furthermore, the high visibility interval was chosen according to Lamm et al. [17] who found that accident related to passing maneuvers increase when the sight distance is less than 400 m to 600 m. The intermediate interval was split into two classes (medium-low, from 100 m to 200 m; and medium, from 200 m to 400 m) in order to divide the remaining cross-sections into subsets more numerically homogeneous. In fact, cross-sections with sight distances included between 100 m and 400 m were the most numerous.

From this classification, 53 cross-sections with a low available sight distance, 110 with a medium-low available sight distance, 38 with a medium and 73 with a high available sight distance were obtained. (The same 137 road cross sections were considered two times because sight distances were computed in the two different directions of travel).

The variable "behavioral differences" is based on measured data. In fact, drivers were clustered into three groups by using the K-means algorithm, a nonhierarchical cluster analysis based on speed data. The number of clusters was chosen in order to maximize the silhouette values.

The three drivers' cluster were so defined:

- Cluster 1: drivers showing low speeding inclination (measured speed greater than the average), the "prudent" drivers;
- Cluster 2: drivers with medium speeding inclination (measured speed near the average), the "average" drivers;
- Cluster 3: drivers with high speeding inclination (measured speed above the average), the "aggressive" drivers.

Cluster 1 consists of 6 drivers, cluster 2 consists of 8 drivers and cluster 3 consists of 5 drivers according to the cluster analysis. Therefore, the three obtained clusters can be representative of speed behavioral differences between drivers. However, this classification was employed only for the regression tree, whereas for the ANOVA, an ad-hoc independent variable was defined. More in detail, considering behaviors of aggressive drivers could help in studying their proneness to aberrant behaviors [18], and in defining the driver behavior based on observed parameters (see e.g. [19]).

C. Classification and regression trees (CART)

Tree based methods are non-linear and non-parametric data mining tools for supervised classification and regression problems. They do not require any assumption about data distribution. The so-called tree is an oriented graph formed by a finite number of nodes departing from the root node.

When the tree is binary, each parent node is linked to only two children nodes: the left node and the right node. In tree growing, from the top to the bottom, predictors generate splits at each internal node of the tree. The choice of candidate predictors for the response variable to be studied is simplified by the fact that the method can accept predictor variables of any type (numeric, binary, categorical, etc.). In the tree structure, the conditional interactions among the predictors to explain the behavior of the response variable can be read. The tree is composed of branches; a branch is a subtree obtained by pruning the tree at a given internal node.

In a classification and regression problem, there is a training sample of *n* observations on the response variable Y that takes value 1, 2, …, k and *p* predictor variables, X_1, …, X_j. The aim of the problem is finding a model for predicting the values of Y from other X values. Tree methods yield rectangular sets A_n by recursively partitioning data taking one variable X at a time.

Regression trees are similar to classification trees, except that the Y variable takes ordered values and a regression model gives the predicted values of Y for each node.

The tree construction is made by repeating the following basic steps [20]:

- Start at the root node;
- For each X, find the set S that minimizes the sum of the node impurities in the two child nodes and choose the split [$X^* \epsilon S^*$] that gives the minimum overall X and S.
- If a stopping criterion is reached, exit. Otherwise, apply step 2 to each child node in turn.

So, the partitioning criterion to define the optimality function during choosing of the best split of the objects into homogeneous subgroups and the stopping rule to arrest the tree growing procedure should be defined.

In the regression tree-growing algorithm, the impurity of a node is measured by the *Least-Squared Deviation* (LSD) *R(t)*, which is simply the *within* variance for the node *t,* which is calculated as follows:

$$R(t) = \frac{1}{N_w(t)} \sum_{i \epsilon t} w_i f_i (y_i - \bar{y}(t))^2 \qquad (1)$$

where $N_w(t)$ is the number of sample units in the node t,   is the value of the response variable for the i-th unit and   is the mean (and the predicted value) of the response variable in the node t.

The LSD criterion function for split s at the node t is defined as follows:

$$\Delta R(s,t) = R(t) - p_L R(t_L) - p_R R(t_R) \qquad (2)$$

Where t_L and t_R are the left and right nodes generated by the split s, respectively, while p_L and p_R are the portions of units assigned to the left and right child node. The split s* maximizing the value of $\Delta R(s,t)$ is chosen. This value relates to the "improvement" in the tree, since it expresses the impurity reduction that can be obtained by generating the two child nodes. In other words the split providing the highest improvement in terms of tree homogeneity is chosen.

Tree growing was arrested basing on two criteria:

- The minimum decrease in the impurity is equal to 0.001;
- The maximum size of the tree, that is the maximum number of levels of the tree, is equal to 5.

Furthermore, the objective of the study is to assess the importance of the variables: time, road behavioral differences in the response variable distribution. One of the outputs in CART technique is the variable importance [21], which defines the variable ability to influence the model. The relative importance of variable X_j computes estimates of predictor importance for tree by summing changes in the Mean Squared Error (MSE) due to splits on every predictor and dividing the sum by the number of branch nodes. If the tree is grown without surrogate splits, this sum is taken over best splits found at each branch node. If the tree is grown with surrogate splits, this sum is taken over all splits at each branch node including surrogate splits. At each node, MSE is estimated as node error weighted by the node probability.

Variable importance associated with this split is computed as the difference between MSE for the parent node and the total MSE for the two children.

To reduce the risk of type I error [22], the dataset was randomly split in two equal parts: an exploratory sample and a holdout sample. The exploratory sample was used to generate the regression tree and the holdout sample was used to validate it.

The selection of this analysis technique was made based on its previous use in road safety studies [23], [24], [25], [26]. All the analyses were performed by employing the software SPSS.

D. *One way mixed-design ANOVA*

The mixed-design ANOVA model (also known as Split-plot ANOVA (SPANOVA)) tests for mean differences between two or more independent groups whilst subjecting participants to repeated measures.

These groups have been split on two factors (independent variables), where one factor is a within-subjects factor and the other factor is a between-subjects factor. The mixed ANOVA can be used when a dependent variable (in this study, speed) has been measured over two or more time points or when all subjects have undergone two or more conditions. Time or conditions are the within-subjects or fixed factors. Instead, the between-subjects or random factors are the separate groups in which subjects have been assigned.

The primary purpose of a mixed ANOVA is to understand an interaction between the within-subjects factor and the between-subjects factor exists. Once the interaction has been established, post-hoc tests can be employed in order to recognize where the differences are.

The mixed one-way ANOVA is based on the following assumptions that should be met by the data set on which the analysis is performed:

- The dependent variable should be measured at a continuous level;

- The within-subjects factor should consist of at least two categorical related groups (groups composed of observations belonging to the same subjects);
- The between-subjects factor should each consist of at least two categorical independent groups;
- There should be no significant outliers in any group of each factor;
- The homoscedasticity assumption (homogeneity of variances for each combination of the groups) should be verified;
- The sphericity assumption (variances of the differences between the related groups of the within-subject factor for all groups of the between-subjects factor are equal) should be verified.

The mixed one-way ANOVA was chosen for the purpose of this study since the individual process of speed choice can be influenced by the human factors, and this was thought as an idiosyncratic factor affecting all responses from the same subject. Thus, in this way, the different responses can be rendered as inter-dependent rather than independent.

In more detail, two different analyses were performed considering, separately, all days of testing and all visibility classes, while Bonferroni post-hoc tests were carried out to isolate where the differences are.

In the first analysis, it was tested whether there is a difference in mean speed between the six days of testing, whereas the six days of testing as fixed effect, and the 20 drivers as random effect.

The second analysis was used to test whether there is a difference in speed mean between visibility classes, whereas visibility classes as fixed effect, and the 20 drivers and the six days as random effects.

So, behavioral differences were considered in the performed ANOVAs by introducing an independent variable able to explain differences among the sample of drivers.

All the analyses were performed by employing the software SPSS.

E. Data

The classification tree technique and the mixed one-way ANOVA model were employed in order to study the influence of the defined variables, namely: the day of testing (time), the visibility (road geometry) and the drivers' behavioral differences on the behavior of the response variable, the speed.

Means and standard deviations of speed were computed for each combination of the variables and they are shown in Table 1.

Table I descriptive statistics of speed data

Days	Visibility	Drivers' cluster	Mean speed (km/h)	St. Dev. (km/h)
1	Low	1	65,65	10,95
		2	74,24	9,74
		3	77,96	12,17
		Total	72,19	11,77
	Medium Low	1	71,36	9,67
		2	79,18	9,56
		3	83,59	12,74
		Total	77,54	11,41
	Medium	1	74,50	11,00
		2	80,43	11,44
		3	85,42	14,98
		Total	79,56	12,83
	High	1	80,80	10,19
		2	85,00	10,55
		3	90,99	14,71
		Total	84,93	12,08
	Total	1	73,57	11,51
		2	80,22	10,80
		3	85,03	14,27
		Total	79,07	12,65
2	Low	1	63,08	10,38
		2	77,18	9,59
		3	83,66	13,61
		Total	75,00	13,60
	Medium Low	1	70,24	11,30
		2	82,91	9,81
		3	89,90	14,94
		Total	81,31	13,99
	Medium	1	72,92	10,92
		2	88,32	10,37
		3	96,69	18,27
		Total	86,37	15,96
	High	1	79,20	9,42
		2	91,83	10,13
		3	101,45	16,28
		Total	91,00	14,59
	Total	1	72,00	11,89
		2	85,34	11,22
		3	93,22	16,91
		Total	83,80	15,46
3	Low	1	67,71	11,97
		2	77,67	10,96
		3	90,33	14,09
		Total	77,84	14,87
	Medium Low	1	74,72	11,31
		2	84,60	10,44
		3	98,16	13,00
		Total	85,08	14,48
	Medium	1	78,46	11,82
		2	87,12	12,33
		3	107,06	13,66
		Total	89,65	16,70
	High	1	84,97	10,85
		2	91,59	10,64
		3	110,06	14,60
		Total	94,43	15,41
	Total	1	77,00	12,82
		2	85,79	11,83

		3	101,60	15,49
		Total	87,20	16,19
4	Low	1	70,76	10,46
		2	80,06	12,52
		3	88,99	14,49
		Total	79,45	14,25
	Medium Low	1	77,09	9,48
		2	87,58	11,50
		3	99,21	12,70
		Total	87,35	14,02
	Medium	1	80,17	11,22
		2	89,42	15,30
		3	101,30	15,15
		Total	89,62	16,20
	High	1	85,93	10,47
		2	95,71	13,30
		3	107,89	15,85
		Total	95,83	15,63
	Total	1	79,04	11,45
		2	88,92	13,84
		3	100,33	15,61
		Total	88,80	15,84
5	Low	1	67,96	12,73
		2	82,16	11,56
		3	92,28	15,57
		Total	79,67	15,83
	Medium Low	1	75,57	13,42
		2	87,45	11,15
		3	101,78	12,73
		Total	86,67	15,60
	Medium	1	77,92	13,49
		2	91,15	12,36
		3	106,85	13,46
		Total	90,23	16,76
	High	1	84,85	12,09
		2	94,22	12,77
		3	112,30	13,50
		Total	95,14	16,22
	Total	1	77,28	14,13
		2	89,04	12,59
		3	103,96	15,17
		Total	88,44	16,83
6	Low	1	73,27	10,42
		2	81,40	11,29
		3	90,84	15,38
		Total	81,74	13,98
	Medium Low	1	79,46	11,65
		2	86,72	9,79
		3	98,59	14,45
		Total	88,01	13,81
	Medium	1	82,76	10,98
		2	88,70	11,34
		3	100,10	17,88
		Total	90,25	14,95
	High	1	89,74	11,60
		2	93,16	9,60
		3	106,02	14,98
		Total	95,86	13,57
	Total	1	81,80	12,68
		2	87,95	10,98
		3	99,65	16,10
		Total	89,52	14,73
Total	Low	1	68,06	11,65
		2	78,79	11,31
		3	87,50	15,04
		Total	77,67	14,46
	Medium Low	1	74,73	11,60
		2	84,75	10,82
		3	95,39	14,80
		Total	84,36	14,43
	Medium	1	77,77	12,05
		2	87,53	12,74
		3	99,82	17,20
		Total	87,65	16,09
	High	1	84,21	11,30
		2	91,91	11,75
		3	105,01	16,46
		Total	92,89	15,17
	Total	1	76,77	12,84
		2	86,22	12,31
		3	97,50	16,80
		Total	86,17	15,78

III. RESULTS AND DISCUSSION

The response variable assessed in the study is the driving speed. It will be analyzed with both the regression tree technique and the mixed one-way ANOVA.

A. Regression tree

The analysis of the obtained regression tree is made in this section. The obtained regression tree was attached to this paper.

Tree growing obtained from the exploratory sample produced 36 nodes. Of these, all nodes were validated. Therefore, the validated tree consisted of 36 nodes too. Furthermore, the regression tree produced 19 validated terminal nodes.

The primary split of the regression tree is the drivers' cluster. The tree keep away the cluster 3, that is the cluster of aggressive drivers (node 1) from the cluster 2 and 1, namely the clusters of average and prudent drivers (node 2). After, the tree splits the clusters (node 2) into the cluster 1, prudent drivers (node 5) and the cluster 2, average drivers (node 4). These two consecutive splits show the highest values of improvement (44.874 and 16.534) and these improvements are very large compared to all the others. This was expected, since the drivers' cluster was based on measured speed. However, the interesting result is that, clearly the speed behavior of aggressive drivers appears as completely different in comparison with the behavior of all other drivers. Therefore, in turn, speed behavior of prudent drivers is less evidently different from the speed behavior of average drivers.

Visibility is responsible of all the splits at the successive level. For the aggressive drivers and the average drivers (node 1 and node 6), the tree splits produced nodes representative of low and medium-low visibility on the left and medium and high visibility on the right (namely nodes 3 and 4 and nodes 13 and 14). Instead, for prudent drivers, the only difference is that medium visibility class is joined to the low and medium-low visibility classes and separated from the high visibility class (namely nodes 11 and 12). In those three splits, the values of the improvement measures are similar (7.204 for the split at node 1, 6.878 for the split at node 5, 5.964 for the split at node 6). Hence, some considerations about the influence of visibility on speed for different categories of drivers can be made. Results from the regression tree allows to consider that visibility influences drivers in the speed choice process in the same way independently from their proneness to speeding. However, for prudent drivers (cluster 1), speed data belonging to the medium visibility condition are clustered with the low and medium-low visibility after the split at node 5. This could mean that the drivers who are less prone to speeding, are also less prone to go faster unless visibility is high.

The variable time (day) is associated to the further splitting level, after the first improvement given by visibility, only for the aggressive drivers (split at nodes 3 and 4) and the average drives in medium and high visibility conditions (split at node 14). In these cases, the improvement given by the variable day (namely 3.681 at node 3, 4.200 at node 4 and 1.629 at node 14) is greater than an additional splitting into single visibility classes. Instead, for the prudent drivers in low, medium-low and medium visibility conditions (split at node 11) and the average drivers in low and medium-low visibility conditions (split at node 13), the improvement given by further dividing data based on visibility is greater than considering the variable time. This occurrence could be explained by the fact that aggressive drivers are more inclined to change their speed over time than the prudent drivers, independently from the visibility condition.

In all the considered splits associated to the variable day, the day 1 or the couple composed by the day 1 and the day 2 were kept away from the other days. This could mean that there is a clear difference between speed data of the first and second (where applicable) day of testing and all the other days. This tendency is confirmed by looking at the further splitting of the tree for prudent and average drivers (nodes 12, 21, 22, 25).

By looking at the terminal nodes, they result from further splitting based on the variables visibility and day. Generally, the last improvement is based on the splitting of speed data into single visibility classes (if the father node resulted from a split based on day, that is the case of aggressive drivers) or based on the variable time (if the father node resulted from a split based on visibility, that is the case of prudent drivers). An interesting exception to this rule is given by the case of aggressive drivers in medium and high visibility conditions and for days different from the day 1 (split at node 10). In fact, only in this case, a greater improvement is given by a further splitting of speed data based on the variable time. This could mean that, for aggressive drivers, the difference between medium and high visibility is perceived as very thin in comparison with their proneness to clearly change speed over the various days of testing.

Finally, the results from the analysis of the variable importance (Vim) are shown in Table 2.

Table II results of the variable importance analysis

Xj (jth variable)	Vim (Variable importance)[1]
Behavioral difference	61,408
Visibility	26,839
Day	13,706

[1]Dependent Variable: Speed

As expected, the most influencing variable is the behavioral difference between drivers. Furthermore, visibility is slightly more important than the variable day for the explanation of the response variable speed. However, as explained, the influence of time is more evident for some categories of drivers (the aggressive drivers) compared to others (the prudent drivers).

Instead, the influence of visibility is more homogeneously distributed among the drivers' categories.

B. *One-way mixed ANOVA*

The results from the one-way mixed ANOVA will be discussed in this section.

The first analysis consisted in testing differences in mean speed between the six days of testing, whereas the six days of testing as fixed effect, and the 20 drivers as random effect. A significant effect of days of testing on speed at the $p < .05$ level was found [F (5, 90.612) = 14.939, $p < 0.001$].

Furthermore, results from the Bonferroni test (Table 3) revealed that speed is statistically significantly lower in the first day of testing (79.068 ± 12.649 km/h) compared to all other days. Similarly, speed is statistically significantly lower in the second day of testing (83.802 ± 15.459 km/h) compared to days 3, 4, 5 and 6, in the third day of testing (87.205 ± 16.195 km/h) compared to days 4, 5 and 6 and in the fifth day of testing (88.442 ± 16.832 km/h) compared to day 6. Instead, there are no statistically significant differences between the fourth day (88.802 ± 15.845 km/h) and days 5 and 6 (89.515 ± 14.735 km/h).

Findings from the statistical analysis can be verified by looking at boxplots of speeds in the six days of testing (see Fig. 4).

A significant increase of mean speed over days can be noted while going from the first to the fourth day of testing. Instead, there are only slight differences between days 4, 5 and 6.

Moreover, a significant effect of the driver factor on speed at the $p < .05$ level was found [F (18, 88.165) = 16.795, $p < 0.001$]. Furthermore, there was a statistically significant interaction between drivers and days of testing on speed, F(88,27682) = 36.615, p < 0.001.

These results did not consider the road geometric layout as a variable able to predict speed. Therefore, cross-sections clustering into four classes with regard to their value of sight distance was considered.

Table III results from the Bonferroni post-hoc tests (difference between days of testing)

Day	**1**	**2**	**3**	**4**	**5**	**6**
1	1	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
2		1	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
3			1	**< 0.001**	**< 0.001**	**< 0.001**
4				1	1	0.083
5					1	**0.001**
6						1
ANOVA test: F-statistic = 14.939 p-value < 0.001						

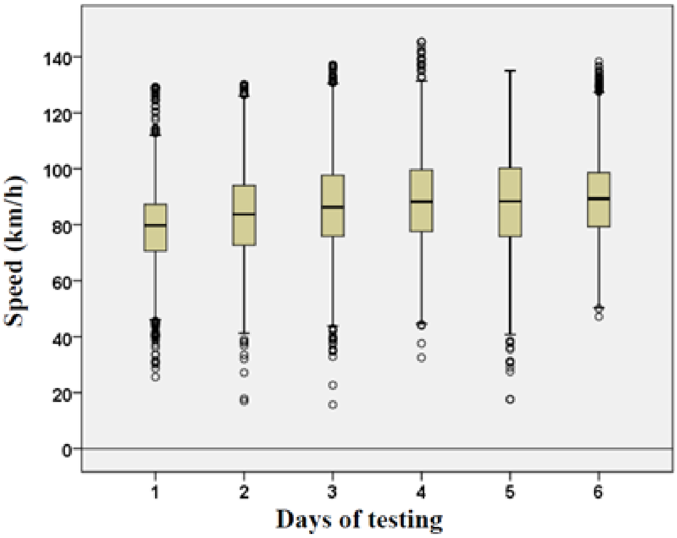


Fig. 4 boxplots of speeds in the six days of testing

The second analysis consisted in testing differences in mean speed between visibility classes, whereas visibility classes as fixed effect, and the 20 drivers and the six days as random effects. A significant effect of visibility on speed at the p <.05 level was found [F (3, 26.869) = 217.599, p < 0.001].

Furthermore, a Bonferroni post-hoc test showed that differences between the speed in each visibility class and the speed in all other classes are statistically significant.

Those results confirm that the chosen clustering of cross-sections into visibility classes is consistent and that road geometric characteristics have a strong impact on speed.

Nevertheless, there is no statistically significant interaction between visibility and days of testing on speed, F(15,258.032) = 1.111, p = 0.346. Therefore, even if globally speed is affected by days of testing and visibility, the way in which drivers modify their speed over days with the acquired route familiarity seems to be not influenced by the different visibility classes. Instead, there is a statistically significant interaction between drivers and visibility on speed, F(54,1078.929) = 1.924, p < 0.001 and a statistically significant interaction between drivers, days of testing and visibility on speed, F(258,560.960) = 4.631, p < 0.001. This means that behavioral differences between drivers are influential in the process of speed choice while acquiring familiarity with the route.

Results showed that, on average, as expected, speed increases with the sight distance, confirming results from the regression tree. At the same time, speed increases over days, but it happens independently from the visibility class (no interaction was found between visibility and days of testing). This means that, on average, acquiring familiarity with a route leads drivers to increase their speed in both higher and lower visibility conditions. In fact, for example, in low visibility condition (cross-sections in which available sight distance is less than 100 m), in the fourth day of testing (when familiarity seems to be already acquired) mean speed (79.449 ± 14.252 km/h) is significantly higher compared to the first day of testing (72.195 ± 11.768 km/h). Moreover, the increased speed is maintained over time in both the fifth and the sixth days of testing.

IV. CONCLUSIONS

A data set belonging to an on-road experiment was critically inquired into this study by using the regression tree technique and the ANOVA mixed models.

The variables considered were: behavioral differences, visibility and time.

Behavioral differences between drivers seem to be very influencing in explaining speed. In more detail, a clear difference between speed data belonging to drivers prone to speeding (aggressive drivers) and speed data belonging to more prudent drivers (the average and prudent drivers) was highlighted.

The way in which visibility influences speed seem to be independent from the drivers' proneness to speeding, except from the case of prudent drivers. In fact, the regression tree splits data basing on visibility into subsets formed by low and medium-low visibility on the left and by medium and high visibility on the right. Only for prudent drivers, medium visibility was clustered together with low and medium-low visibility. This occurrence can be explained by a low inclination to increase speed for prudent drivers, apart from the extreme case of high visibility condition. The existence of an interaction between visibility and driver behavioral differences on speed is confirmed by looking at results from the one-way ANOVA.

Moreover, differences between days were systematically highlighted by both the regression tree (in an advanced stage) and the post-hoc tests. In fact, speed changes over time: on average, speed increases over days as can be verified by looking at data in Table 1. In more detail, splits made by the regression tree and results from the ANOVA show that speeds of the first days of testing are clearly different from speeds in all other days. However, these differences are more evident for aggressive drivers than for prudent drivers. Therefore, aggressive drivers seem to be more prone to change speed over time than the others, especially in the higher visibility conditions, where speed choice is less conditioned by the road geometric layout. This was highlighted by both the regression tree and the mixed ANOVA. This finding, obtained by a more "naturalistic" study, confirms what found in literature: route familiarity leads to less cautious behaviors [8].

Therefore, both the use of the regression tree and of the mixed one-way ANOVA seem to suit well to the demand of explaining driving speed data belonging to an on-road experiment.

The regression trees technique is characterized by its flexibility and its simplicity in both the application and interpretation stages. It shows immediately where the main differences are and which variables are more influential than others. Moreover, the variable importance index is a powerful tool to estimate the explanatory capability of a variable. However, regression trees could need to be integrated with other quantitative measures in order to represent a complete analysis (see e.g. [23]).

Instead, the ANOVA mixed models requires some conditions to be met and they could be more demanding in both the application and interpretation stages. However, they can give quantitative estimates of the interactions between variables. Moreover, differences between each group can be easily found by using post-hoc tests. These techniques are more commonly used by researchers in all fields of study.

Thus, a further use of these techniques, in addition to the formerly cited employments in the field of road safety, was highlighted for future research.

References

[1] Fuller, R. Driver control theory: from task difficulty homeostasis to risk allostasis. In: Porter, B. E., *Handbook of traffic psychology*, Academic Press, Waltham, 2008.

[2] Näätänen, R., Summala, H. A model for the role of motivational factors in drivers' decision-making. *Accident Analysis and Prevention*, Vol. 6, 1974, pp. 243-261.

[3] Wilde, G. J. S. The theory of risk homeostasis: implications for safety and health. *Risk Analysis*, Vol. 2, 1982, pp. 209-225.

[4] Michon, J. A. Explanatory pitfalls and rule-based driver models. *Accident Analysis and Prevention*, Vol. 21, Issue 4, 1989, pp. 341-353.

[5] Vaa, T. Proposing a driver behavior model based on emotions and feelings: exploring the boundaries of perception and learning. In: Regan, M. A., Lee, J. D., Victor, T. W. (eds.). *Advances in research and countermeasures*, Vol. 1, Ashgate, Farnham, 2013.

[6] Tarko, A. P, Figueroa Medina, A. M. Implications of Risk Perception Under Assumption of Rational Driver Behavior, *Transportation Research Record: Journal of the Transportation Research Board*, 1980, Washington, D.C., 2006, pp. 8–15.

[7] Slovic, P., Fischhoff, B., Liechtenstein, S. Why study risk perception? *Risk Analysis*, Vol. 2, Issue 2, 1982, pp. 83-93.

[8] Yanko, M. R., Spalek, T. M. Route familiarity breeds inattention: A driving simulator study. *Accident Analysis and Prevention*, Vol. 57, 2013, pp. 80-86.

[9] Martens, M. H., Fox, M. R. J. Do familiarity and expectations change perception? Drivers' glances and response to changes. *Transportation Research Part F: Traffic Psychology and Behavior,* Vol. 10, 2007, pp. 476–492.

[10] Young, M., Stanton, N. Malleable attentional resources theory: a new explanation for the effects of mental underload on performance. *Human Factors*, Vol. 44, Issue 3 , 2002, pp. 365–375.

[11] Godley, S. T., Triggs, T. J., Fildes, B. N. Driving simulator validation for speed research. *Accident Analysis and Prevention*, Vol. 34, Issue 5, 2002, pp. 589-600.

[12] Colonna, P., Berloco, N., Intini, P., Ranieri, V. Route familiarity in road safety: speed choice and risk percetpion based on a on-road study. *94th Annual Meeting of the Transportation Research Board Compendium of Papers.* Washington, DC, 2015.

[13] Colonna, P., Aquilino, A., Berloco, N., Intini, P., Ranieri, V. The influence of memory on road user behavior. *Paper accepted for poster presentation at the 93rd Annual Meeting of the Transportation Research Board*. Washington, DC, 2014.

[14] Colonna, P., Aquilino, A., Berloco, N., Ranieri, V. Relationships between road geometry, drivers' risk perception and speed choice: an experimental study. *92nd Annual Meeting of the Transportation Research Board Compendium of Papers*. Washington, DC, 2013.

[15] Guidelines for the Design of Road Infrastructures. Italian Ministry of Infrastructures and Transport. D. M. n. 6792, 5/11/2001. Rome, 2001.

[16] Fambro, D. B., Fitzpatrick, K., & Koppa, R. J. Determination of stopping sight distances (No. 400). Transportation Research Board, 1997.

[17] Lamm, R., Psarianos, B., & Mailaender, T. *Highway design and traffic safety engineering handbook*, 1999.

[18] Tsironis, L., Moustakis, V., Mavropoulos, H., Maravelakis, E., & Bilalis, N. Discovering characteristics of aberrant driving behavior. In *Proceedings of the 5th WSEAS international conference on Simulation, modelling and optimization*, 2005, pp. 220-224. World Scientific and Engineering Academy and Society (WSEAS).

[19] Kamaruddin, N., & Wahab, A. Heterogeneous driver behavior state recognition using speech signal. In *The 10th WSEAS International Conference on System Science and Simulation in Engineering*, 2011, pp. 207-212.

[20] Wei-Yin Loh. *Classification and regression trees. WIREs Data Mining Knowledge Discovery*, Vol. 1, 2011, pp. 14-23, John Wiley & Sons, Inc.

[21] Strobl, C., Malley, J., Tutz, G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, Vol. 14(4), 2009, pp. 323-348.

[22] Webb, G. I. Discovering significant patterns. *Machine learning*, Vol. 68, 2007, pp. 1-33.

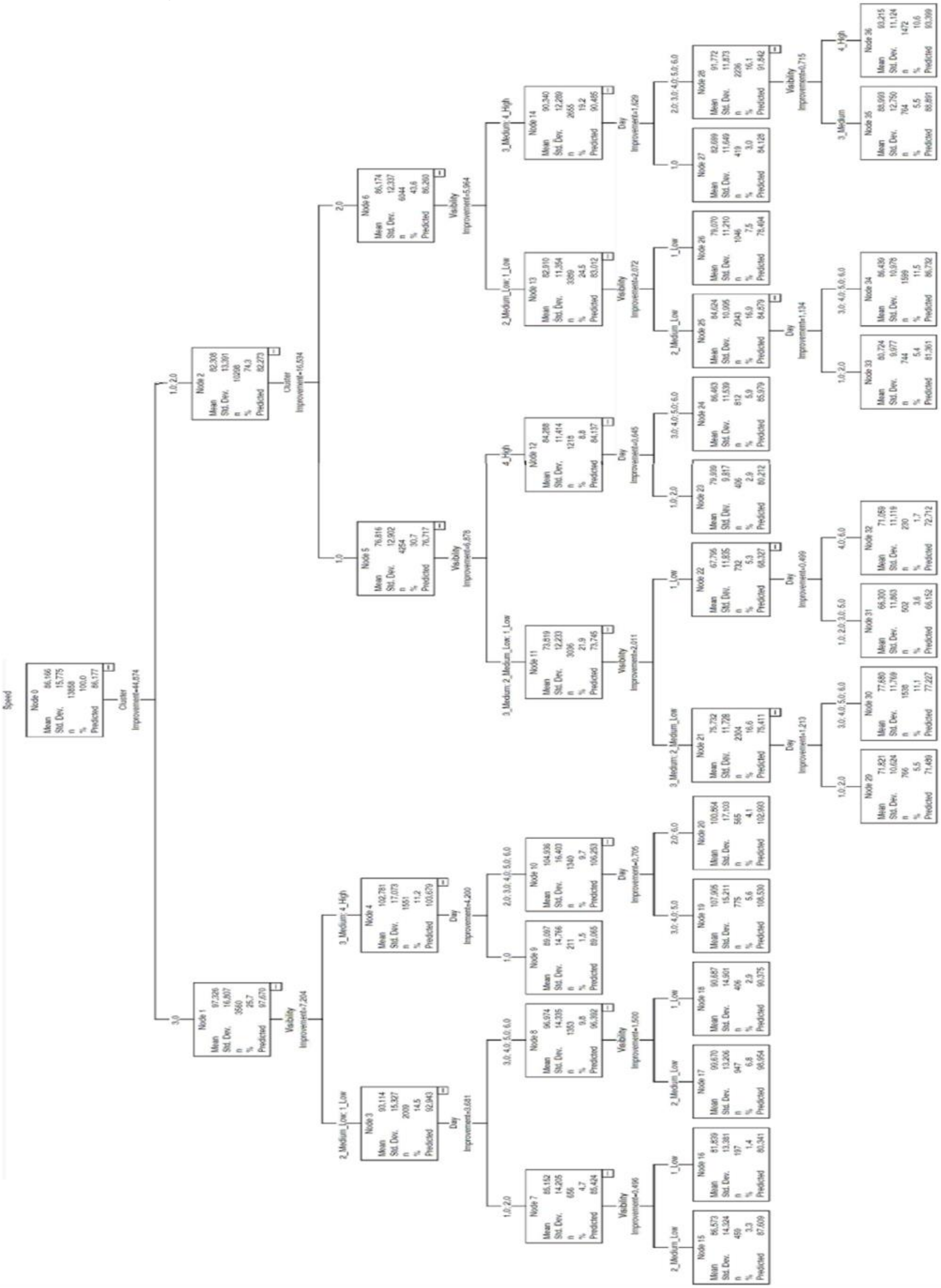
[23] Montella, A., Aria, M., D'Ambrosio, A., Mauriello, F. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. *Accident analysis and Prevention*, Vol. 49, 2012, pp. 58-72.

[24] Montella, A., Aria, M., D'Ambrosio, A., Mauriello, F. Data mining techniques for exploratory analysis of pedestrian crashes. *Transportation Research Record No. 2237*, Transportation Research Board of the National Acadamies, Washington, DC, 2011, pp. 107-116.

[25] Harba, R., Yanb, X., Radwanc, E., Sud, X. Exploring pre-crash maneuvers using classification trees and random forests. *Accident Analysis and Prevention*, Vol. 41, 2009, pp. 98-107.

[26] Kaur, D., & Pulugurta, H. Comparative analysis of fuzzy decision tree and logistic regression methods for pavement treatment prediction. *WSEAS Transactions on Information Science and Applications*, 2008, 5(6), pp. 979-990.

Attachment 1 – The regression tree.


Speed
Node 0 Mean 86,166 Std. Dev. 15,775 n 13858 % 100,0 Predicted 86,177
Cluster Improvement=44,874
1,0; 2,0
Node 2 Mean 82,308 Std. Dev. 13,391 n 10298 % 74,3 Predicted 82,273
Cluster Improvement=16,534
3,0
Node 1 Mean 97,326 Std. Dev. 16,807 n 3560 % 25,7 Predicted 97,670
Visibility Improvement=7,204
2,0
Node 6 Mean 86,174 Std. Dev. 12,337 n 6044 % 43,6 Predicted 86,260
Visibility Improvement=5,964
1,0
Node 5 Mean 76,816 Std. Dev. 12,902 n 4254 % 30,7 Predicted 76,717
Visibility Improvement=6,878
3_Medium; 4_High
Node 4 Mean 102,781 Std. Dev. 17,073 n 1551 % 11,2 Predicted 103,679
Day Improvement=4,200
2_Medium_Low; 1_Low
Node 3 Mean 93,114 Std. Dev. 15,327 n 2009 % 14,5 Predicted 92,943
Day Improvement=3,681
3_Medium; 4_High
Node 14 Mean 90,340 Std. Dev. 12,289 n 2655 % 19,2 Predicted 90,485
Day Improvement=1,629
2_Medium_Low; 1_Low
Node 13 Mean 82,910 Std. Dev. 11,354 n 3389 % 24,5 Predicted 83,012
Visibility Improvement=2,072
4_High
Node 12 Mean 84,288 Std. Dev. 11,414 n 1218 % 8,8 Predicted 84,137
Day Improvement=0,645
3_Medium; 2_Medium_Low; 1_Low
Node 11 Mean 73,819 Std. Dev. 12,233 n 3036 % 21,9 Predicted 73,745
Visibility Improvement=2,011
2,0; 3,0; 4,0; 5,0; 6,0
Node 10 Mean 104,936 Std. Dev. 16,403 n 1340 % 9,7 Predicted 106,253
Day Improvement=0,705
1,0
Node 9 Mean 89,097 Std. Dev. 14,766 n 211 % 1,5 Predicted 89,065
3,0; 4,0; 5,0; 6,0
Node 8 Mean 96,974 Std. Dev. 14,335 n 1353 % 9,8 Predicted 96,392
Visibility Improvement=1,500
1,0; 2,0
Node 7 Mean 85,152 Std. Dev. 14,205 n 656 % 4,7 Predicted 85,424
Visibility Improvement=0,496
2,0; 3,0; 4,0; 5,0; 6,0
Node 28 Mean 91,772 Std. Dev. 11,873 n 2236 % 16,1 Predicted 91,842
Visibility Improvement=0,715
1,0
Node 27 Mean 82,699 Std. Dev. 11,649 n 419 % 3,0 Predicted 84,128
1_Low
Node 26 Mean 79,070 Std. Dev. 11,210 n 1046 % 7,5 Predicted 78,494
2_Medium_Low
Node 25 Mean 84,624 Std. Dev. 10,995 n 2343 % 16,9 Predicted 84,679
Day Improvement=1,134
3,0; 4,0; 5,0; 6,0
Node 24 Mean 86,463 Std. Dev. 11,539 n 812 % 5,9 Predicted 85,979
1,0; 2,0
Node 23 Mean 79,939 Std. Dev. 9,817 n 406 % 2,9 Predicted 80,212
1_Low
Node 22 Mean 67,795 Std. Dev. 11,835 n 732 % 5,3 Predicted 68,327
Day Improvement=0,499
3_Medium; 2_Medium_Low
Node 21 Mean 75,732 Std. Dev. 11,728 n 2304 % 16,6 Predicted 75,411
Day Improvement=1,213
2,0; 6,0
Node 20 Mean 100,864 Std. Dev. 17,103 n 565 % 4,1 Predicted 102,993
3,0; 4,0; 5,0
Node 19 Mean 107,905 Std. Dev. 15,211 n 775 % 5,6 Predicted 108,530
1_Low
Node 18 Mean 90,687 Std. Dev. 14,901 n 406 % 2,9 Predicted 90,375
2_Medium_Low
Node 17 Mean 99,670 Std. Dev. 13,206 n 947 % 6,8 Predicted 98,954
1_Low
Node 16 Mean 81,839 Std. Dev. 13,381 n 197 % 1,4 Predicted 80,341
2_Medium_Low
Node 15 Mean 86,573 Std. Dev. 14,324 n 459 % 3,3 Predicted 87,609
4_High
Node 36 Mean 93,215 Std. Dev. 11,124 n 1472 % 10,6 Predicted 93,399
3_Medium
Node 35 Mean 88,993 Std. Dev. 12,750 n 764 % 5,5 Predicted 88,891
3,0; 4,0; 5,0; 6,0
Node 34 Mean 86,439 Std. Dev. 10,978 n 1599 % 11,5 Predicted 86,732
1,0; 2,0
Node 33 Mean 80,724 Std. Dev. 9,977 n 744 % 5,4 Predicted 81,361
4,0; 6,0
Node 32 Mean 71,059 Std. Dev. 11,119 n 230 % 1,7 Predicted 72,712
1,0; 2,0; 3,0; 5,0
Node 31 Mean 66,300 Std. Dev. 11,863 n 502 % 3,6 Predicted 66,152
3,0; 4,0; 5,0; 6,0
Node 30 Mean 77,680 Std. Dev. 11,769 n 1538 % 11,1 Predicted 77,227
1,0; 2,0
Node 29 Mean 71,821 Std. Dev. 10,624 n 766 % 5,5 Predicted 71,489