TRAFFIC LEVEL PREDICTION

A study on the city of Torino

A project by **Four pandas**

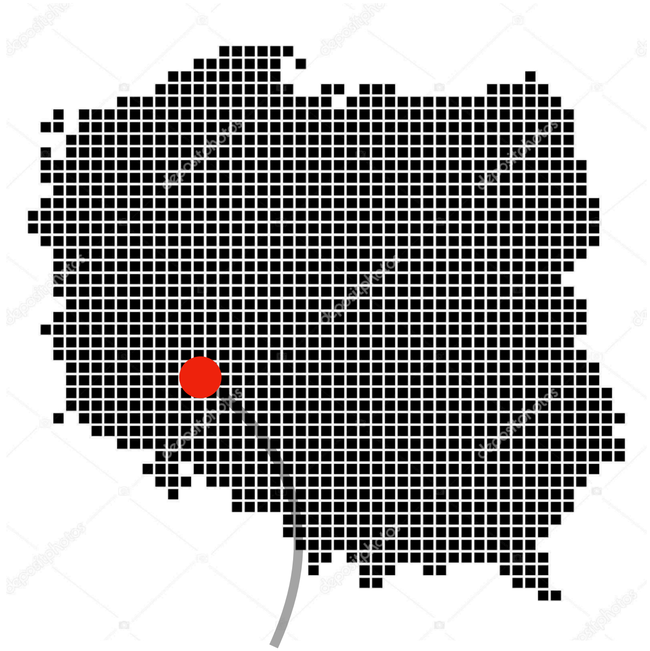
BIG DIVE

We are… Four pandas!

Marco Martellacci
#Scientist

Stefano Calderan
#DataScientist
#MangaLover

Daniele Morano
#Developer
#Murdock



Marek Kufel
#BIDeveloper

BIG DIVE

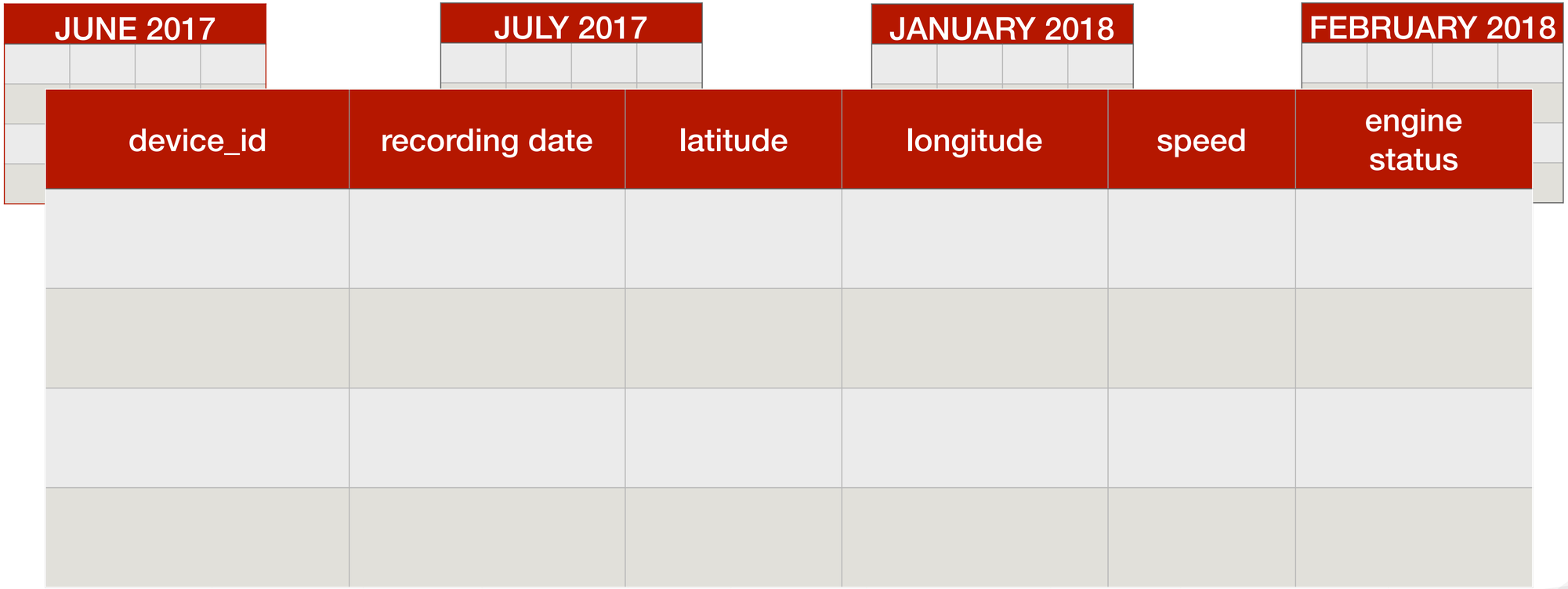
Our **Goal**

Check if it is possible to predict **traffic levels** in a **specific area** of the city, with a **reasonable forecast window**

- How is traffic level defined?
- What is a specific area?
- What do we mean by reasonable forecast window?

Starting Point

VEM dataset → **28 GB** across 4 datasets



BIG DIVE

Quick **Insight on Data**

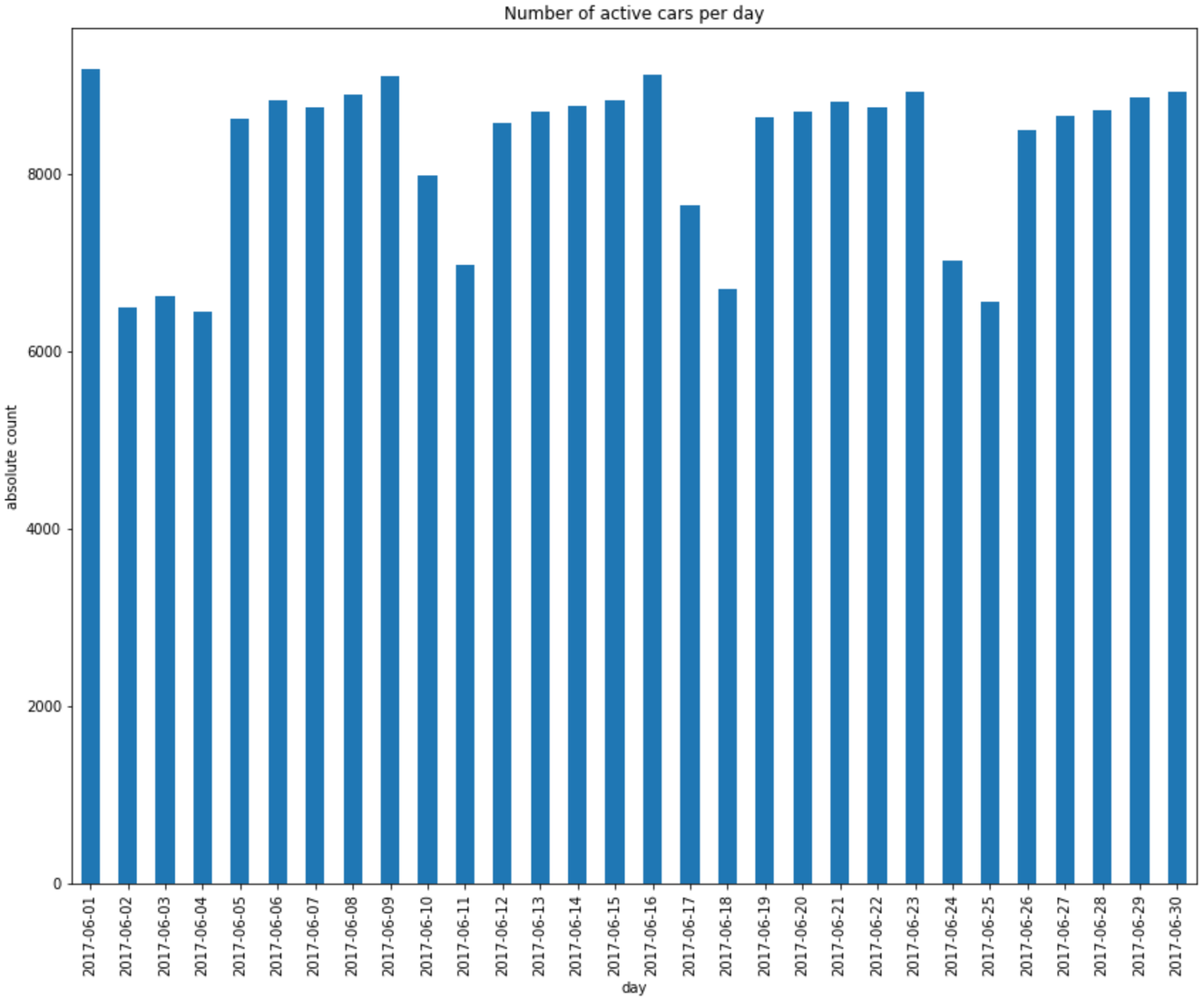
We selected the Torino area as a **rectangle** with minimum and maximum borders:

(7.57785 ; 7.77337)

(45.00678 ; 45.1402)

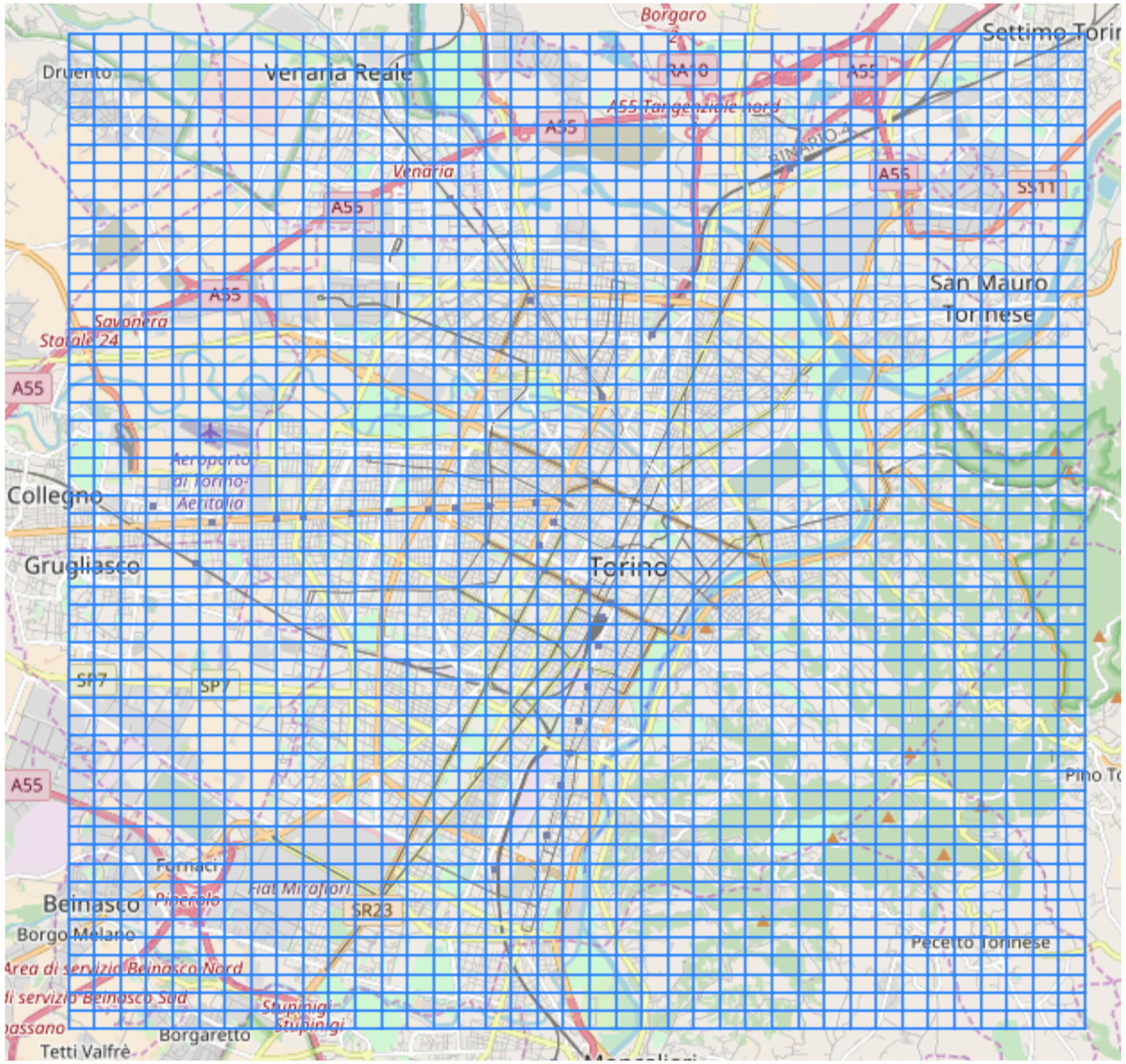
Total number of active cars:

20158



BIG DIVE

Grid **Creation**



Our **Goal**

Check if it is possible to predict **traffic levels** in a **specific area** of the city, with a **reasonable forecast window**

Initial choice for rectangle shaped zones.

Size of 400 m x 250 m

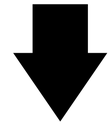
Started with 55 x 40 = 2220 zones

Original idea: predict traffic levels for every zone in the grid

BIG DIVE

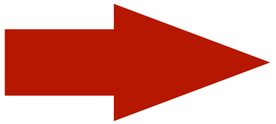
A **metric** for traffic levels

Our **Goal**



Check if it is possible to predict **traffic levels** in a **specific area** of the city, with a **reasonable forecast window**

$$TL(l,t) = \frac{N(l,t)}{\langle v \rangle_N}$$

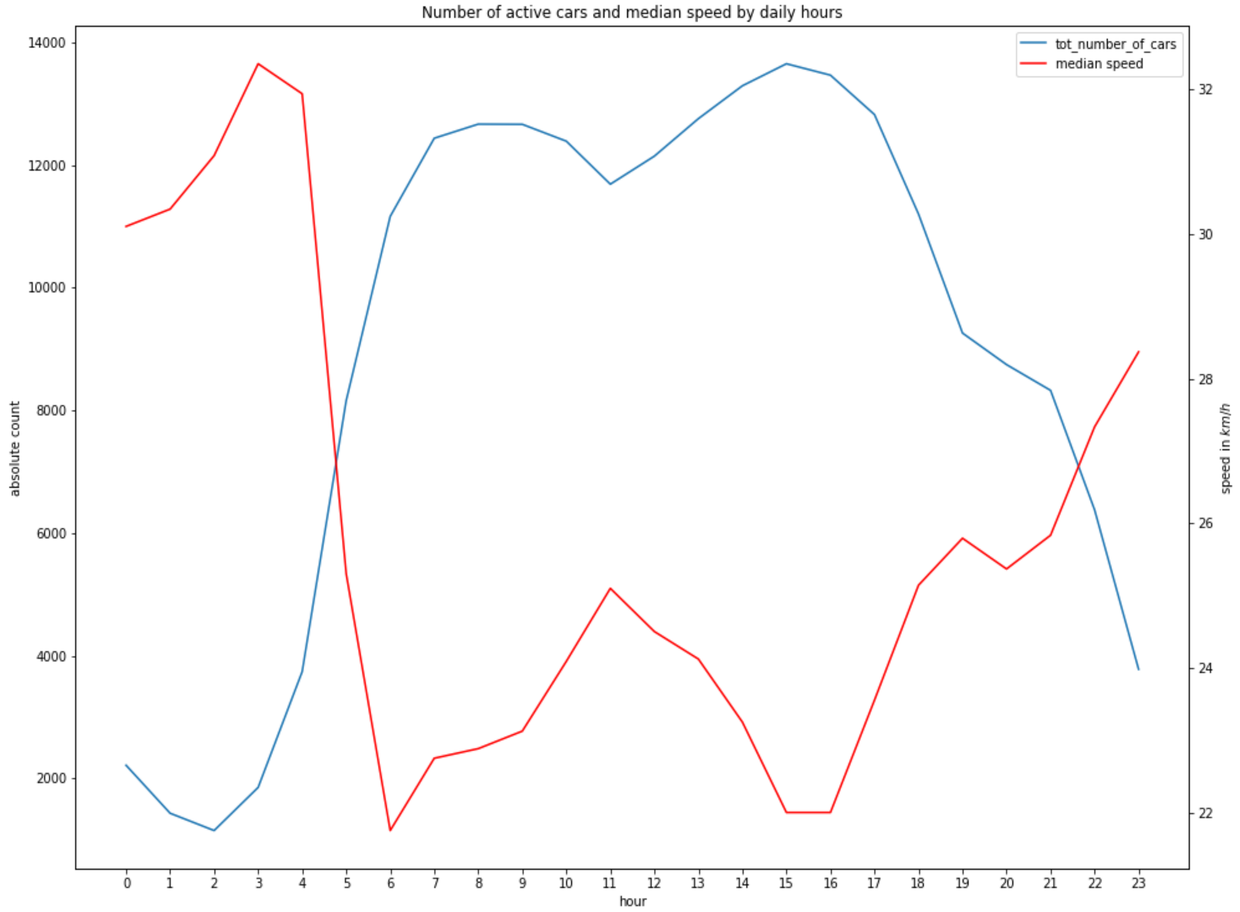


Is this good?

BIG DIVE

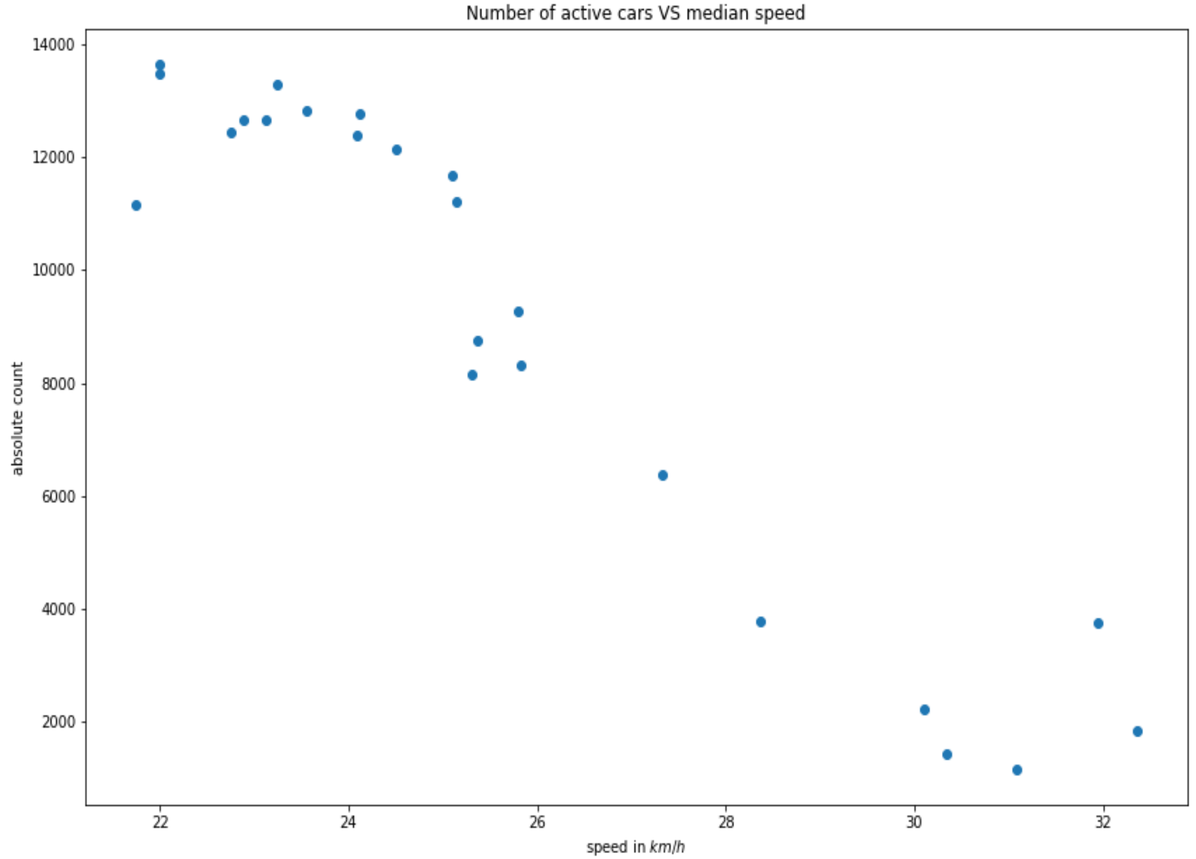
Exploratory analysis: daily hours

Number of active cars and median speed by daily hour



Number of active cars VS median speed

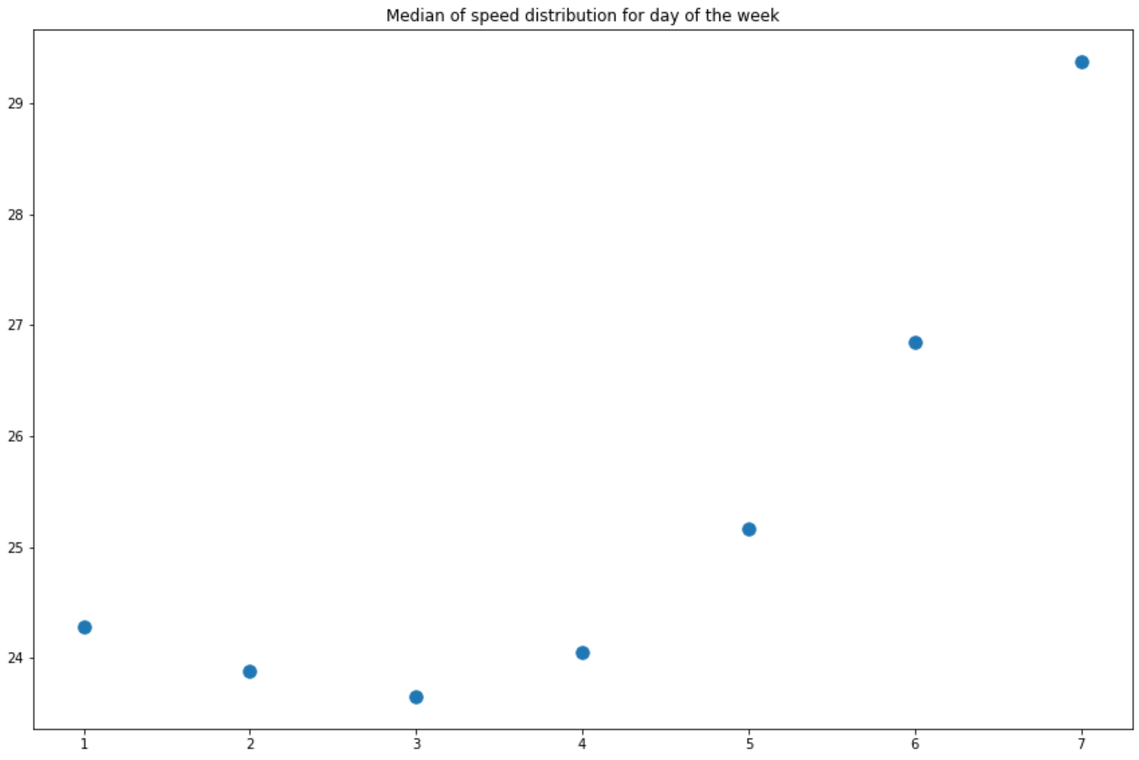
Pearson coeff: **r = - 0.95**



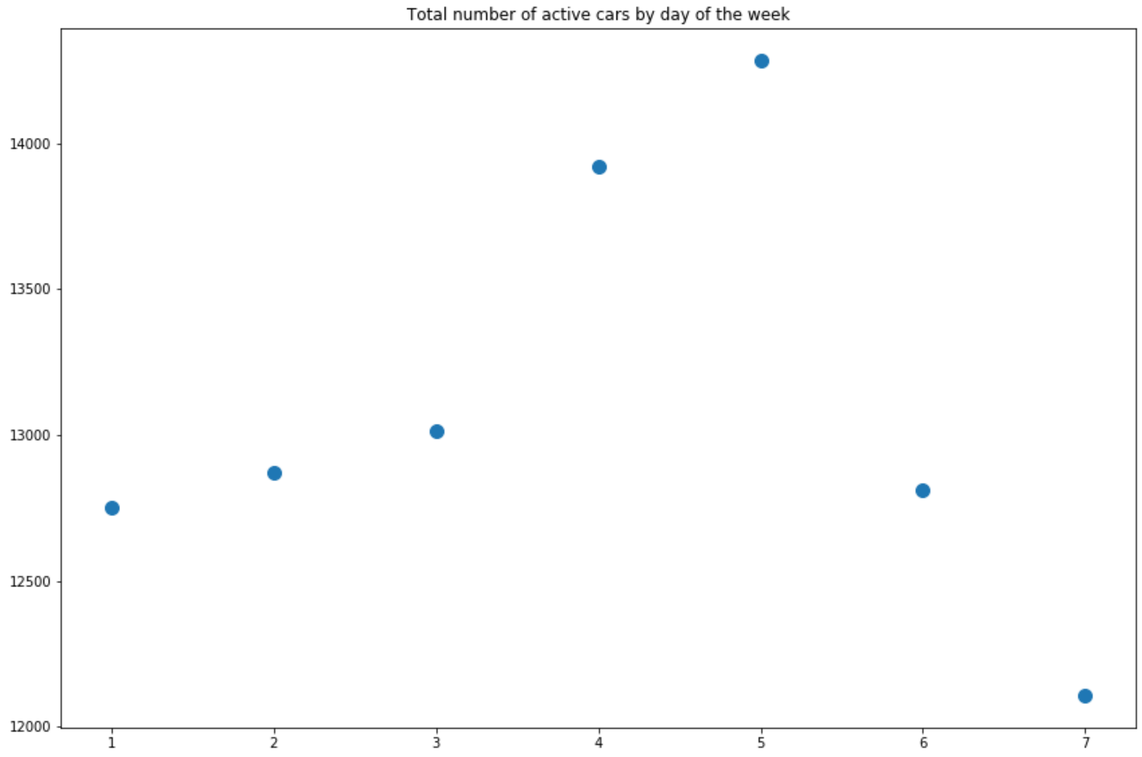
BIG DIVE

Exploratory analysis II: weekdays

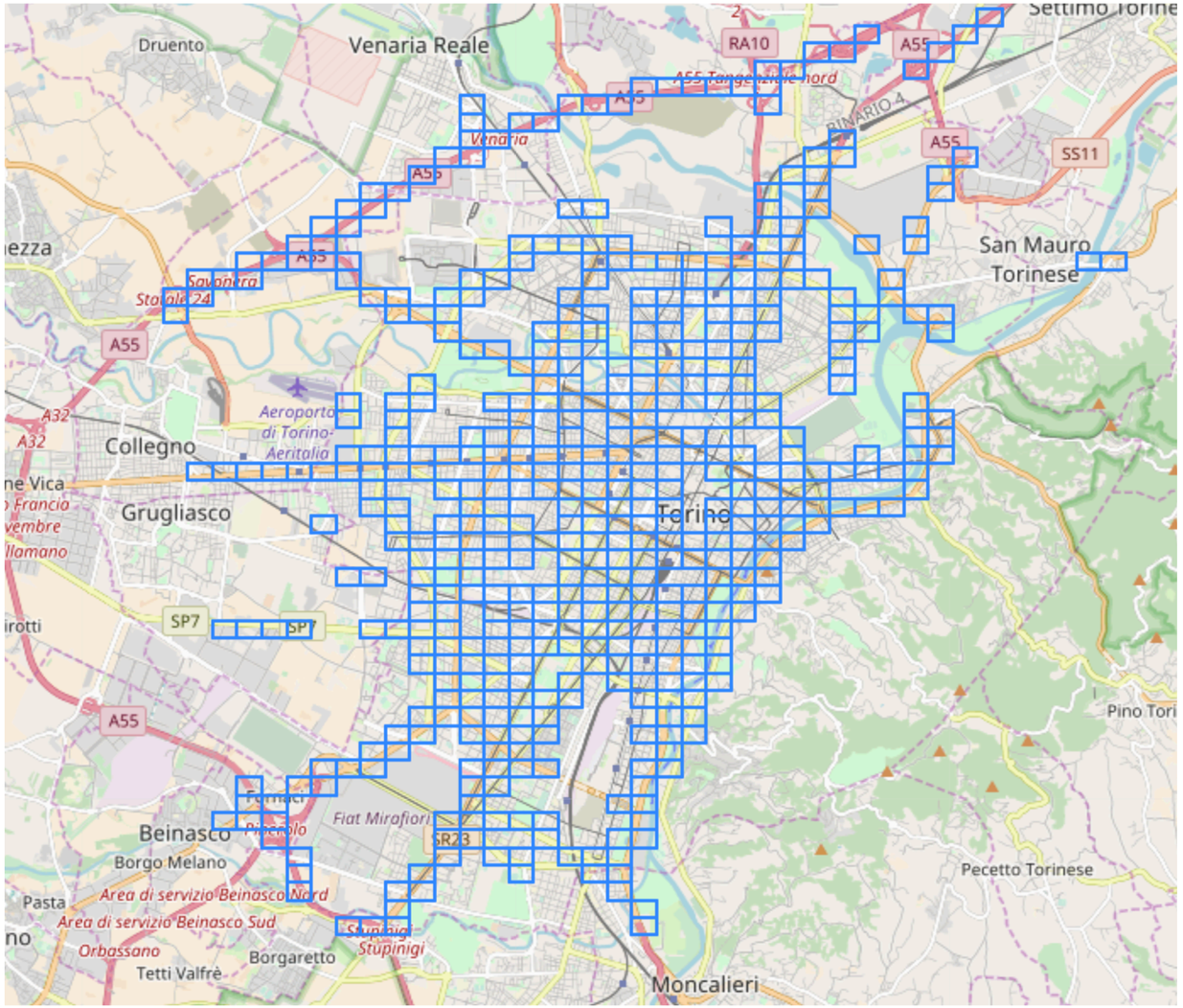
Median speed by day of the week



Total active cars by day of the week



Focus on grid zones selection



We decided to discard cells with data that could lead to **unreliable traffic estimations**.

The selection criterion was based on the **average daily count** of distinct cars crossing the zone.

Furthermore, we requested at least one neighbour for each zone.

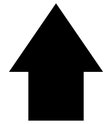
Final number of zones:

500

Facing a time **tradeoff**

Our Goal

Check if it is possible to predict **traffic levels** in a **specific area** of the city, with a **reasonable forecast window**



$$TL(l,t) = \frac{N(l,t)}{\langle v \rangle_N}$$

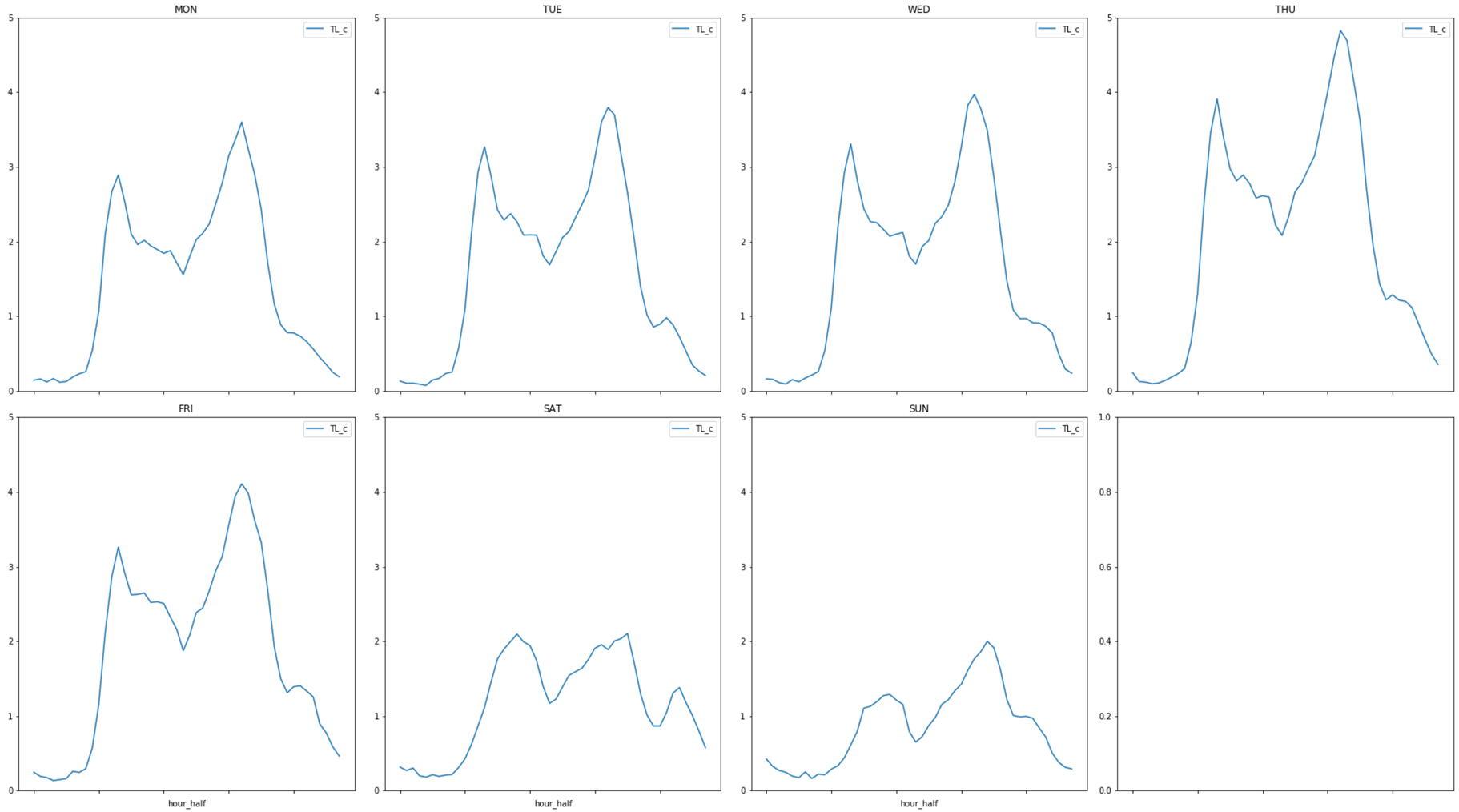
At the same time, we want a forecast range that is

- **large enough** -> collects a statistically significant amount of data
- **small enough** -> detects actual fluctuations of traffic levels

We selected a **30 minutes** range

BIG DIVE

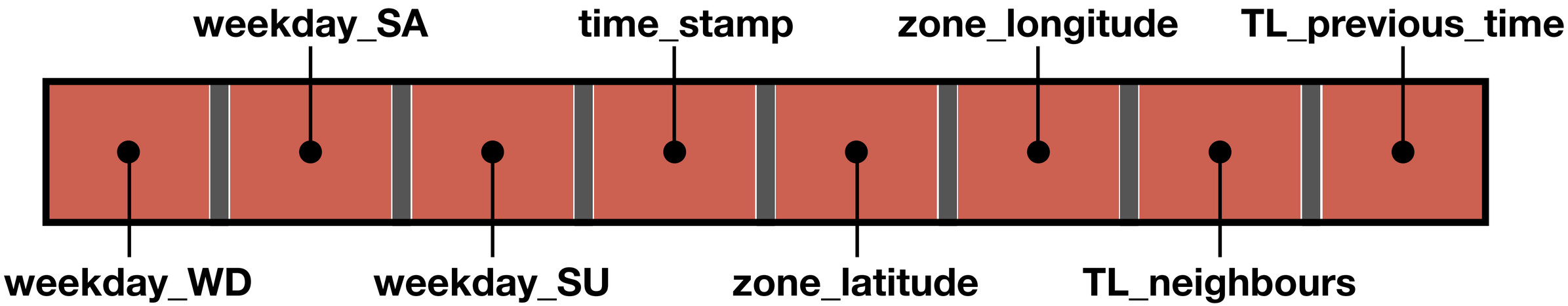
A panoramic of TL values



The values are aggregated for the whole city of Turin.

- TL seems like a **good metric** for traffic levels
- Working days looks very similar to one another

Time for **data preparation**



Data was aggregated by type of weekday, time stamp and zone id: each input vector is a **snapshot** in time

Ready for building the dataset! **But** …

BIG DIVE

The first problem

- Lack of traffic values data for many zones, across a lot of time windows
- Majority of the problematic hours in the range between **1:30 AM** and **7:00 AM**

Solution ➡ keep data for time stamps between **7.00 AM and 1.30 PM**

BIG DIVE

Algorithm selection

- **Regression problem**
- Select an informative and comprehensible metric → **r2 score**
- Baseline model for comparison: always output **traffic level in previous time window**
- Different models for starting point:

Random Forest Regressor

Extremely Randomized Trees Regressor

Extreme Gradient Boosting (XGBoost)

Support Vector Machines

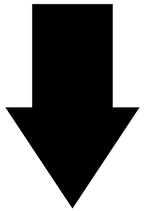
BIG DIVE

The **second problem**

- **Baseline model** score: **0.28**
- After quick hyper-parameter tuning via CV, good r2 for **trained model**: **0.61**

But …

- **Very** poor performance on test set: **0.13** (strange even for overfitting)

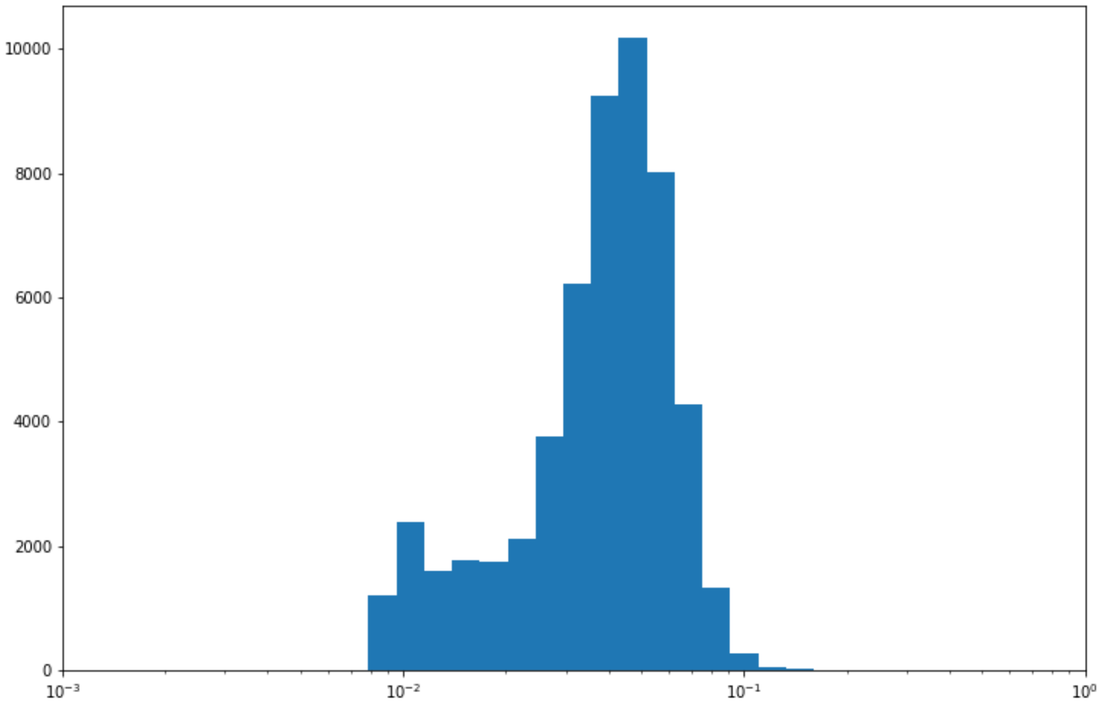


Dataset had **“false” outliers**: TL values computed with **very low number of cars**, **very slow** for some reason (**MSE** heavily weights outliers).

BIG DIVE

Further enhancing… why not?

- All outliers had **number of cars < 7** → discard them
- Lost ~ **1200 records** out of **55450**



- Further improvement: log transform the target TL values → **normal log**

BIG DIVE

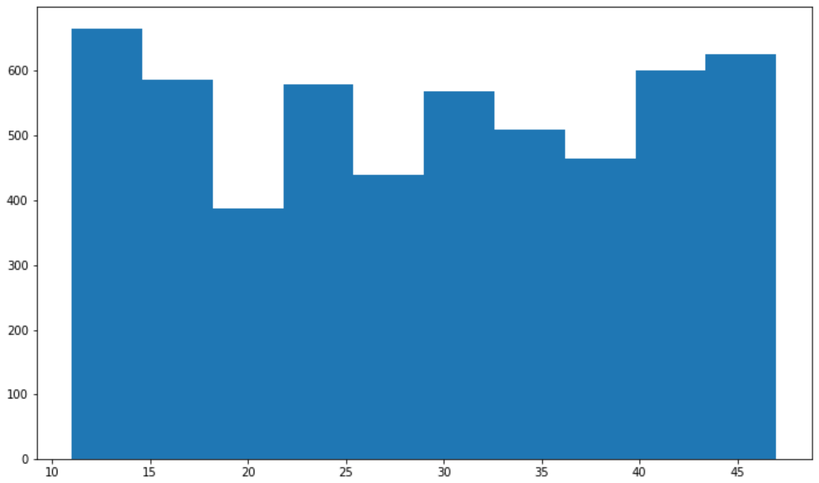
And at last… the results

Trained model (Extree Regressor):

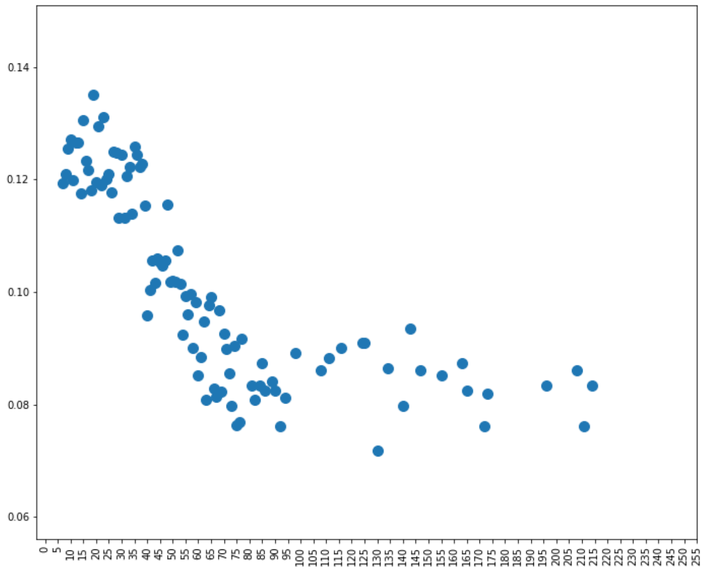
0.917

(**Baseline** model score: **0.535**)

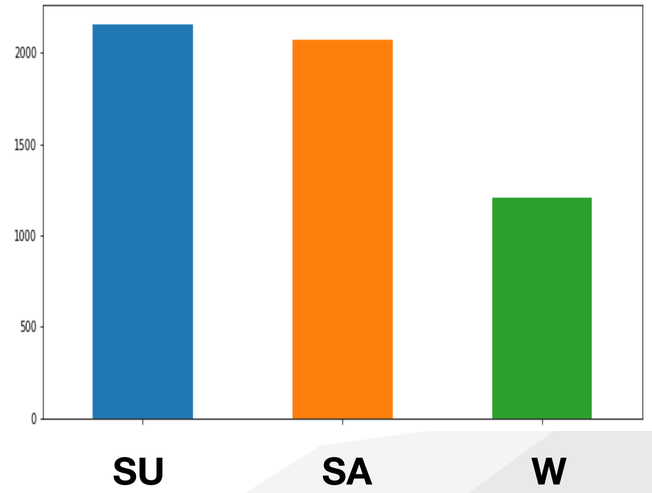
Big error chance by time stamp



Big error chance by car number



Big error count by weekday



Happy ending: **conclusions**

- It seems to be possible to have a nice estimate of traffic levels **for a very specific area**, and with good anticipation

- **Model merit**: by collecting traffic jam values for a certain moment, one can have a good idea of future traffic, and **know it 30 minutes ahead**

Decision tool

Car drivers: “Should I take the car?”,”Should I rethink the route?”

City administration: “Should I do something about public transport?”

BIG DIVE

Epilogue: future developments

- Explore the model with different grid and time window selection (either smaller or bigger)
- Find a way to include discarded hours
- Try with less features (location info maybe?)
- Test performance on unseen months

Thanksgivings

A thanks to all our charming princes:

- **All BIGDIVE staff** for support and patience
- **Alan, Paolo and Andrè** from ISI for very precious suggestions and guidance in very dark hours
- **Laura** for the very nice slide design

BIG DIVE

Thanks for your attention

BIG DIVE