STATISTICAL LEARNING PROJECT

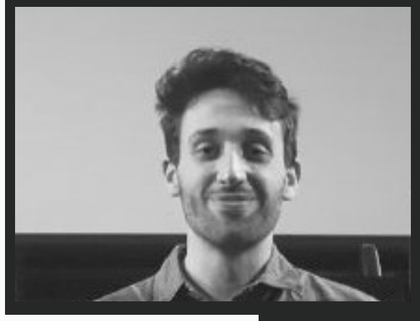
ACADEMIC YEAR 2018/2019

Road quality detection with smartphone sensors

Meet the Team

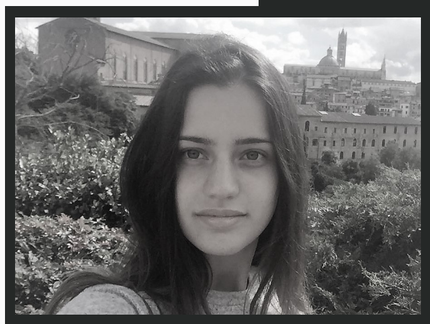
Group 12

ALESSANDRO FLABOREA



DAVIDE MANFREDINI

GIULIA SCIKIBU MARAVALLI



ILARIA SERVADIO

Our Goals

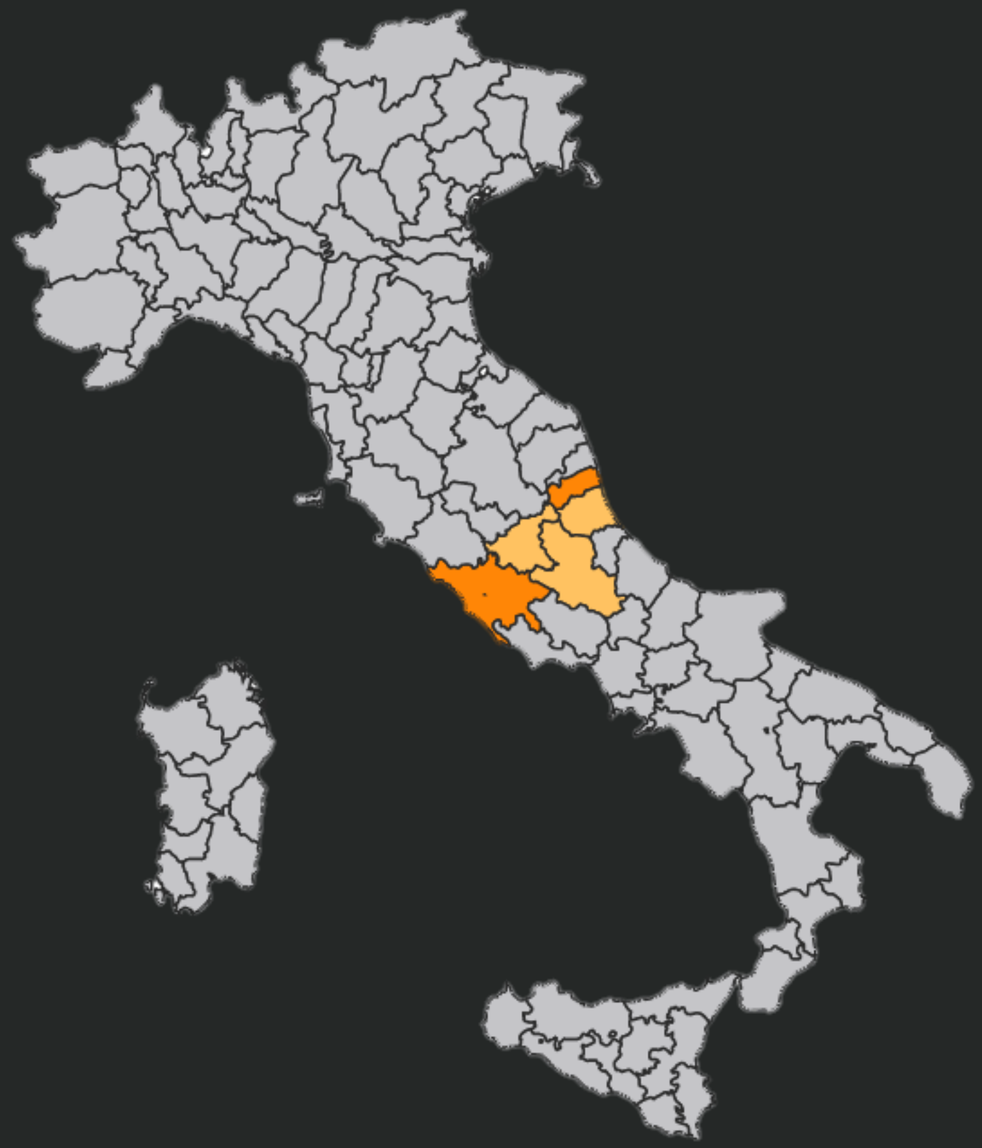
PREDICT:

- MEAN
- STANDARD DEVIATION
- AUTOCORRELATION

WITH REGRESSION

- ROAD TYPE

WITH CLASSIFICATION



Data Collection

Road data have been collected directly by us during our car trips. Data where recorded mainly in provinces of Ascoli Piceno and Rome using six different vehicle:

- Fiat Panda
- Toyota Yaris
- Opel Karl
- Peugeot 107
- Renault Megane
- Bus

In order to take records, we exploited sensors of Android and iOS smartphones.

Data Collection

GPS Logger:

Android app to collect additional data as coordinates and speed.

Tracks Logger:

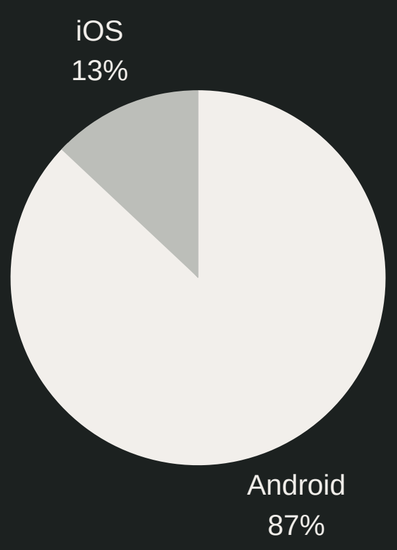
iOS app correspondent to GSP Logger for iPhones.

Science Journal:

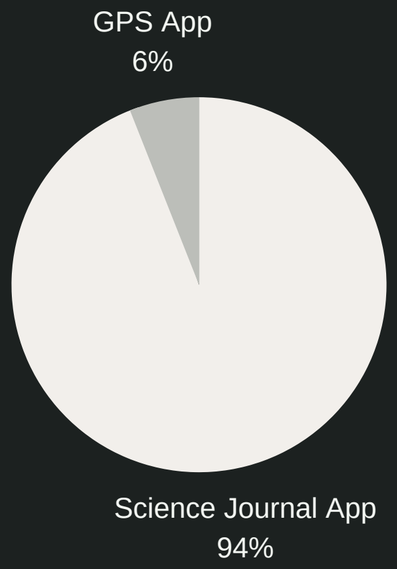
Android/iOS app to collect acceleration data in Z or Y.

Data Collection

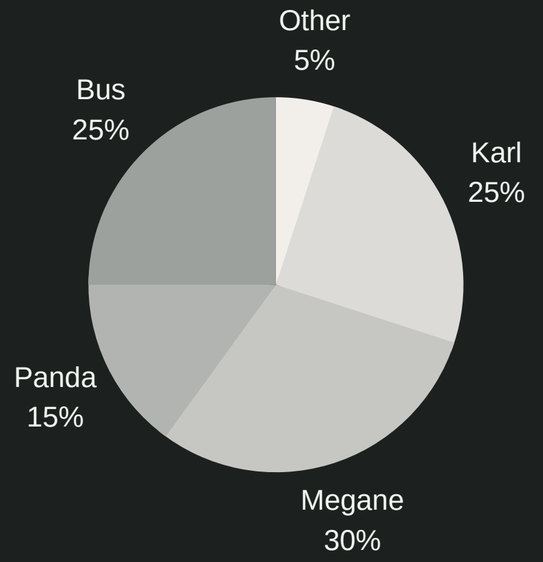
We recorded 107 MB of data, for a total of 556 unique roads. Let's discover something more about them:



Smartphone OS



Sj vs Gps App



Vehicle

Data Preprocess

Time	Latitude	Longitude	Elevation	Speed	AccX	AccY	AccZ	Bearing	Accuracy
19:43:29	42.850721	13.623339	153.241078	0.0	-0.002373	0.022186	0.141744	0.0	8.000
19:43:30	42.850675	13.623707	102.044242	0.0	0.006046	0.063188	0.144003	0.0	22.539
19:43:42	42.850717	13.623436	116.855100	0.0	-0.167065	-0.024882	-0.779089	0.0	24.000

relative_time	AccZ
0	8.178593
60	8.082825
123	8.140285

Data	Elapse (s)	Longitudine	Latitudine	Altitudine (m)	Corso	Velocità (km/h)	Distanza totale (km)
2019-04-22 10:59:17.172	0,0	12°41'25.74" E	41°46'52.17" N	397,6	0,0	0.0	0,000
2019-04-22 10:59:26.172	9,0	12°41'24.71" E	41°46'51.37" N	404,7	338,3	2.8	0,053
2019-04-22 10:59:28.172	11,0	12°41'24.69" E	41°46'51.30" N	404,8	340,1	2.5	0,056

Data Preprocess

Create two different script (one for Android and one for iPhone).

Delete/rename columns to make Android and iPhone files equal.

Convert coordinates to street name with reverse Geopy and Yandex.

Merge gps and science journal dataframes.

Output:

Date	Elapse(ms)	Latitude	Longitude	Street	Elevation	Speed(Km/h)	Acc
2019-04-25 19:44:00	31002	42.850754	13.623386	Via dei Narcisi	144.169198	1.842176	2.604892
2019-04-25 19:44:00	31012	42.850754	13.623386	Via dei Narcisi	144.169198	1.842176	2.595315
2019-04-25 19:44:00	31022	42.850754	13.623386	Via dei Narcisi	144.169198	1.842176	2.585738

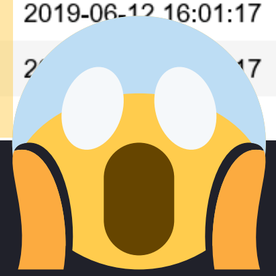
Remove inconsistent speed

IN ORDER TO OVERLOOK TIMES OF:

- LAUNCHING GPS/SJ APPS BEFORE START DRIVING
- WAITING AT TRAFFIC LIGHTS
- TRAFFIC QUEUE
- ALL OTHER POSSIBLE SITUATIONS IN WHICH THE CAR IS NOT MOVING (I.E. STATIONARY AND DOUBLE ROW)

Remove very high speed

	Date	Elapse(ms)	Latitude	Longitude	Street	Elevation	Speed(Km/h)	Acc	N_person	Car_type
0	3.0	9.901011	41.912503	12.451414	Via Giovanni Bovio	15.5	701.7	2019-06-12 16:01:17	3	0
1	3.0	10.008787	41.912503	12.451414	Via Giovanni Bovio	15.5	701.7	2019-06-12 16:01:17	3	0
2	3.0	11.233989	41.912503	12.451414	Via Giovanni Bovio	15.5	701.7	2[illegible]17	3	0
3	3.0	10.025701	41.912503	12.451414	Via Giovanni Bovio	15.5	701.7	[illegible]	3	0



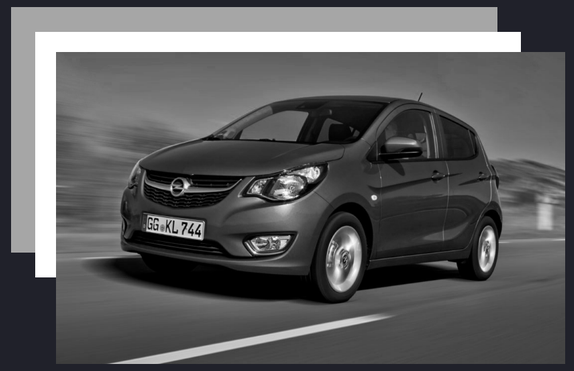
The sensors did not work properly:
remove all measurements that report speeds above 200 km/h to avoid any adverse impact on the analysis.

Feature engineering

Handmade Features

 # Person:

For each trip we saved the number of people in the car -> Payload



Car Type:

- 0 City car (superutilitaria);
- 1 Sedan car (berlina);
- 2 Crossover suv;
- 3 Bus;

Road Surface:

- 0 street without sampietrini (asphalt road);
- 1 street with sampietrini;

Automated Feature

For each street name identify the road type with reverse Geopy (using latitude and longitude) and OpenStreetMaps.

If the algorithm does not find the type of road (**error!**), We save the name of that road.

Error -> Check how many measurements we have: < 1000 we drop the road otherwise we try to manually replace the type of road.

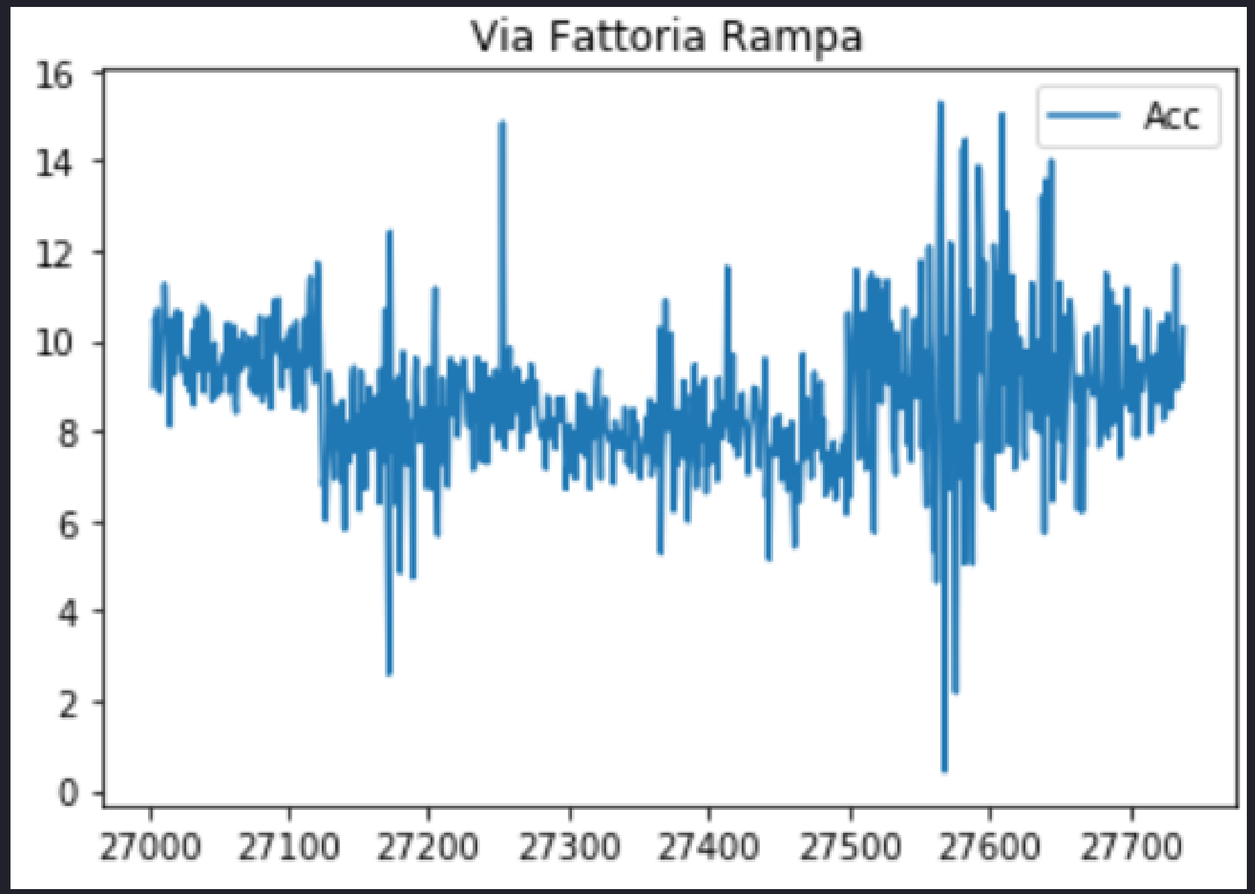
Automated Feature

type	
motorway	31
primary	122
residential	268
secondary	262
tertiary	70

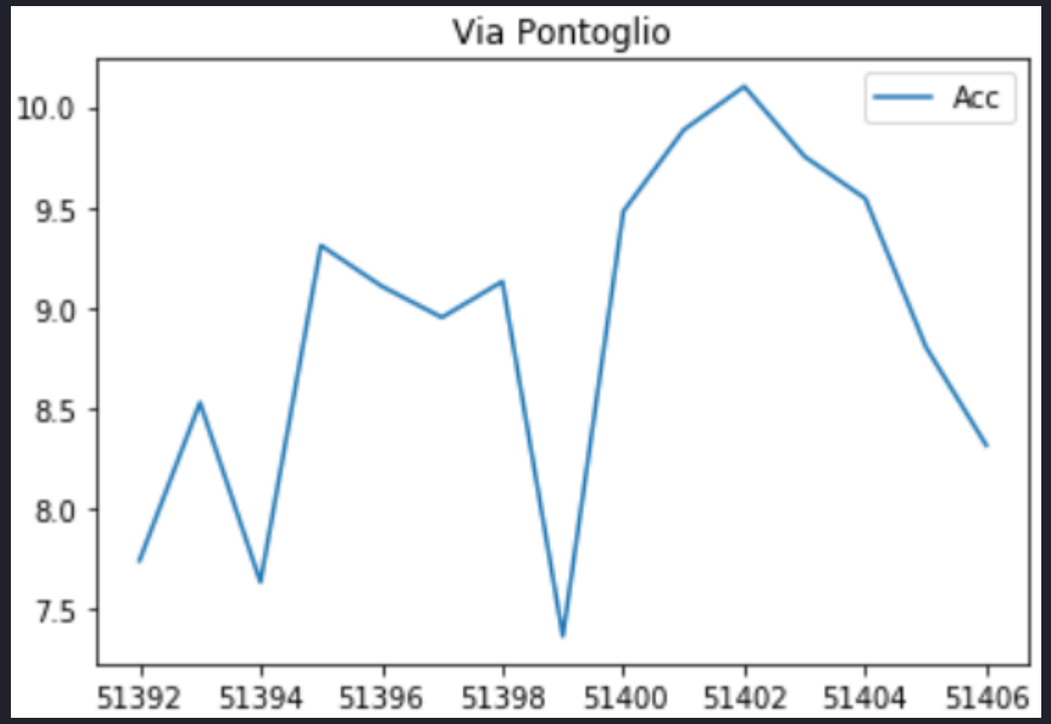
Take a look at the dataset with the new features

Acc	Car_type	Date	Elapse (ms)	Elevation	Latitude	Longitude	N_person	Speed (Km/h)	Street	type	Road_Surface
5.554548	1	2019-04-25 19:44:04	36533.0	152.455377	42.850746	13.623430	4	2.362088	Via dei Narcisi	residential	0
6.167464	1	2019-04-25 19:44:04	36543.0	152.455377	42.850746	13.623430	4	2.362088	Via dei Narcisi	residential	0
6.426037	1	2019-04-25 19:44:04	36557.0	152.455377	42.850746	13.623430	4	2.362088	Via dei Narcisi	residential	0
6.339846	1	2019-04-25 19:44:04	36562.0	152.455377	42.850746	13.623430	4	2.362088	Via dei Narcisi	residential	0

"Correct" Trend

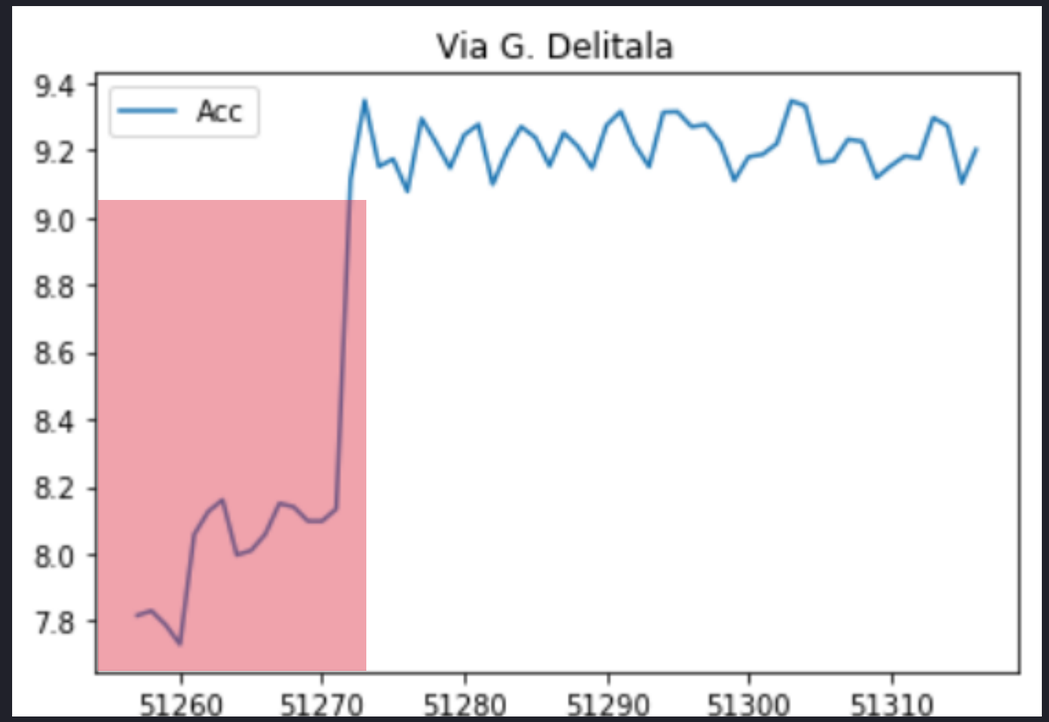


"Outliers"



Too few measurements

We do not know if the phone has malfunctioned or if the car has passed only a few moments on such a street.



Badly positioned phone

We removed only the "wrong" part of the measurements.

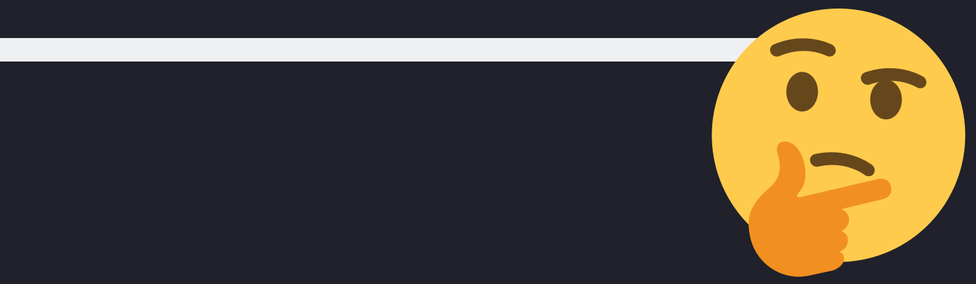
Final Dataset: each row, road trip

The data was taken on different road in several days and many of these was obtained by retracing the same route several times. Therefore we have decided to divide them. The first division consists in using the street names, the second exploits the date and time of beginning and end of a "experiment" on a particular street.

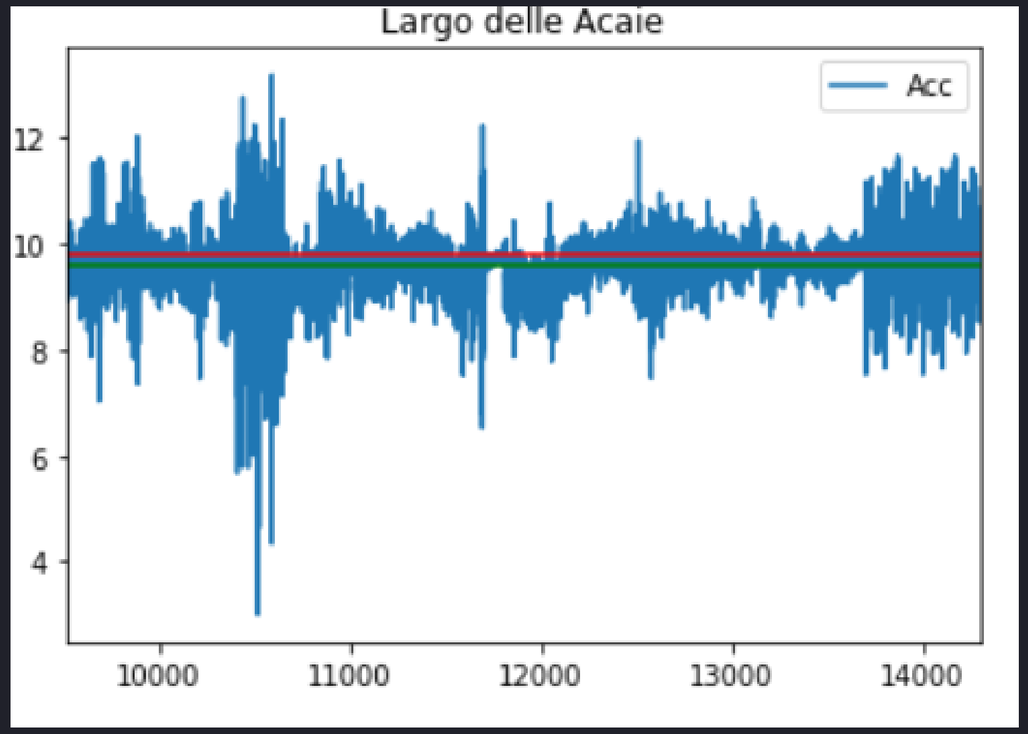
Car_type	Date	Latitude	Longitude	N_person	Street	type	Road_Surface	Mean_Acc	Mean_Speed (Km/h)	Std.Dev	Autocorr	Avg_Elevation
1	2019-04-25 19:44:04	42.850746	13.623430	4	Via dei Narcisi	residential	0	9.379572	8.720296	0.968745	0.726477	139.846453
1	2019-04-25 23:27:20	42.850326	13.623381	4	Via dei Narcisi	residential	0	9.567276	70.884363	0.485227	-0.246530	126.529794
1	2019-04-25 18:00:27	42.850345	13.623469	1	Via dei Narcisi	residential	0	9.632656	22.789475	0.227014	-0.355578	111.543697

753 trips

Why these statistics?



Why mean?

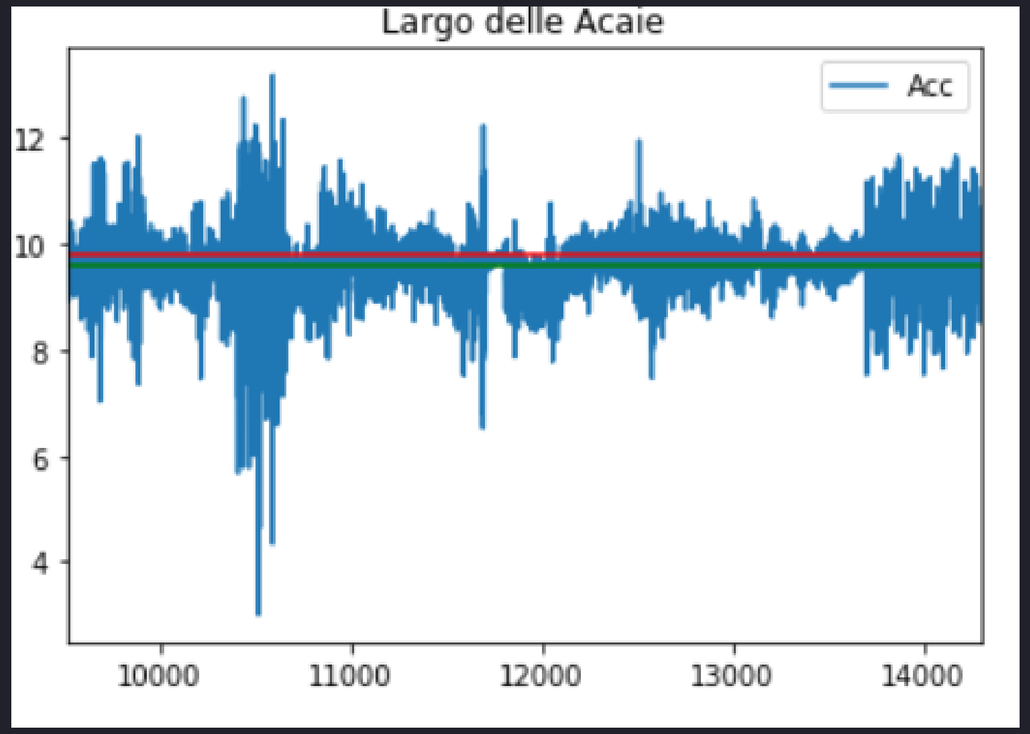


Gravitational attraction

Measure how much the mean is distance respect to the gravitational attraction

The ideal street, with the highest possible quality, have the mean near to 9.81.

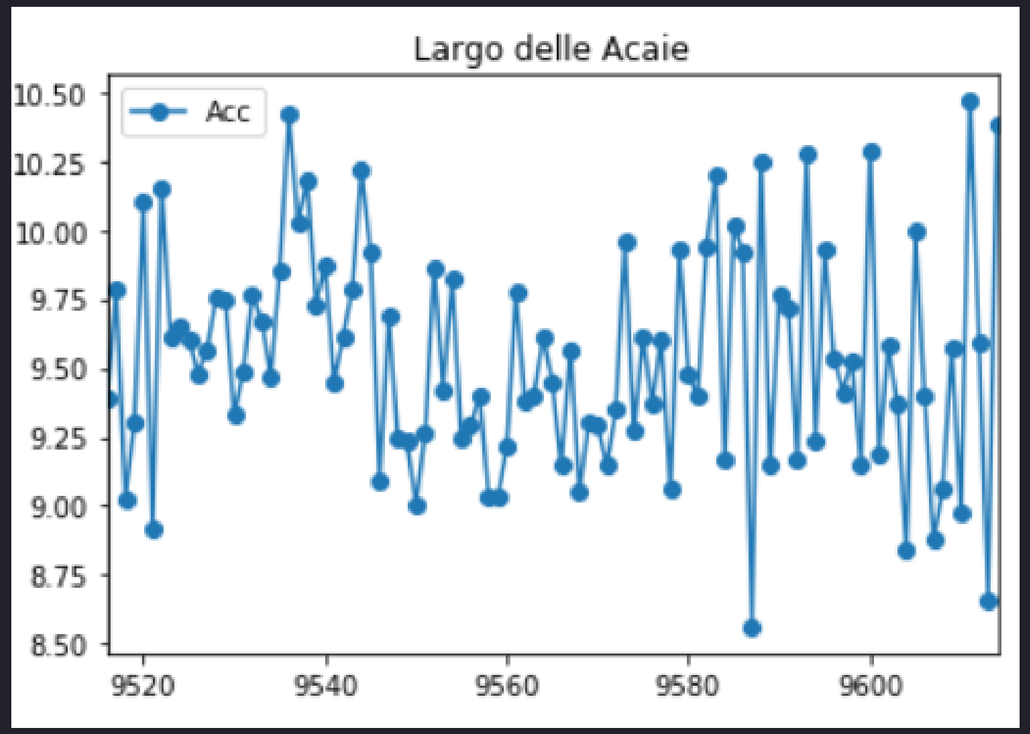
Why Standard Deviation?



We want a measure of data variability.

The higher the variance, the greater the deterioration of the road

Why Autocorrelation?



We want a measure of data stability.

If it is high, then the signal is not stable, that is a low quality street

PREDICTION

DATA SPLITTING

70%

TRAINING SET

30%

TESTING SET

MEAN - AUTOCORREALTION - ST. DEVIATION

Data

Standard Scaler

Get dummies

Model Selection

REGRESSION

MODEL SELCTION

CHOOSING RMSE AS METRIC

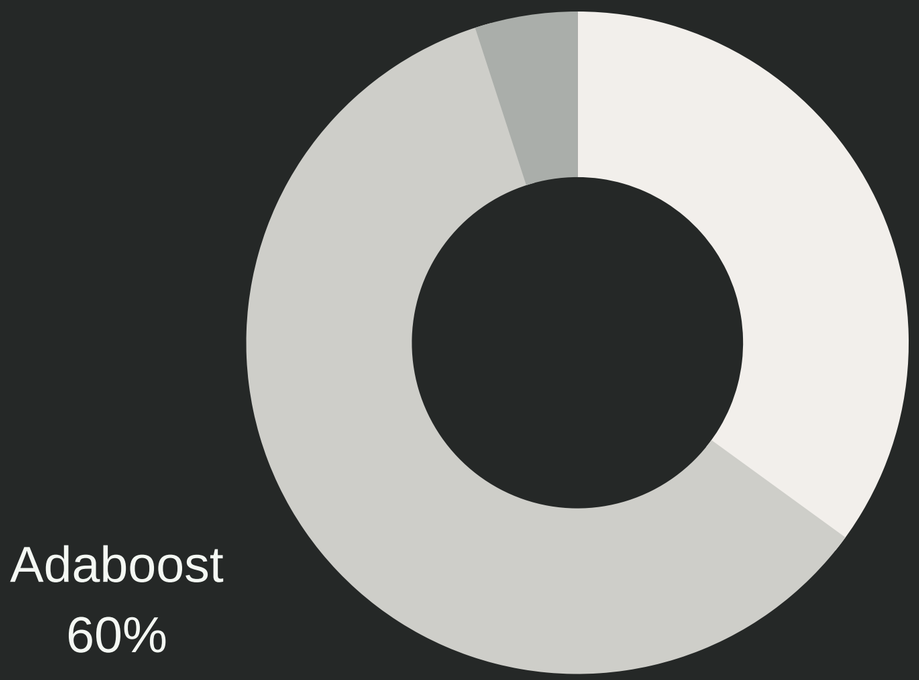
TUNING OF PARAMETERS WITH GRID SEARCH ON SPECIFIC MODELS: RANDOMFOREST, ADABOOST, XGBOOST, GBOOST, LASSO

ENSEMBLE MODELS WITH WEIGHTED AVERAGE

BEST MODEL MINIMIZE RMSE.

REGRESSION

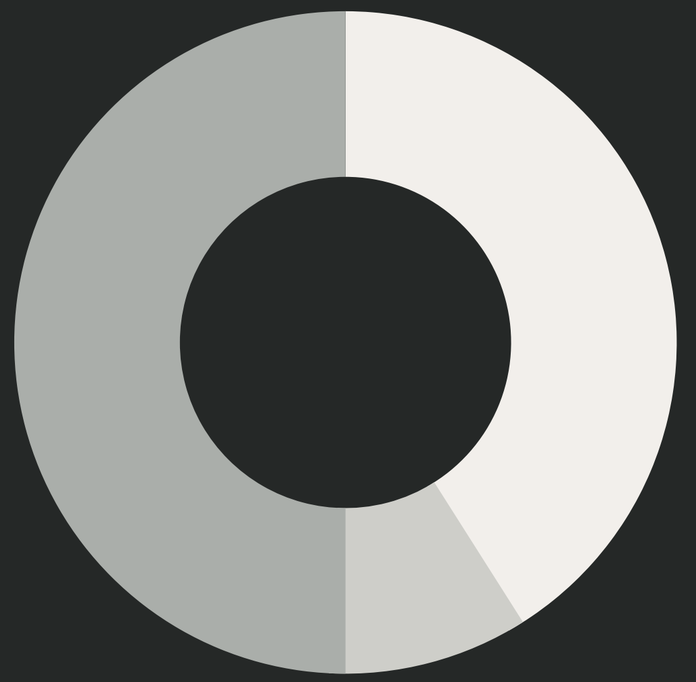
MEAN



0.1664333941

RMSE

AUTO-CORREALTION



0.24769410273

RMSE

STANDARD DEVIATION

RandomForest

100%

0.48925954380

RMSE

Latitude	Longitude	N_person	Car_type	Road_Surface	Mean_Speed(Km/h)	Avg_Elevation	type_motorway	type_primary	type_residential	type_secondary
41.912494	12.451478	3	0	0	17.116667	21.700000	0	0	1	0
41.912614	12.451197	3	0	0	13.444118	22.700000	0	0	1	0
41.913206	12.451742	3	0	0	19.791892	22.664865	0	0	0	1
41.912428	12.452481	3	0	0	15.100000	22.700000	0	0	1	0
41.911789	12.453039	3	0	0	18.720000	22.700000	0	0	0	0

	Std.Dev	Autocorr	Mean_Acc
0	0.715155	0.344985	9.504113
1	0.814061	0.219001	9.482833
2	0.911436	0.111042	9.501237
3	1.068644	0.460154	9.391031
4	0.708683	0.081764	9.535375

RESULTS

ROAD TYPE

Data

Standard Scaler

Model Selection
(the best one that minimize ACCURACY)

CLASSIFICATION

ROAD TYPE

RandomForest
100%

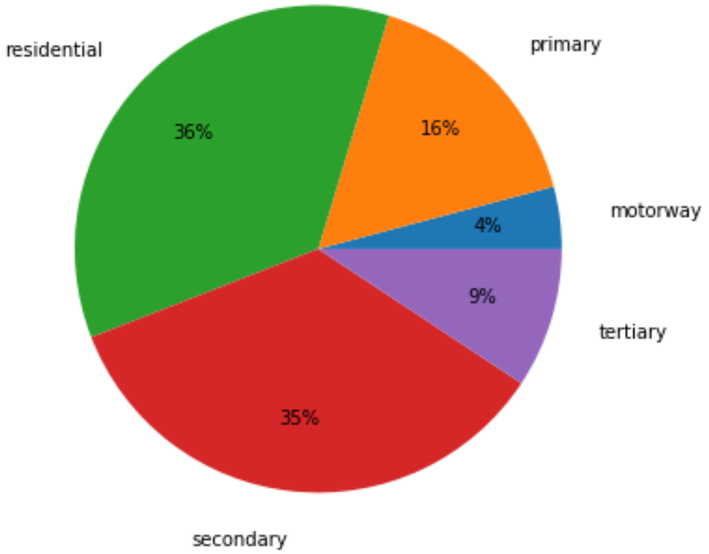
ACCURACY 67%

Latitude	Longitude	N_person	Car_type	Road_Surface	Mean_Acc	Mean_Speed(Km/h)	Std.Dev	Autocorr	Avg_Elevation
41.912494	12.451478	3	0	0	9.504113	17.116667	0.715155	0.344985	21.700000
41.912614	12.451197	3	0	0	9.482833	13.444118	0.814061	0.219001	22.700000
41.913206	12.451742	3	0	0	9.501237	19.791892	0.911436	0.111042	22.664865
41.912428	12.452481	3	0	0	9.391031	15.100000	1.068644	0.460154	22.700000
41.911789	12.453039	3	0	0	9.535375	18.720000	0.708683	0.081764	22.700000

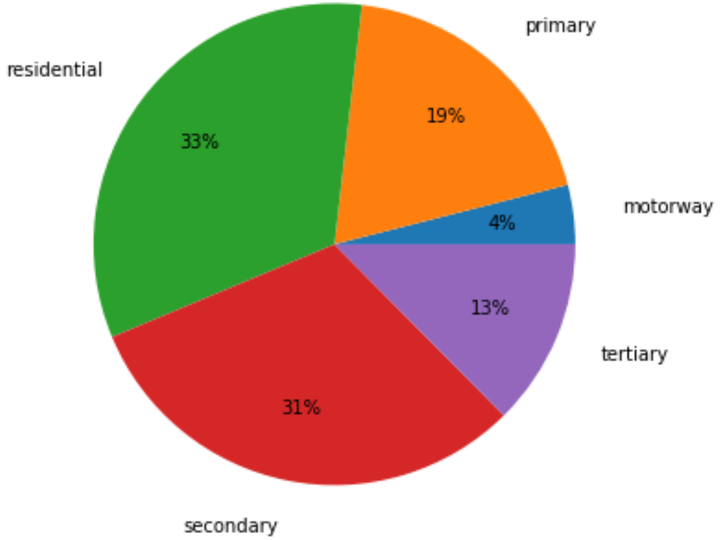
	type
0	residential
1	residential
2	secondary
3	residential
4	tertiary

Results in detail

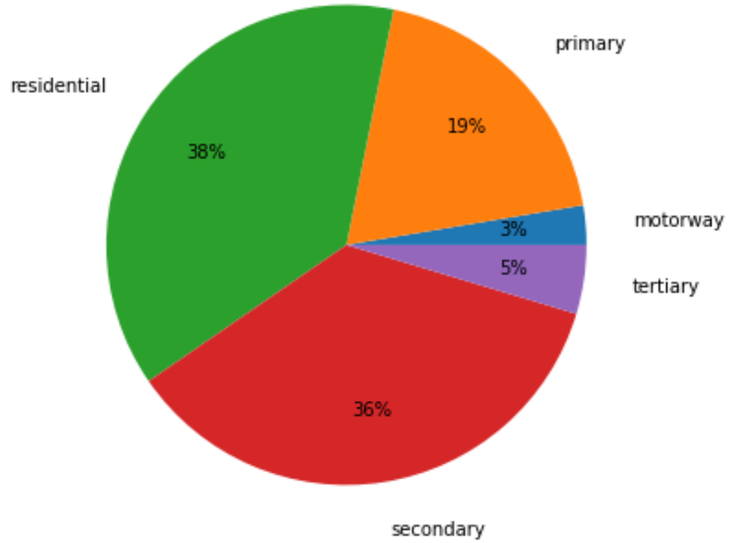
Typology



Train



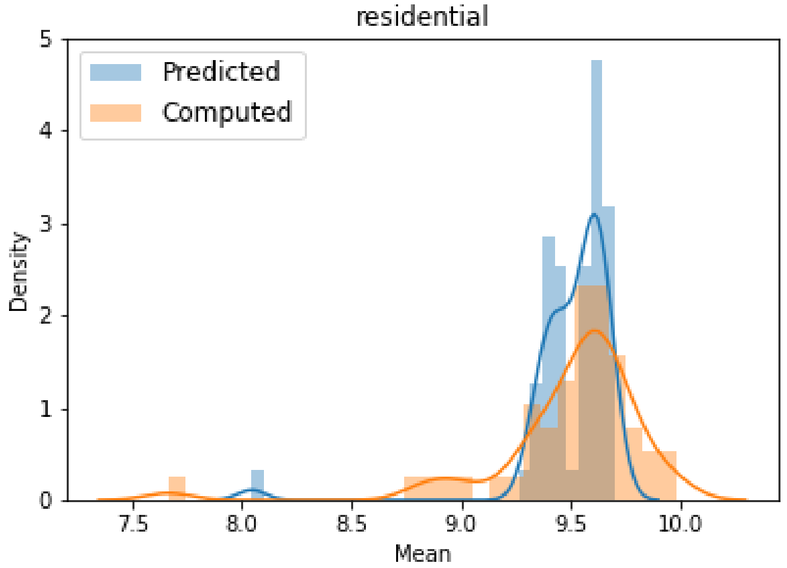
Test



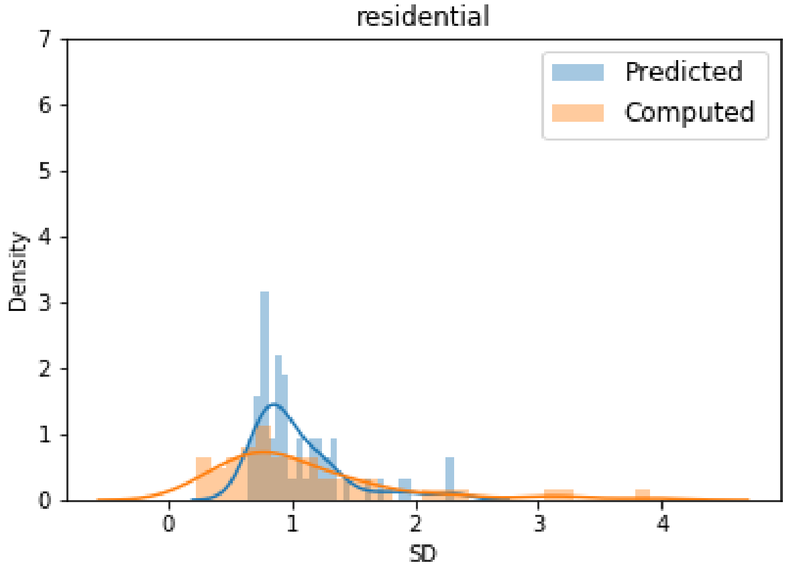
Predictions

Distributions

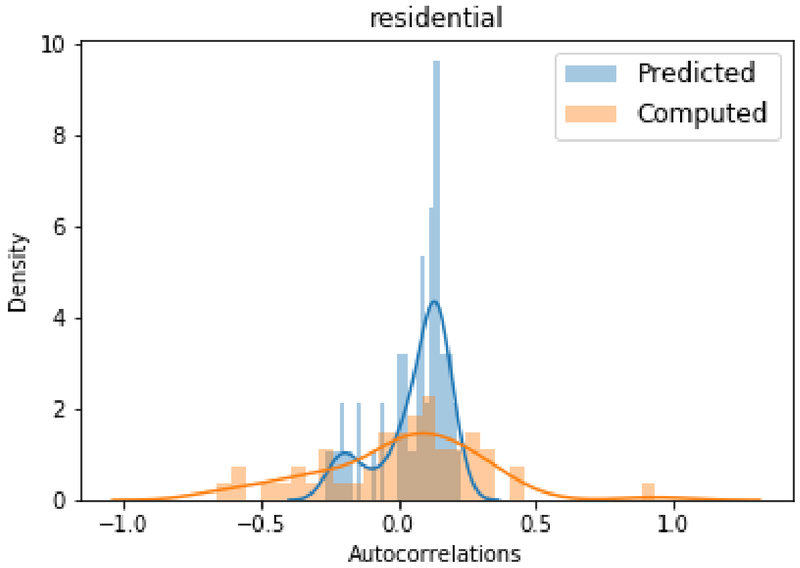
Residential



Mean



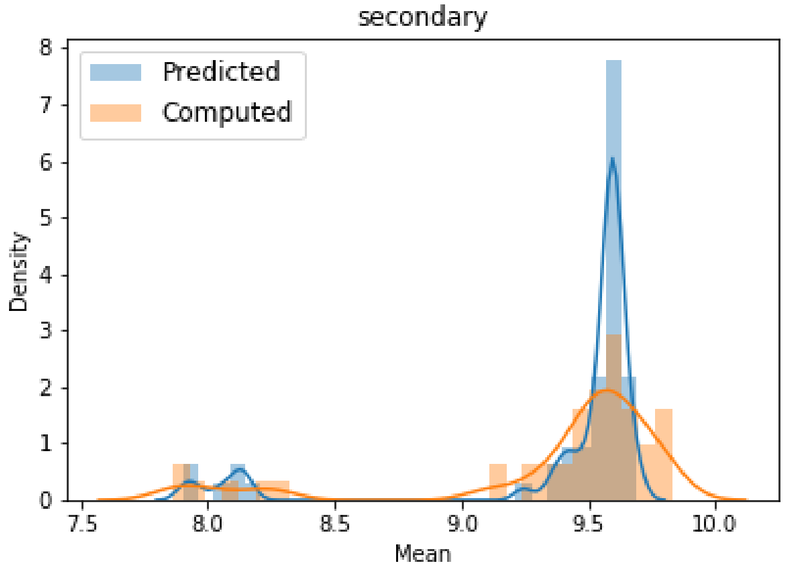
Standard Deviation

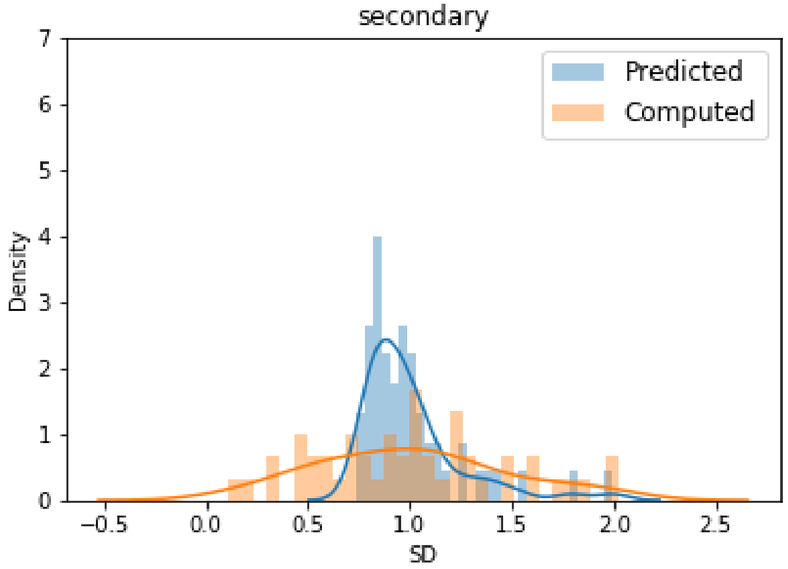


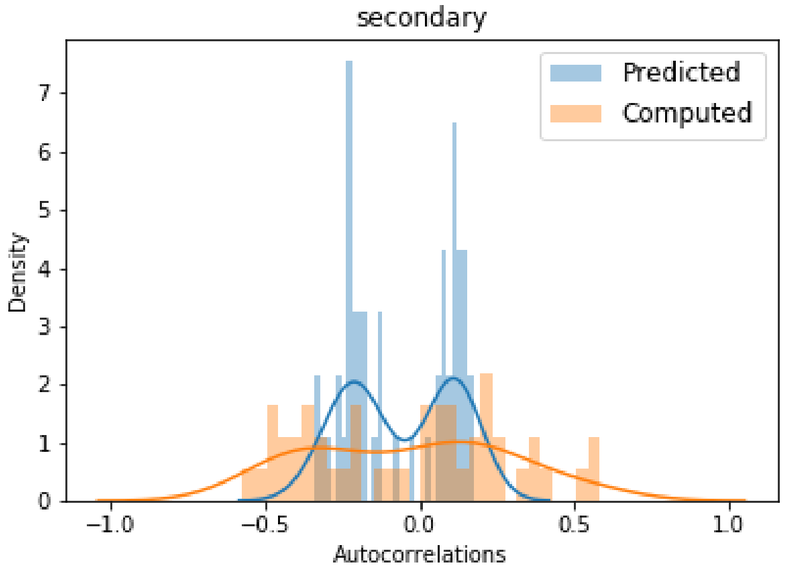
Autocorrelation

Distributions

Secondary







Mean

Standard Deviation

Autocorrelation

THE END