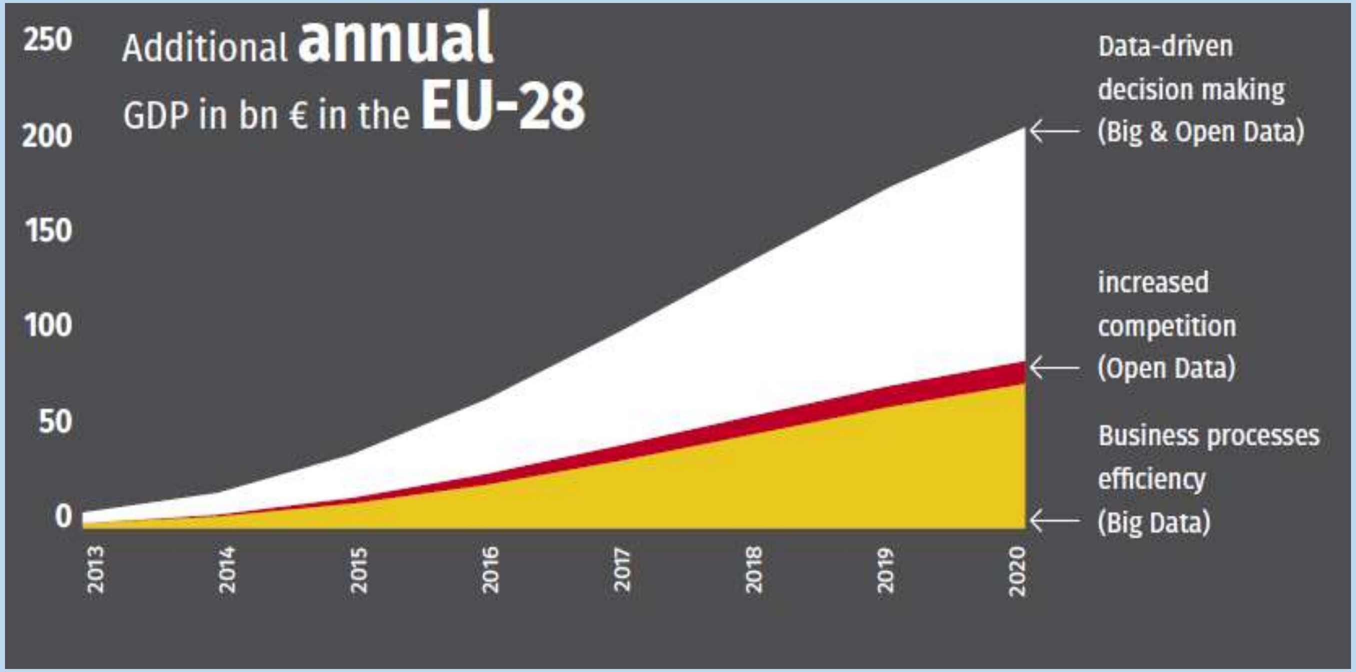
# Data Analytics

**Josep Maria Salanova Grau**

**Neofytos Boufidis**

# Big and Open Data Importance

## *Data: Big & Open (free, reusable & redistributable)*




**Source**: Big and open data in Europe

- *Public*
- *Accessible*
- *Described*
- *Reusable*
- *Complete*
- *Timely*
- *Managed post-release*

# Data Analytics in Transport

## Problem

???

## Solution

## Public Administration

What to?

## Consultant (internal or external)

Scenario 1
Scenario 2
Scenario 3
Scenario 4

What if?

# Decision Support in Transport Policy

- *What to?*
  - *High-level experts*
  - *Experience*
  - *Best practices*
- *What if?*
  - *Models*

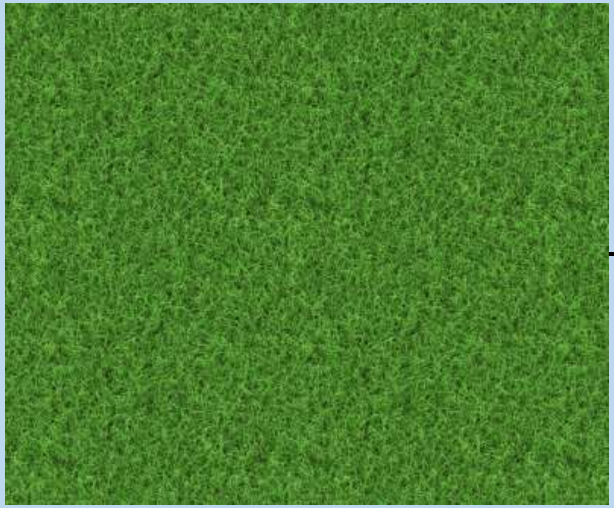
THE GAME OF POLITICS

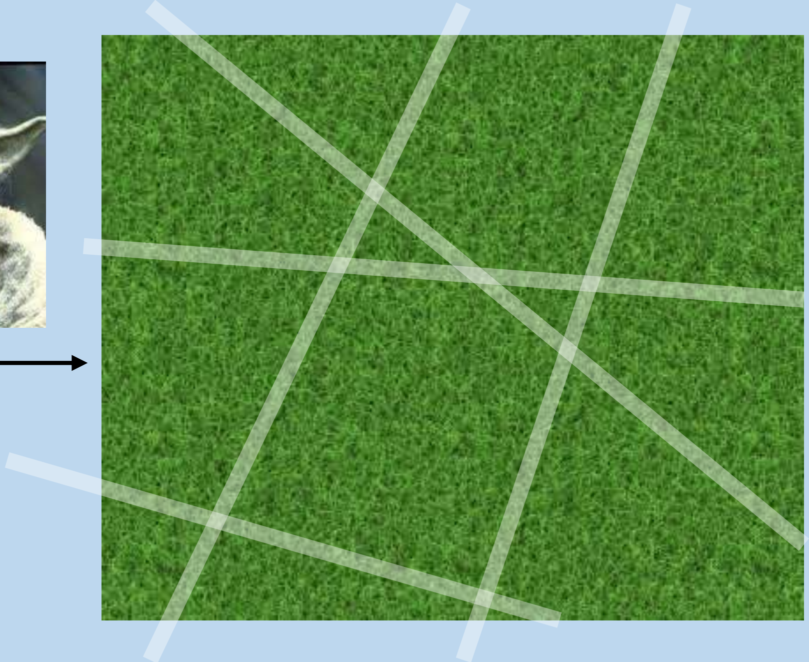
By ROBERT J. AUMANN

Anatol Rapoport, *Fights, Games, and Debates*, Ann Arbor, The University of Michigan Press, 1960, 400 pp. $6.95.

THE last decade has seen a tremendous outpouring of work on the applications of mathematics to the non-exact sciences. In the accepted jargon, this activity is called "model-building"; mathematical models have been built in disciplines ranging from biology and psychology to economics, sociology, linguistics, and even politics. One of the most engaging of these models is the whimsically named "theory of games." More to the point would have been to call it a "theory of conflict and cooperation"; but mathematicians like to christen their babies according to their appearance, not according to the job they are expected to do. Mathematically, wars and games look

# Decision Support in Transport Policy- User generated content





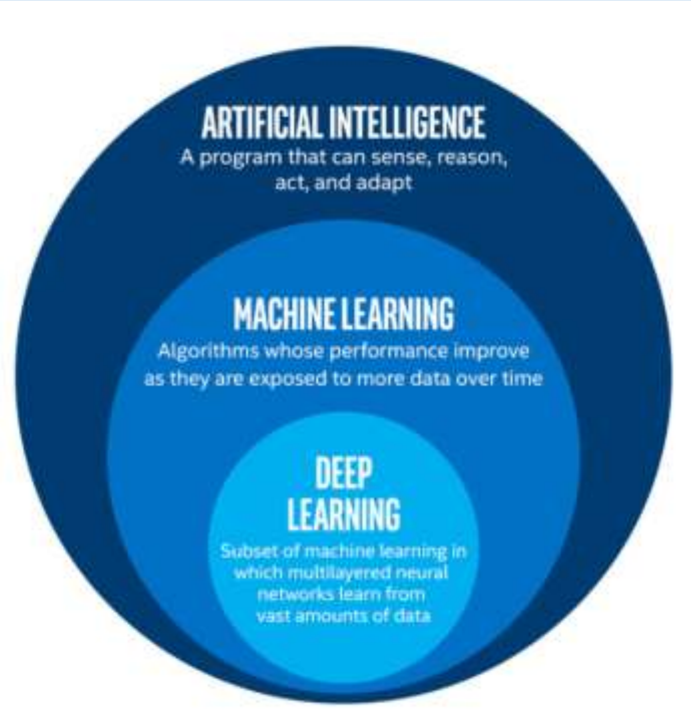
# Data Analytics Workflow


# Data Analytics and AI

## Basic terminology

- **Artificial intelligence** consists of the "thoughts" and conclusions that computers make after receiving data inputs. If human intelligence uses the brain to receive, store, and analyze information, artificial intelligence *uses various technologies to also receive, store, and analyze information.*

- **Machine learning** is a branch within artificial intelligence in which computers use programming as a jumping-off point to create their own processes to <u>analyze vast amounts of data</u>. The machines themselves develop processes and algorithms to take these data and compute observations, trends, and conclusions about the data.
- **Deep leaning** is a subset of machine learning in which multilayer neural networks are able to learn form vast amounts of data.



Co-funded by the Erasmus+ Programme of the European Union

# Big Data and Artificial Intelligence

**Mobility and transport Eco-system**

Mobility service providers
- Traffic info
- In-vehicle services

Data brokers
- Data market
- Automatic contracts

MaaS aggregators
- Ticketing
- Travel planning

**Context-aware smart car**

City authorities
- TMC
- Policy
- PuT

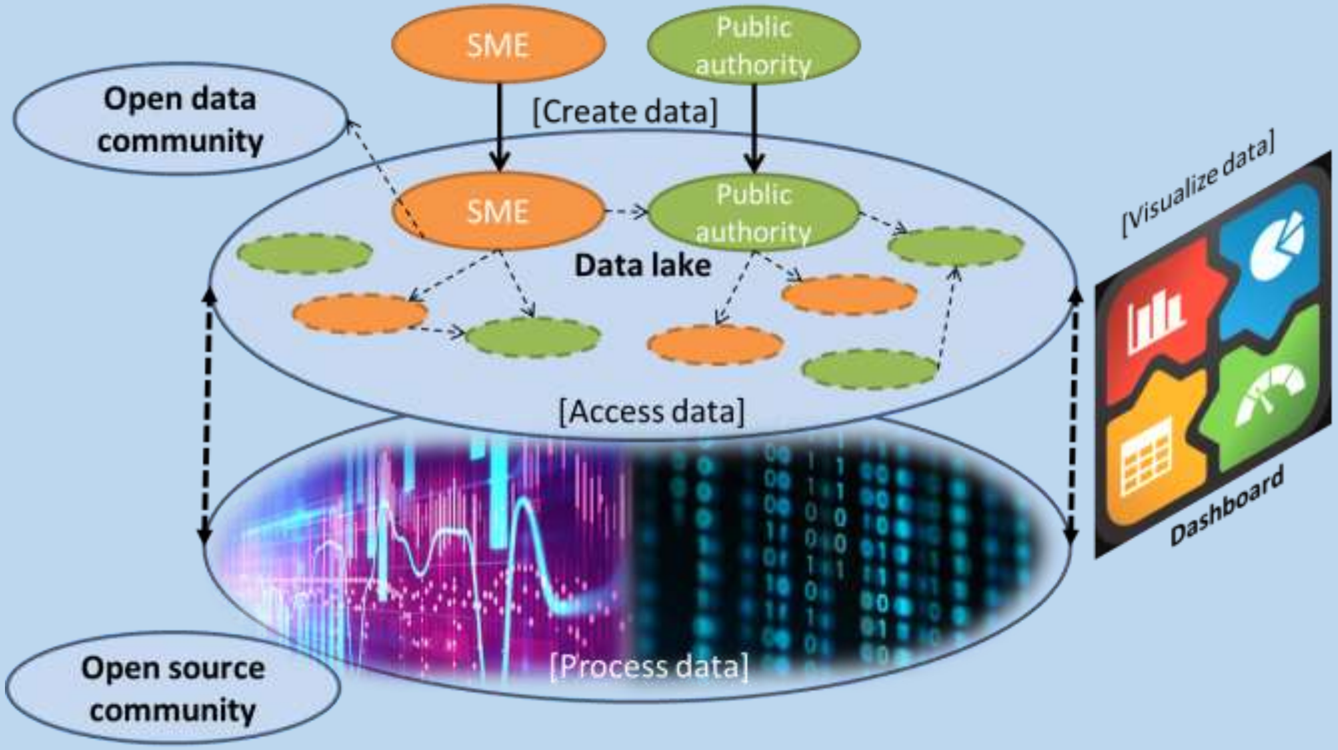
Logistic operators
- Retail services
- Warehouse management

Fleet operators
- Routing
- Maintenance

3rd party mobility providers
- Micro-mobility assets
- Individual PuT

# Thessaloniki mobility and traffic monitoring center

*The project aims at developing a unified platform for fusing, analyzing and processing heterogenous data relevant to the sector of mobility.*



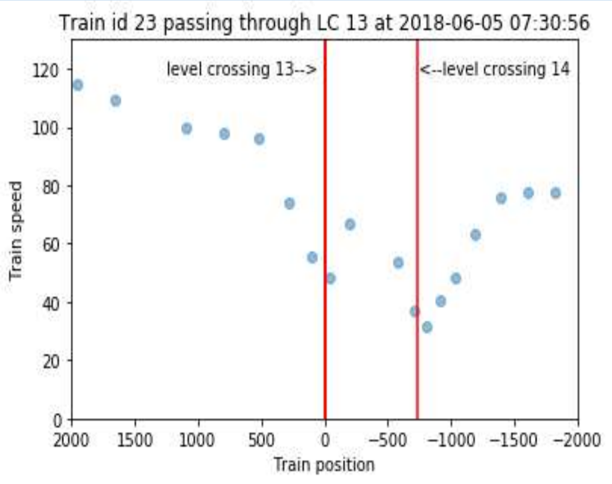
# Estimating Time of Arrival of Trains at Level Crossings (LC) for the Provision of Multimodal Cooperative Services

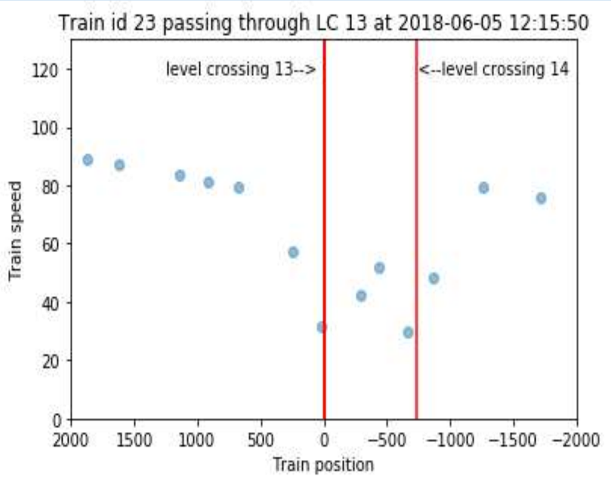
# Data

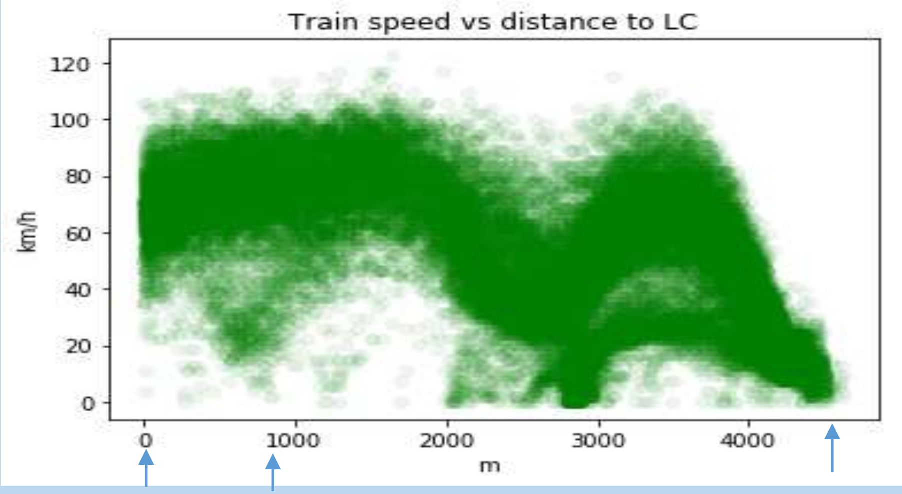
- *Six months of train kinematics data were recorded*
  - *suburban itineraries*
  - *more than 5500 events of trains crossing LCs*
- *Data generated by on-board Global Navigation Satellite System (GNSS) enabled sensors*
  - *coordinates*
  - *instantaneous speed (magnitude and direction)*
  - *timestamp*
  - *metadata (train id, train type, etc)*
- *Recording frequency: 10 seconds*

# Exploratory Data Analysis

- *Atypical train speed profiles around LCs*



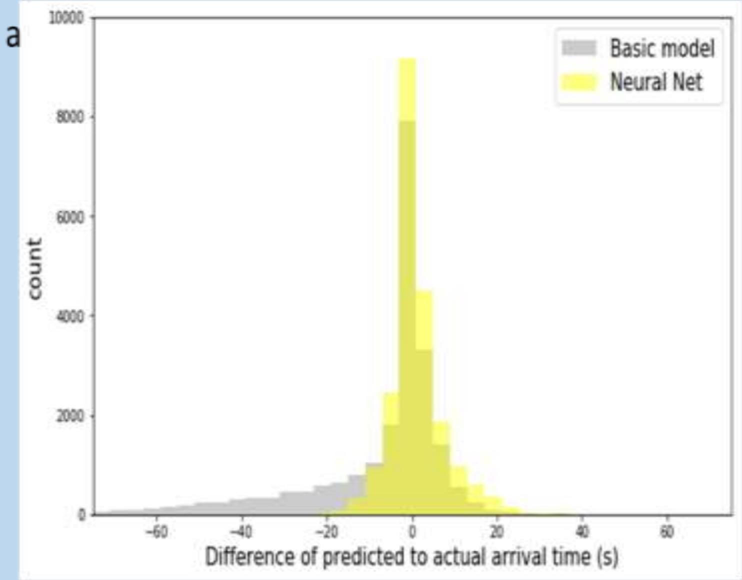



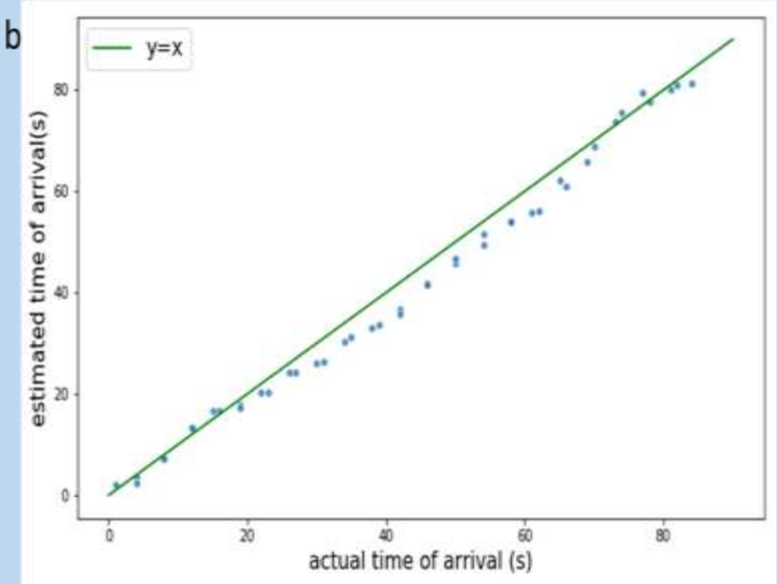

LC 2 LC 1

Thessaloniki Train
Terminal

# Time of Arrival Models

- *Basic (not predictive) model was outperformed by machine learning models*

a


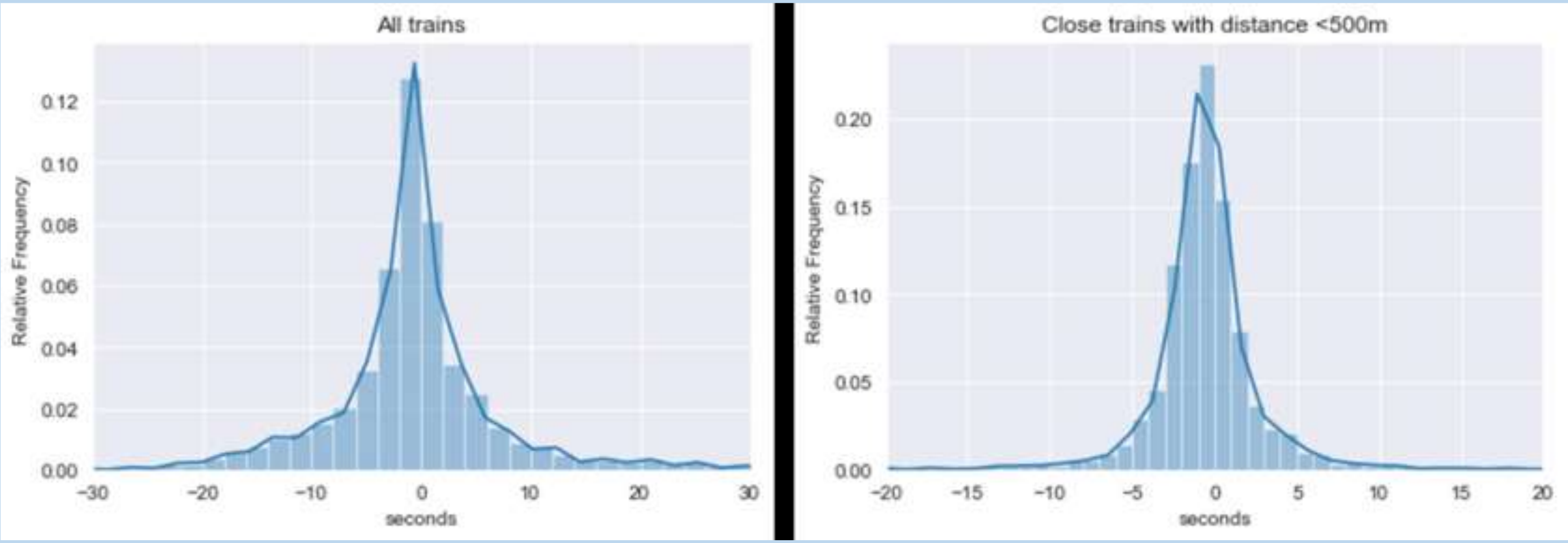

b



  - ***numerous** algorithms were compared*
  - *considered error: **mean absolute error***
  - *model assessment using **unseen test data***
- ***XGBoosting** and **Dense Neural Network** (NN) were the best performing*
  - *after parameter optimization and improvements, the **NN was the most accurate** model*

ADA
ADVANCED DATA ANALYTICS IN BUSINESS
Co-funded by the Erasmus+ Programme of the European Union

# Best Performing Model

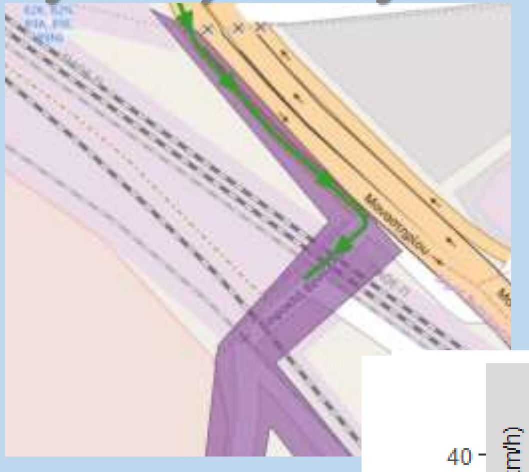
- *Dense Neural Network*
  - *two hidden layers, sigmoid activation function*
  - *p=0.15 dropout and 60 training epochs*
  - *a mean absolute error of under 10 and 5 seconds, when predicting the arrival of trains within a radius of 2000 and 1000 meters, respectively.*

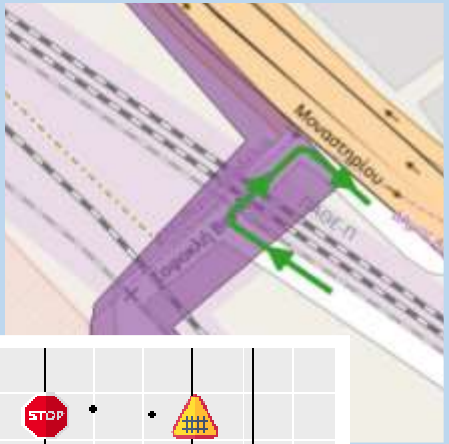


# Impact on Drivers

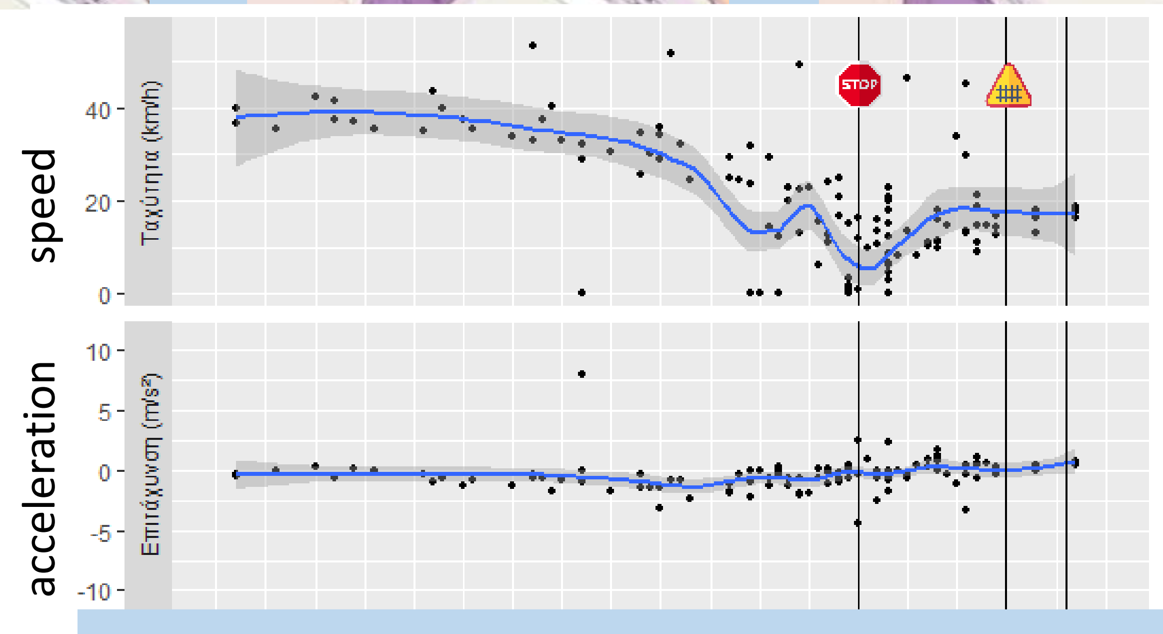
*Analysis on trajectories of taxis around the level crossings.*

*Trajectory classification examples:*





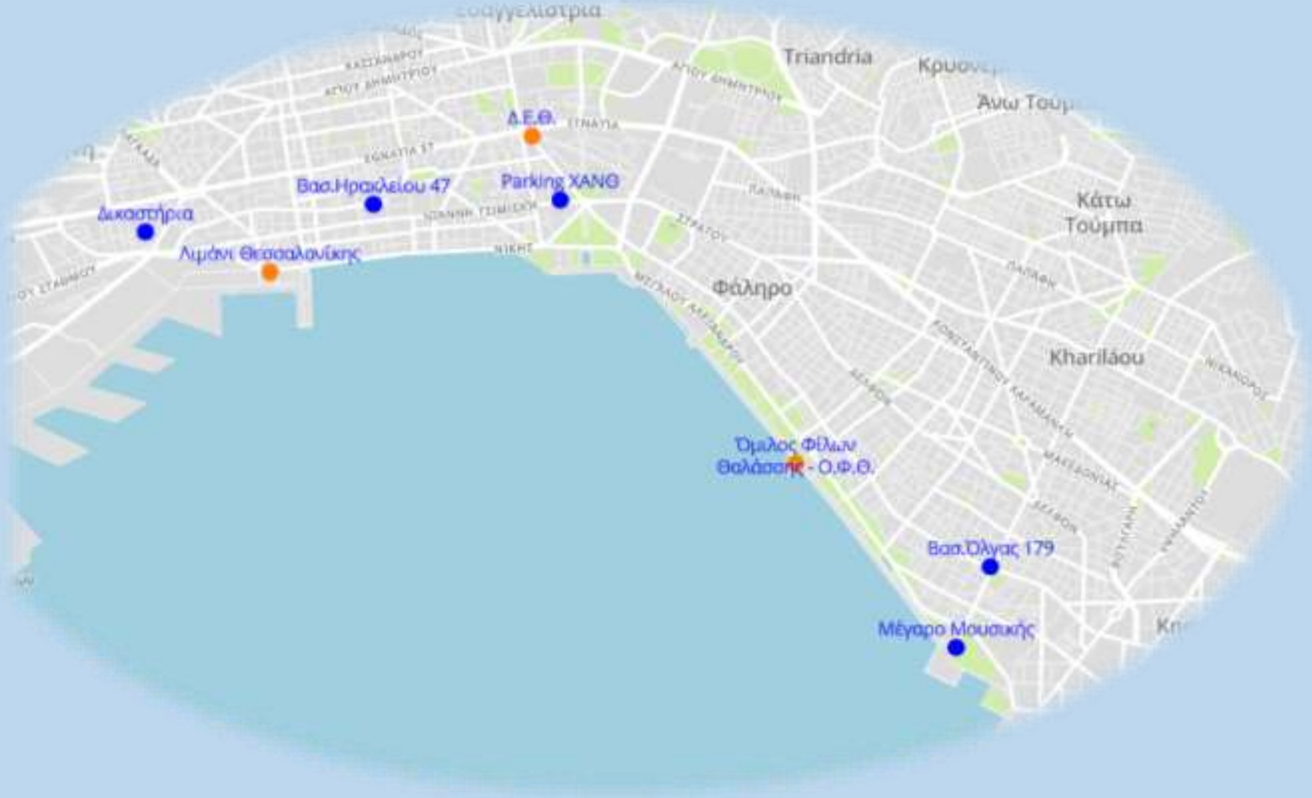
*Analysis results:*



# Development of a station-level demand prediction and visualization tool to support bike-sharing systems' operators

# Introduction and scope

- *study and improve the existing bike sharing system in Thessaloniki, Greece*
  - *5 years of operation – 8 stations – 200 bikes*
- *develop a rentals prediction tool*

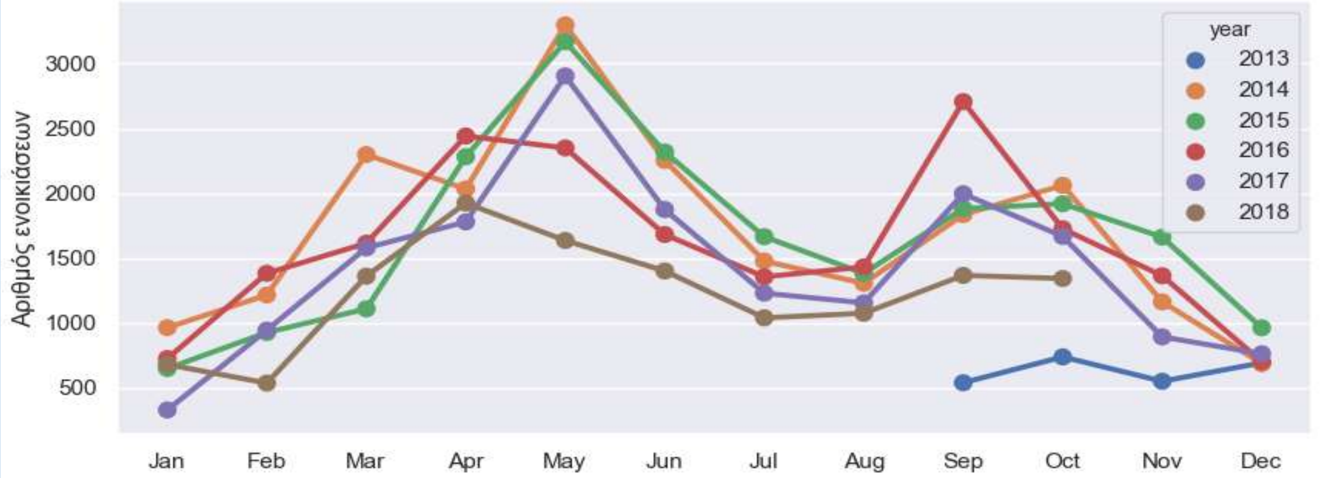


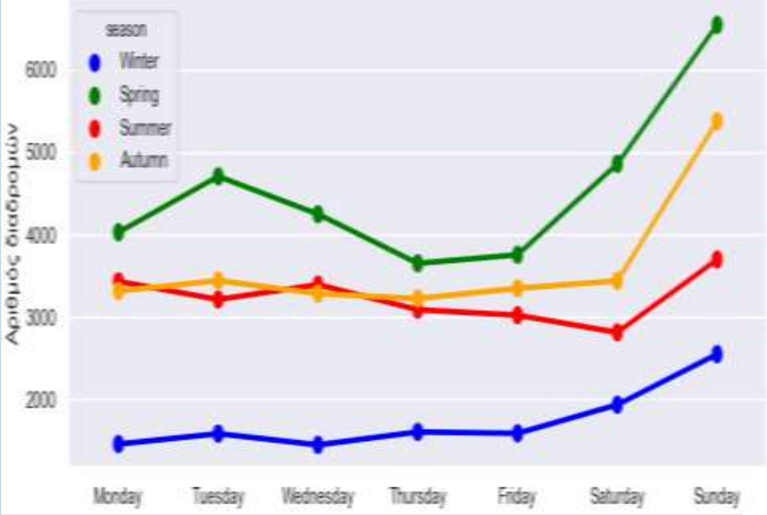
ADA
ADVANCED DATA
ANALYTICS IN BUSINESS

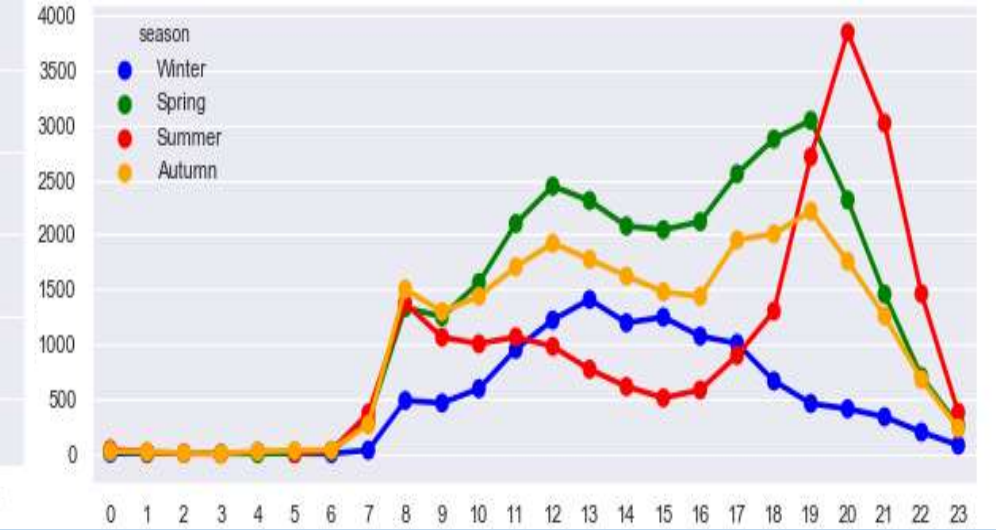
Co-funded by the
Erasmus+ Programme
of the European Union

# Exploratory Data Analysis

## Bike rentals distributions

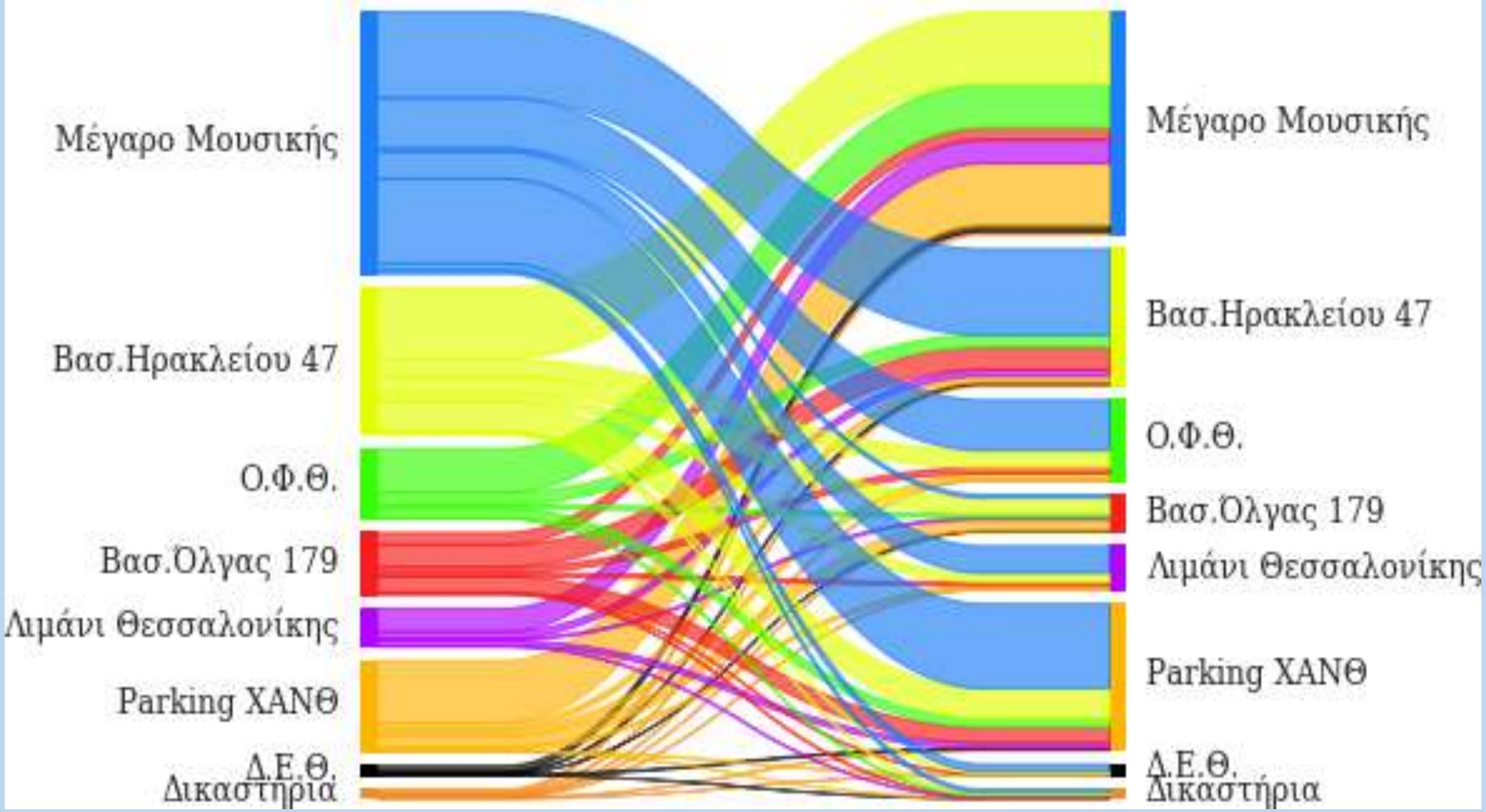


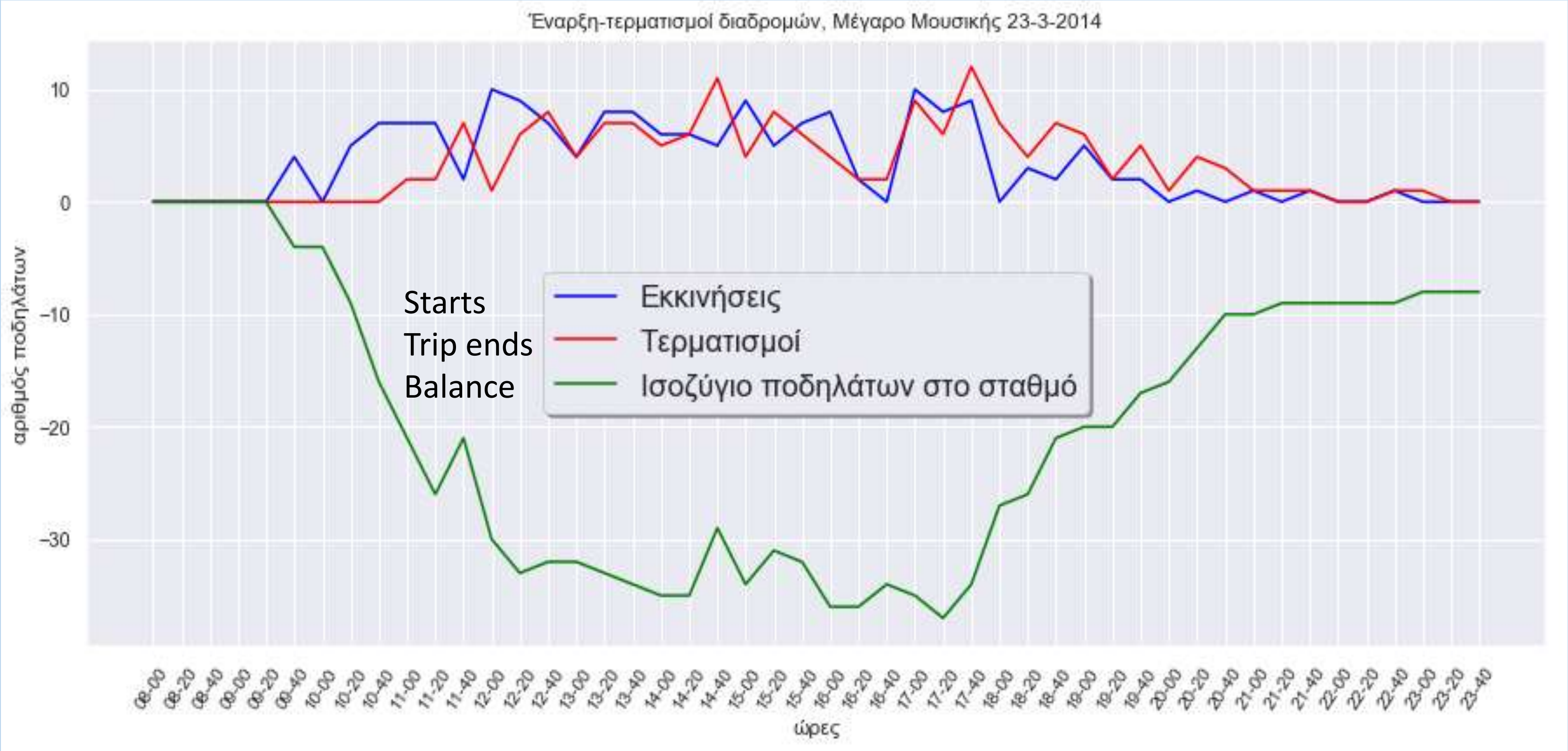
# Exploratory Data Analysis

## *Bike flows amongst the stations*



# Exploratory Data Analysis

## Re-distribution of bikes

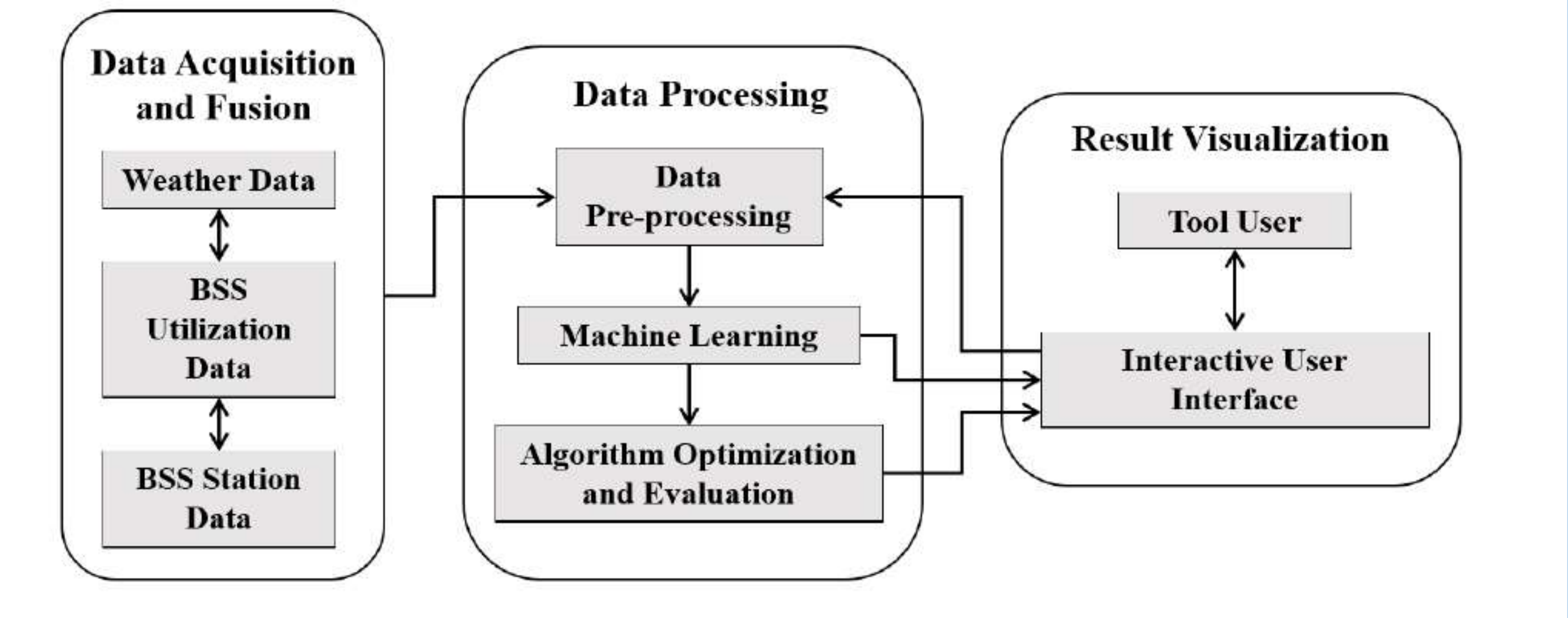


# The Data Analytics Dashboard

*Predictive tool features:*

- *automated training, optimization (grid search cross-validation) and evaluation (10-fold cross-validation of different machine learning algorithms*
- *results on an interactive user interface (dashboard)*

*Tool architecture:*

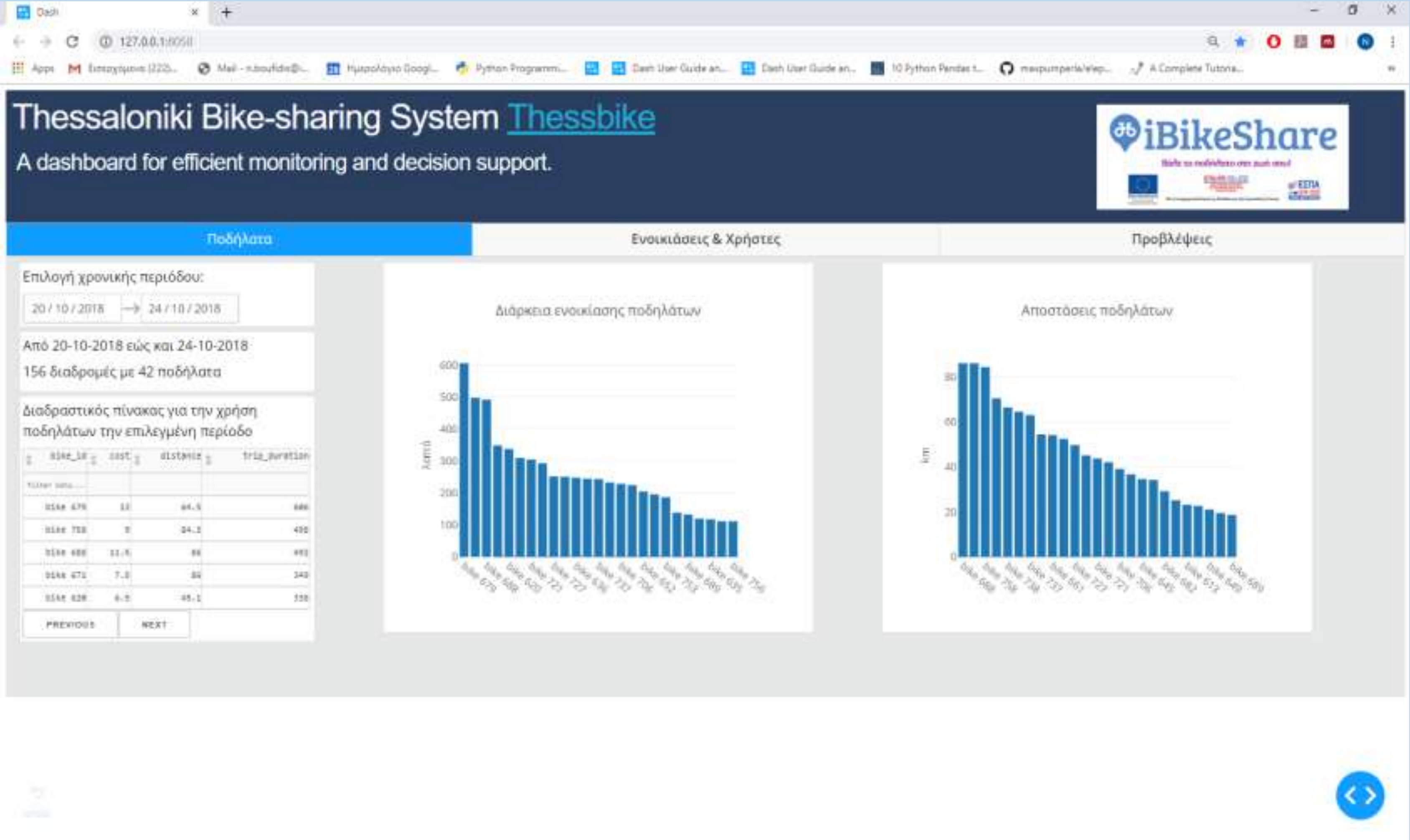


# Model Evaluation Metrics

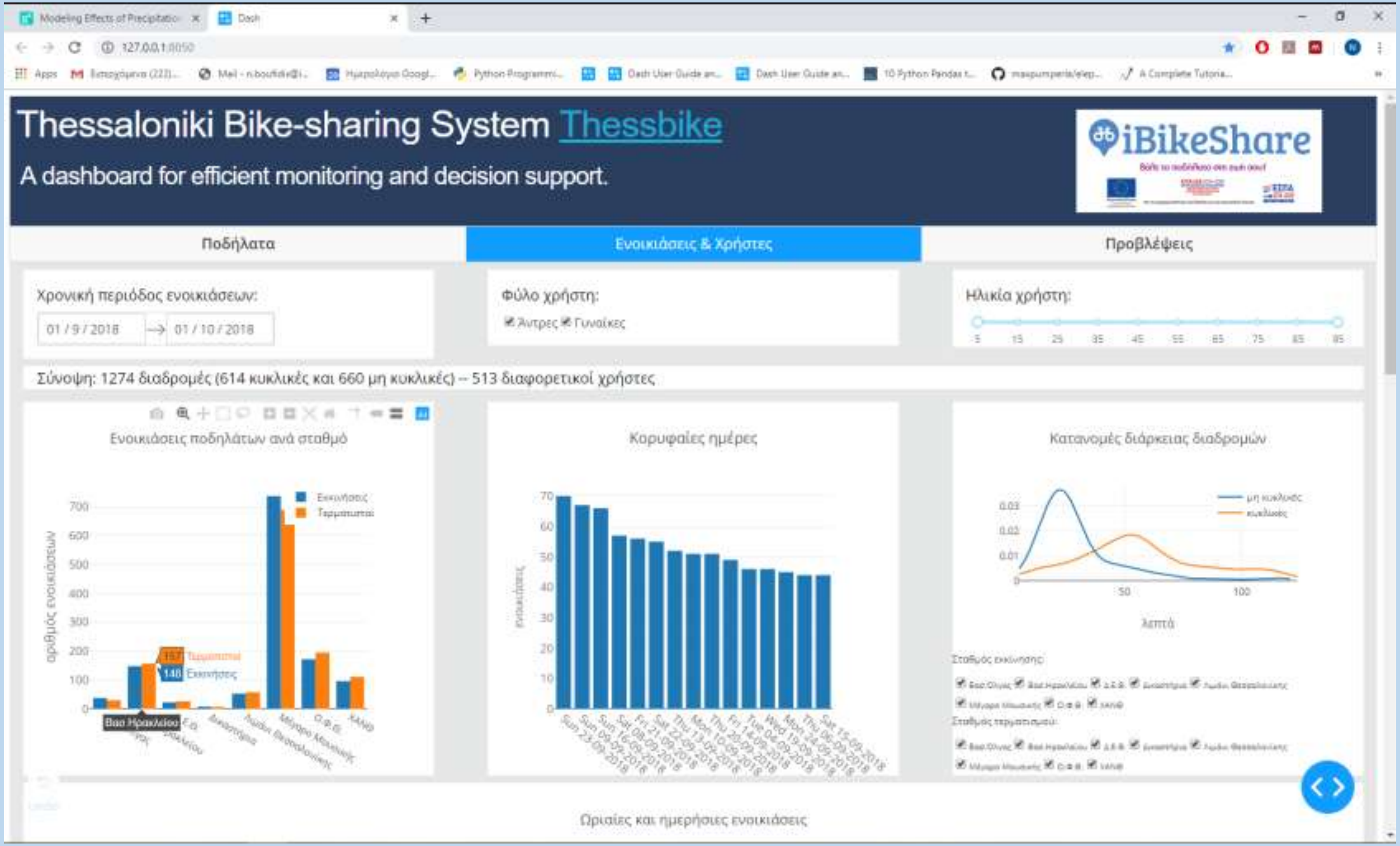
Table 1. Evaluation metrics for all predictive models after GSCV hyperparameter optimization.

| | Train set metrics | | | | Test set metrics | | | |
|---|---|---|---|---|---|---|---|---|
| **Bike rentals** | **MAE** | **MSE** | **RMSLE** | **$R^2$** | **MAE** | **MSE** | **RMSLE** | **$R^2$** |
| Gradient Boosting | 0.75 | 1.81 | 0.42 | 0.76 | 0.85 | 2.69 | 0.46 | 0.64 |
| XGBoost | 0.76 | 1.91 | 0.43 | 0.74 | 0.85 | 2.71 | 0.46 | 0.63 |
| Random Forest | 0.72 | 1.91 | 0.40 | 0.74 | 0.85 | 2.77 | 0.46 | 0.63 |
| Neural Network | 0.89 | 2.66 | 0.49 | 0.64 | 0.91 | 3.00 | 0.49 | 0.6 |
| **Bike returns** | | | | | | | | |
| Gradient Boosting | 0.74 | 1.80 | 0.42 | 0.75 | 0.85 | 2.69 | 0.46 | 0.63 |
| XGBoost | 0.76 | 1.90 | 0.43 | 0.74 | 0.85 | 2.70 | 0.46 | 0.63 |
| Random Forest | 0.71 | 1.90 | 0.40 | 0.74 | 0.84 | 2.76 | 0.46 | 0.62 |
| Neural Network | 0.89 | 2.66 | 0.48 | 0.64 | 0.91 | 3.00 | 0.49 | 0.59 |

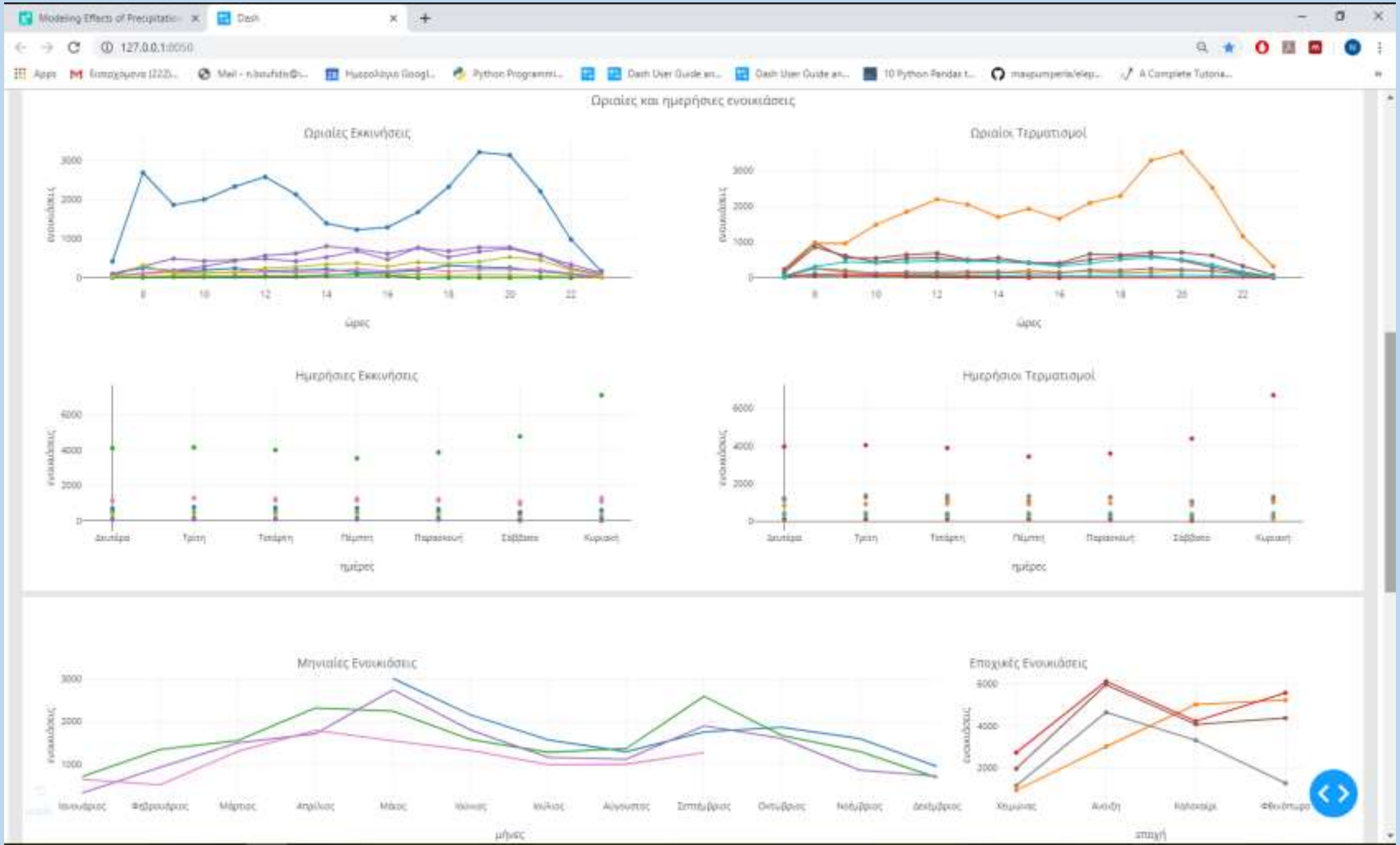
# *Interactive Dashboard*



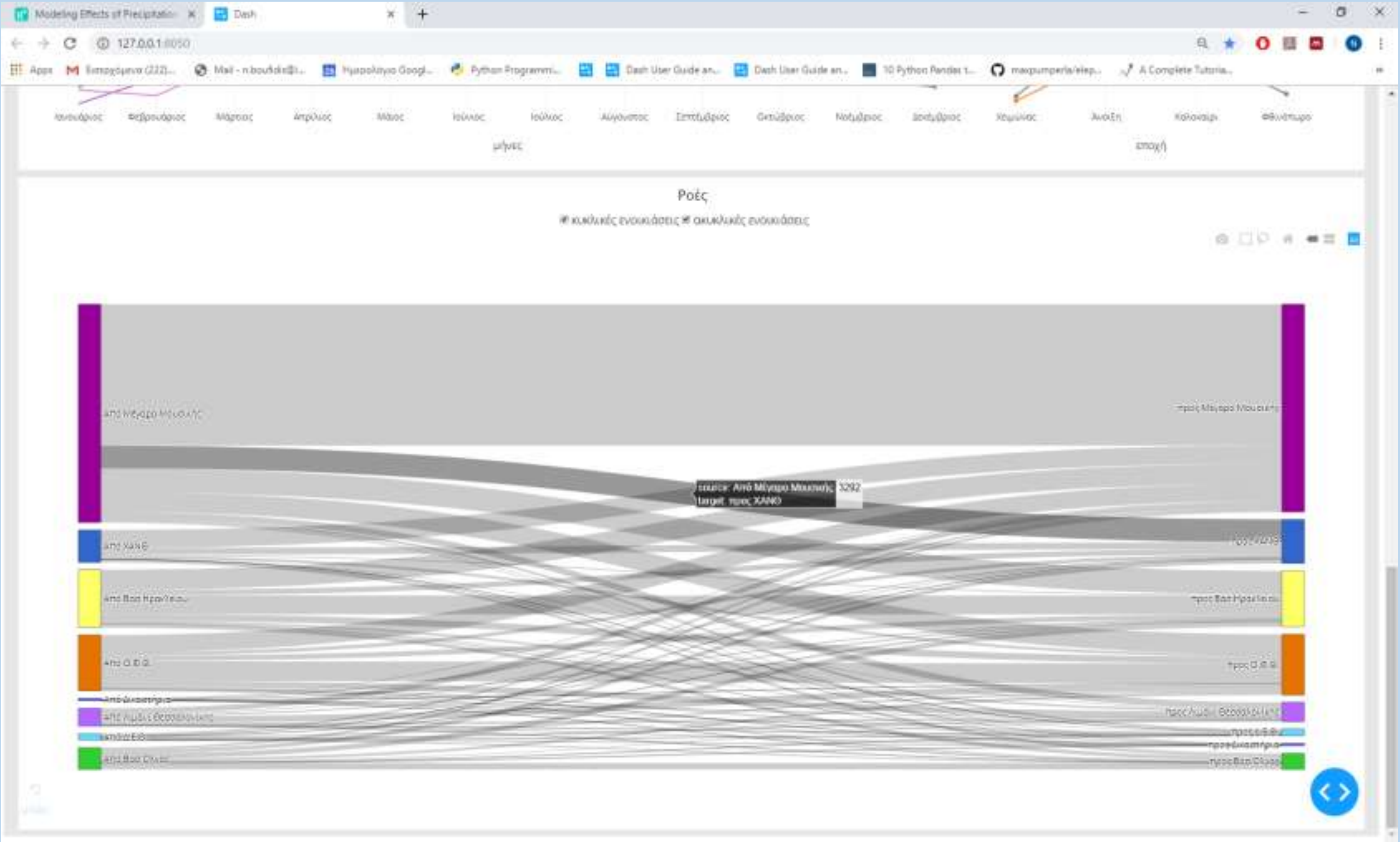
# *Interactive Dashboard*



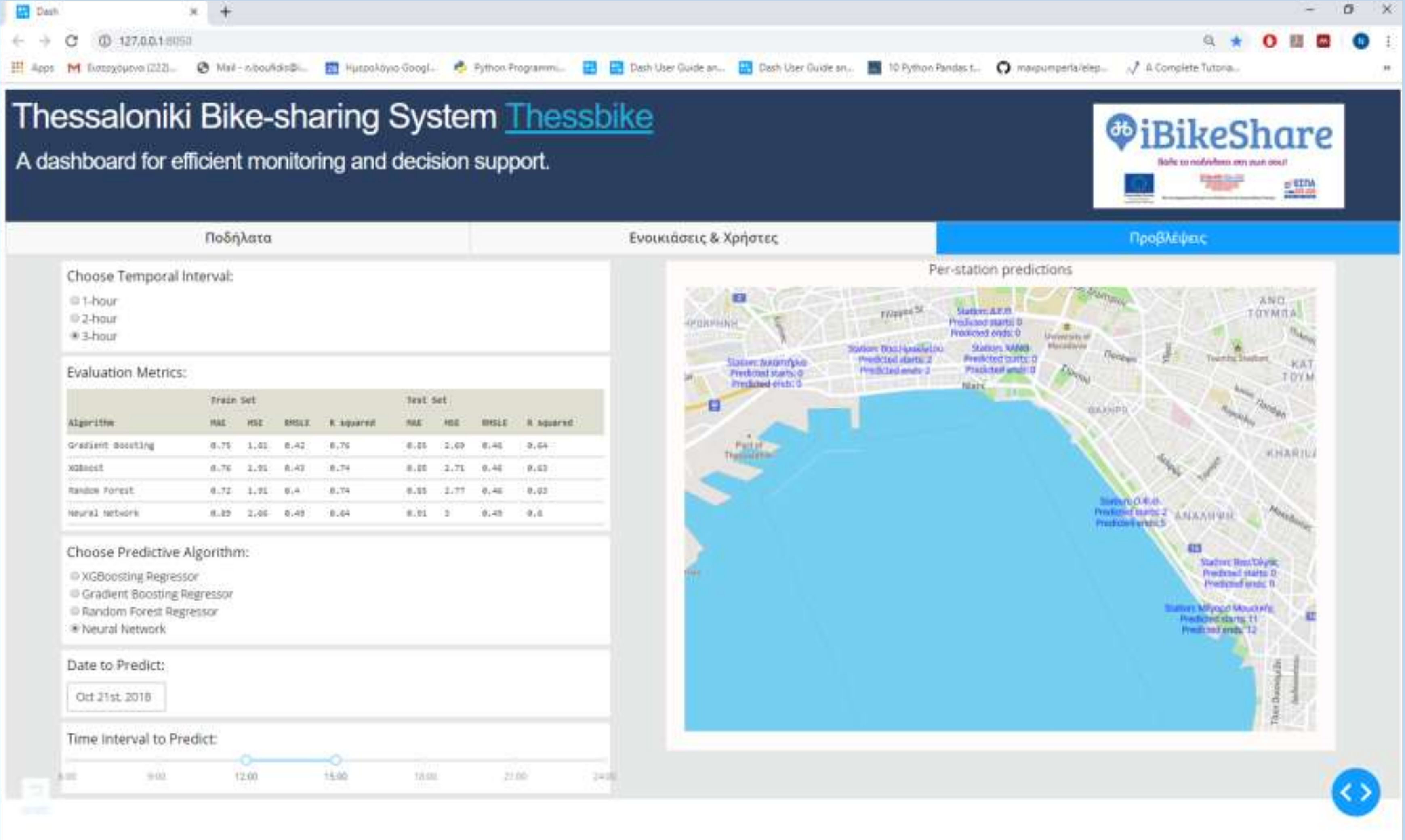
# Interactive Dashboard



# Interactive Dashboard



# *Interactive Dashboard*



# Correlation between digital and physical world, case study in Thessaloniki

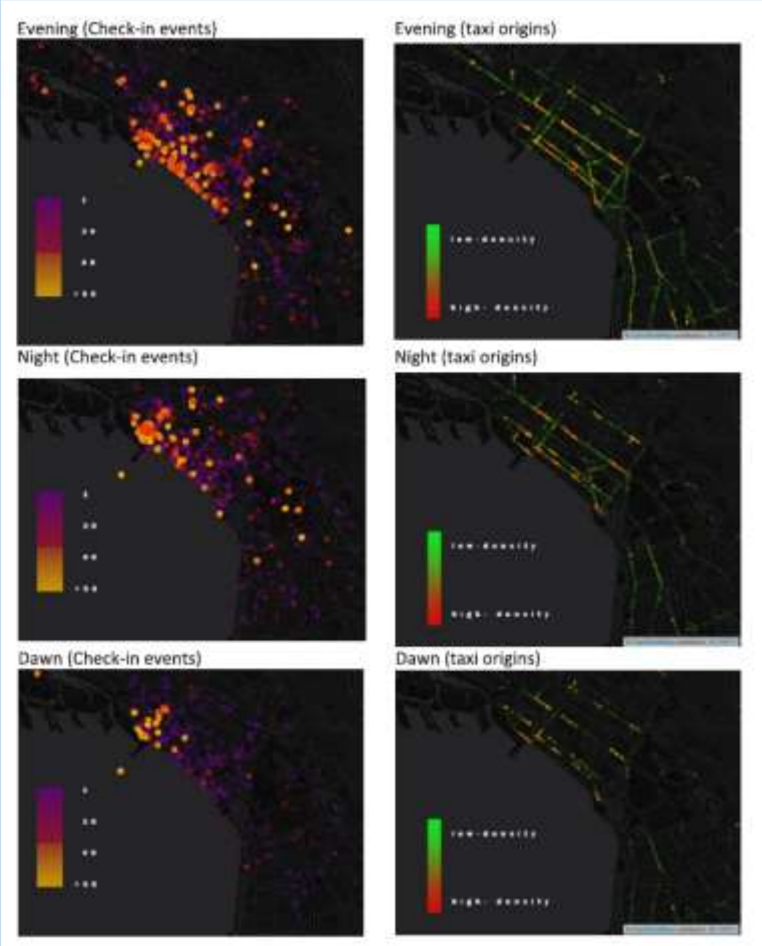
# *Correlation between digital and physical world, case study in Thessaloniki*

*Fusion of heterogenous datasets to study the correlation between activity and transport patterns.*

*Data:*

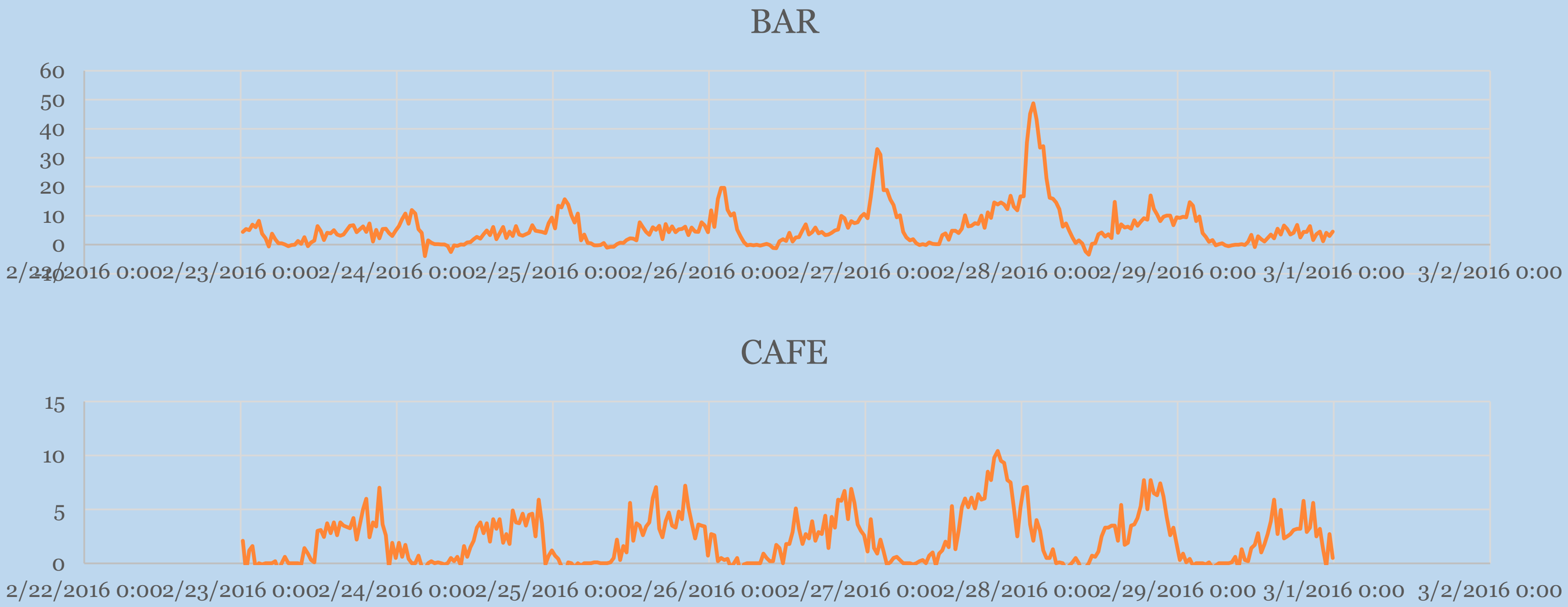
- *FCD from a taxi fleet of 1200 vehicles*
- *check-in events collected from 2951 locations in the city center, generating ~44.000 check-in events per week in the 750 most visited locations with a total of 1265 check-in events during the most active hour*
- *Study period: March 2016*

# Correlation between digital and physical world, case study in Thessaloniki



# More Data Analytics Cases…

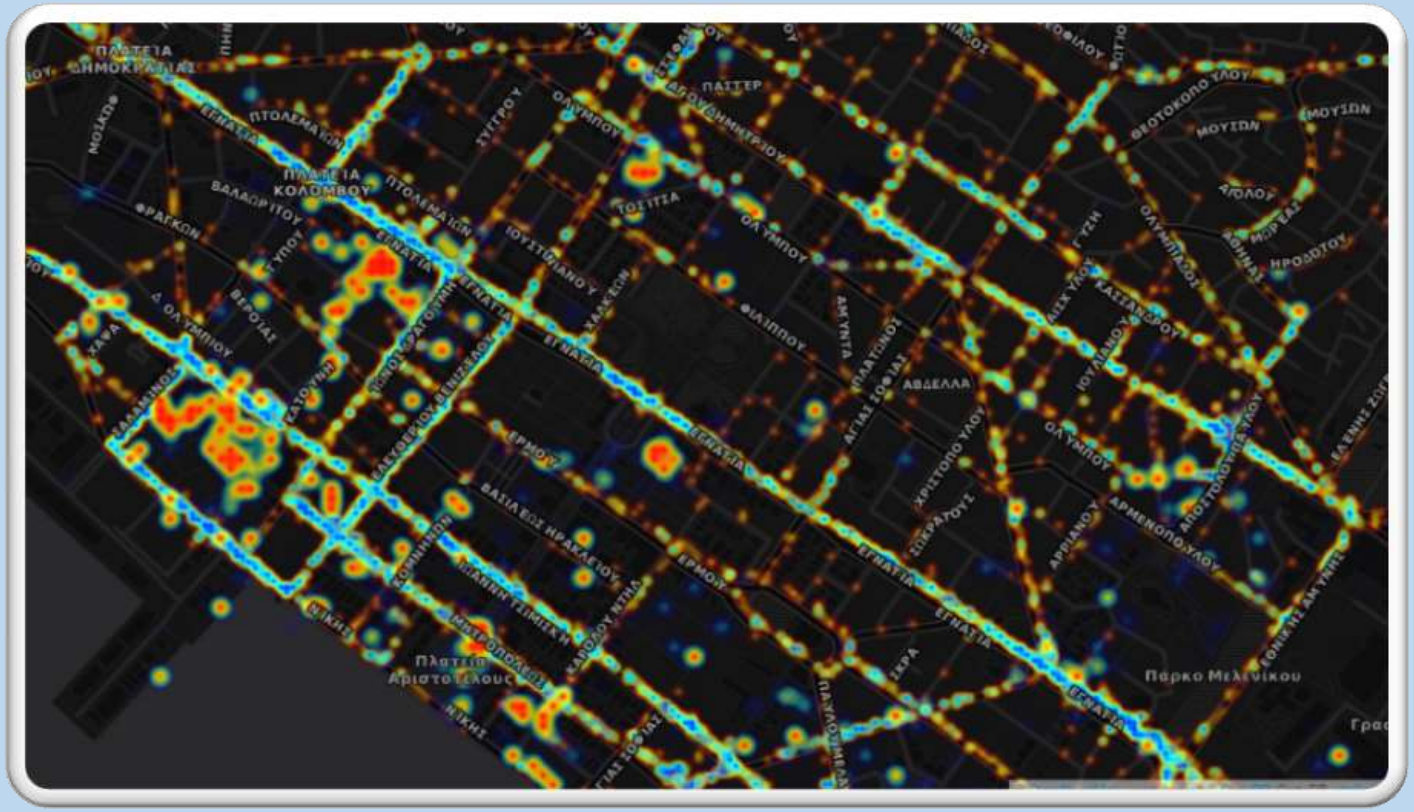
## *Spatial monitoring of Facebook check-ins every 30 minutes.*




Total check-ins by type in a typical Saturday

# Correlation between digital and physical world, case study in Thessaloniki

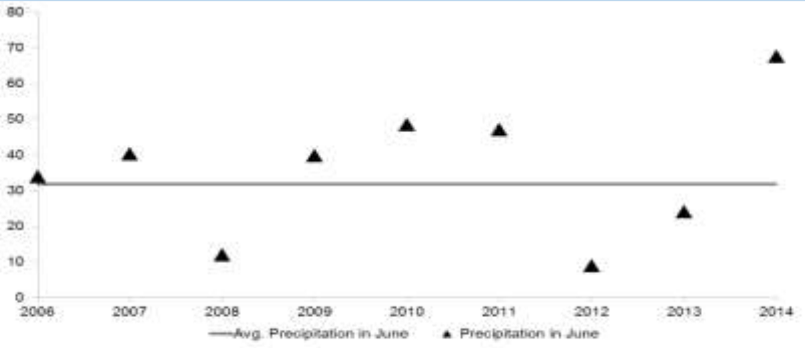
## Concentration of check-in events and taxi trips origins/destinations:

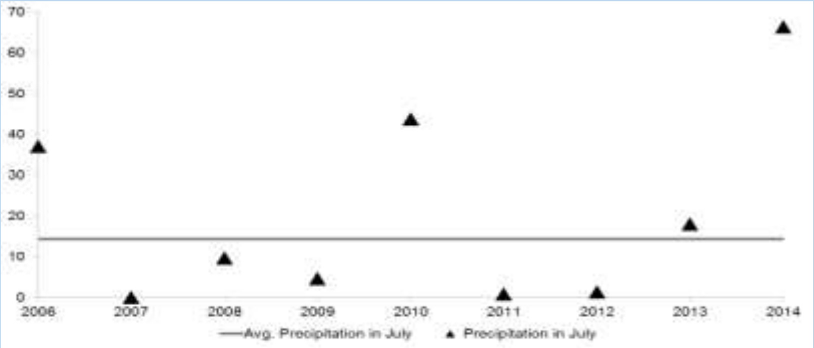


# Modeling the Effects of Precipitation on Vehicle Speed

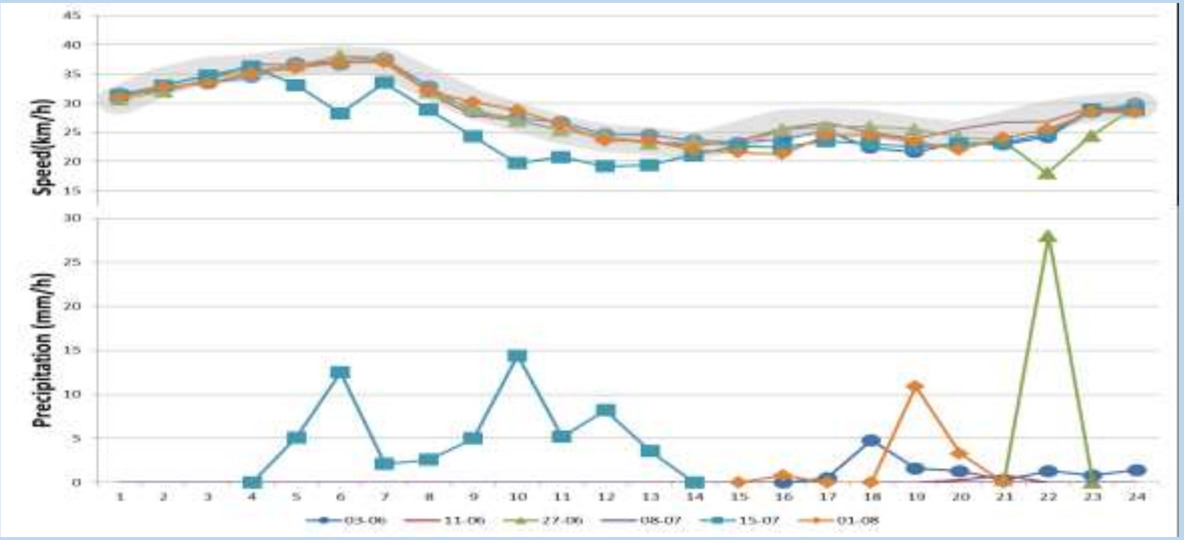
# Modeling the Effects of Precipitation on Vehicle Speed

- *a floating Car Data approach*
- *summer of 2014, a period with extraordinary precipitation*



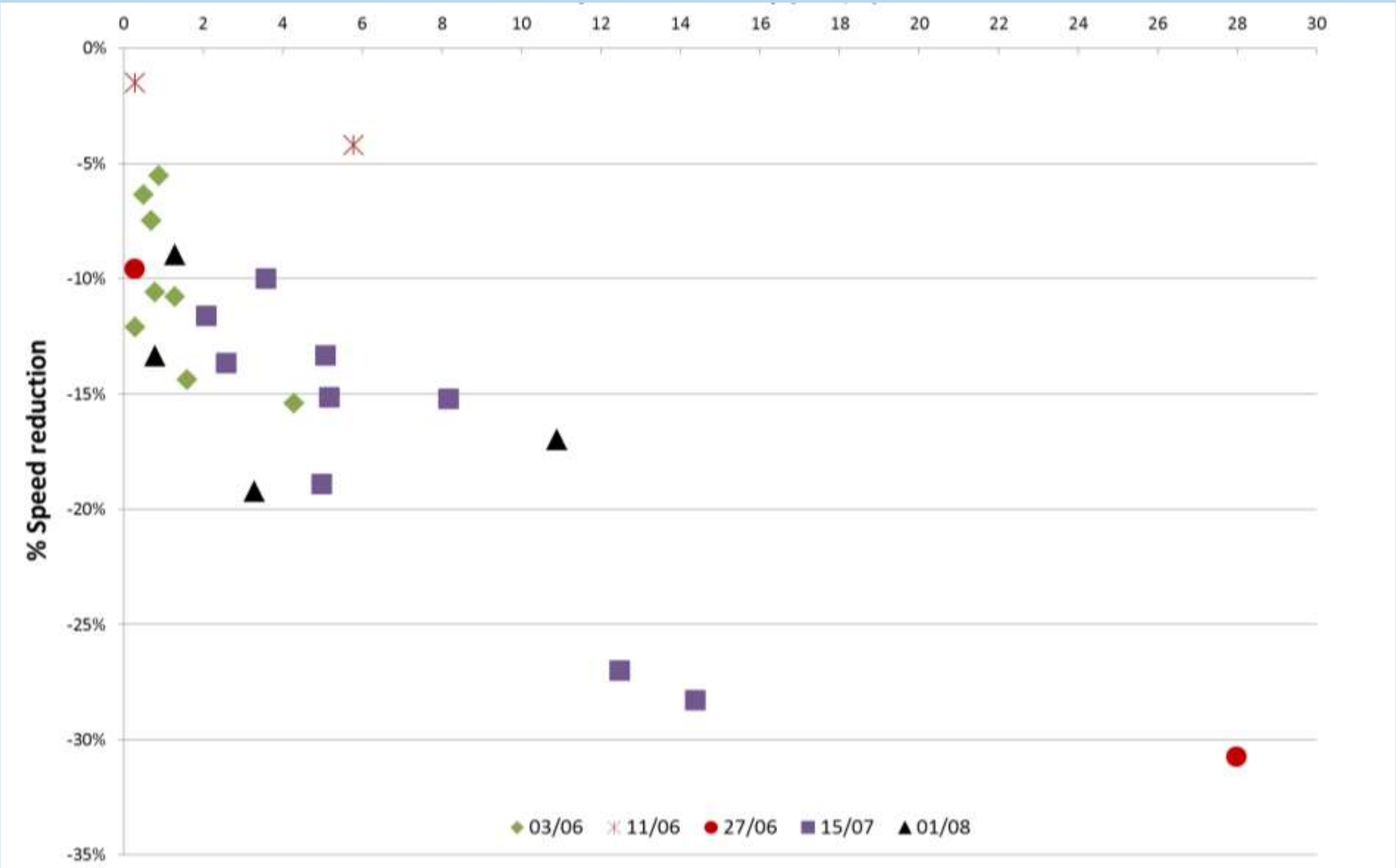




- *avg speed reduction:*



# Modeling the Effects of Precipitation on Vehicle Speed

## Speed reduction in several road sections:



# Modeling the Effects of Precipitation on Vehicle Speed

- *Best fitting model ($R^2$ = 0.78)*

$$v_c = v_0(0{,}984 - 0{,}0563 * \sqrt{i})^2 \ \forall\, i > 0$$

*where:*

*$v_c$: vehicle speed under rain conditions*

*$v_o$: vehicle speed under dry conditions*

*i: precipitation intensity in mm/h*

*…more:*

https://www.researchgate.net/publication/308751627_Modeling_Effects_of_Precipitation_on_Vehicle_Speed_Floating_Car_Data_Approach

ADA
ADVANCED DATA
ANALYTICS IN BUSINESS

CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS

SRPSKA ASOCIJACIJA
MENADŽERA

Co-funded by the
Erasmus+ Programme
of the European Union

# Short-Term Prediction of the Traffic Status in Urban Places Using Neural Network Models

# *Short-Term Prediction of the Traffic Status in Urban Places Using Neural Network Models*

- *A tool for short-term prediction of travel speed on any selected road section*

Table 1. Short description of the inputs in the algorithm.

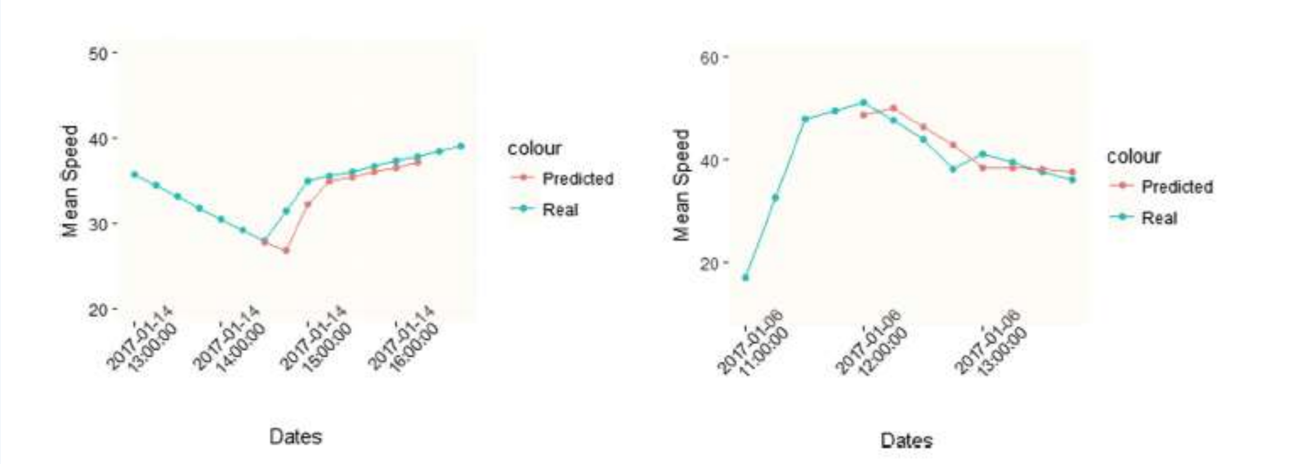
| Input | Description |
|---|---|
| Path | The path with the historical data available |
| Link_id | The Link_id for the road to be predicted |
| Direction | The direction of the road to be predicted |
| Datetime | The time and date for the prediction |
| Steps | How many steps forward will the prediction be |
| Predict | The variable to be predicted. Either "Mean_speed" or "Entries" |

*feature extraction (calculation of input variables for the model)*

| | Min_speed | Max_speed | Stdev_speed | Skewness_speed | Kurtosis_speed | Entries | UniqueEntries | Mean_speed |
|---|---|---|---|---|---|---|---|---|
| 2017-01-16 22:15:00 | 0.04838710 | 0.4915254 | 0.34838094 | 0.5029376 | 0.4147874 | 0.154411765 | 0.42857143 | 0.34693878 |
| 2017-01-16 22:30:00 | 0.32258065 | 0.3898305 | 0.26030382 | 0.5844592 | 0.3162458 | 0.088235294 | 0.24489796 | 0.32653061 |
| 2017-01-16 22:45:00 | 0.11290323 | 0.2542373 | 0.23749948 | 0.2927972 | 0.6167640 | 0.091911765 | 0.28571429 | 0.32653061 |

# Short-Term Prediction of the Traffic Status in Urban Places Using Neural Network Models

- *A neural network model is used to predict the mean speed on the road section of interest*




| Link | Date time | Predicted speed | Real speed | RMSE |
|---|---|---|---|---|
| 1 | 2017-01-12 19:30:00 | 17.07 | 16.71 | 0.35 |
| 1 | 2017-01-12 19:45:00 | 16.88 | 16.14 | 0.74 |
| 1 | 2017-01-12 20:00:00 | 16.02 | 15.57 | 0.45 |
| 1 | 2017-01-12 20:15:00 | 15.69 | 15.00 | 0.69 |
| 2 | 2017-01-14 16:00:00 | 36.75 | 37.28 | 0.53 |
| 2 | 2017-01-14 16:15:00 | 37.26 | 37.85 | 0.59 |
| 2 | 2017-01-14 16:30:00 | 37.77 | 38.42 | 0.65 |
| 2 | 2017-01-14 16:45:00 | 38.31 | 39.00 | 0.68 |

*…more:* https://www.researchgate.net/publication/329578237_Short-Term_Prediction_of_the_Traffic_Status_in_Urban_Places_Using_Neural_Network_Models_Proceedings_of_4th_Conference_on_Sustainable_Urban_Mobility_CSUM2018_24_-_25_May_Skiathos_Island_Greece

# Evaluation framework in Cooperative Intelligent Transport Systems (C-ITS) for freight transport: the case of the CO-GISTICS speed advice service

# Methodology: Festa

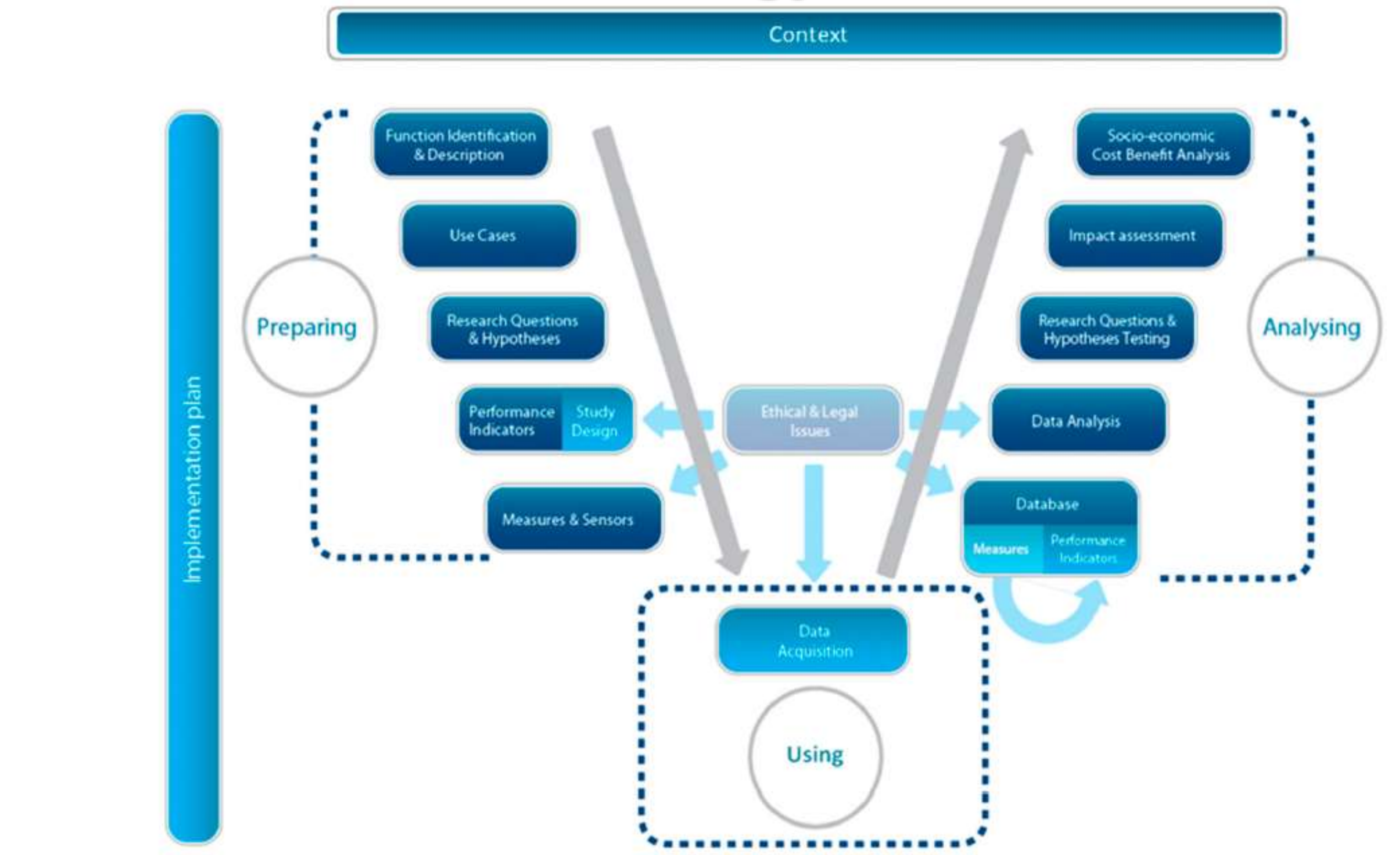
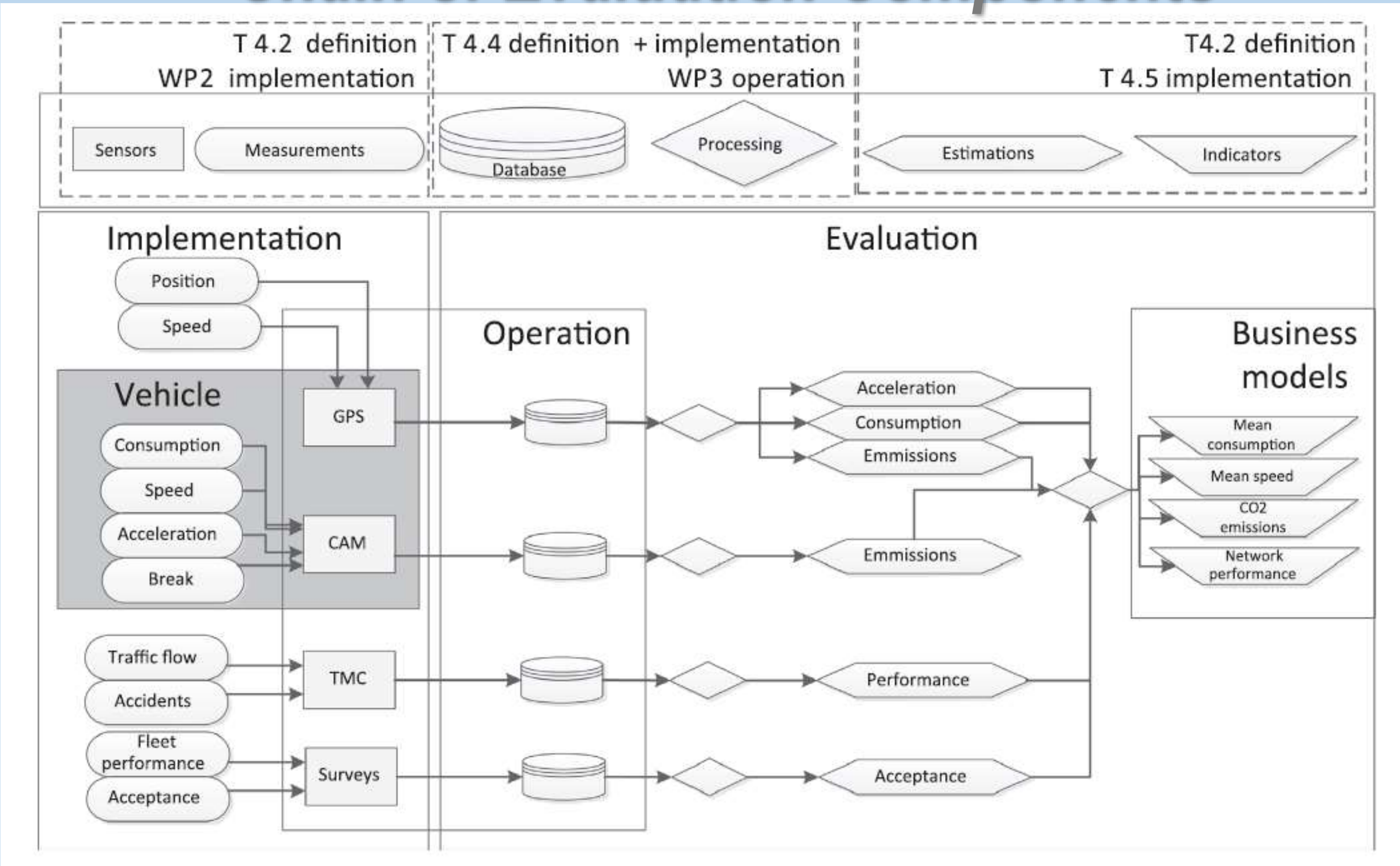



Figure 1. Evaluation methodology proposed by the FESTA handbook [12].

# Chain of Evaluation Components



# KPIs

| Evaluation criteria | KPIs | Description | Unit | Quantitative/ qualitative |
|---|---|---|---|---|
| Network efficiency | Average vehicle speed | Mean speed of the vehicle per route | km/h | Quantitative |
| | Standard deviation speed | Standard deviation of vehicle speed per route | km/h or (m/s) | Quantitative |
| | Approach speed to green traffic light_5 m | Speed at 5 m before a green traffic light | km/h | Quantitative |
| | Approach speed to green traffic light_10 m | Speed at 10 m before a green traffic light | km/h | Quantitative |
| | Approach speed to green traffic light_20 m | Speed at 20 m before a green traffic light | km/h | Quantitative |
| | Approach speed to red traffic light_5 m | Speed at 5 m before a red traffic light | km/h | Quantitative |
| | Approach speed to red traffic light_10 m | Speed at 10 m before a red traffic light | km/h | Quantitative |
| | Approach speed to red traffic light_20 m | Speed at 20 m before a red traffic light | km/h | Quantitative |
| | Maximum acceleration | Peak level of longitudinal or lateral acceleration achieved during a route | $m/s^2$ | Quantitative |
| | Average number of stops and go per route | Mean number of stops and go per vehicle route | Integer | Quantitative |
| | Average travel time (local area) | Mean travel time duration per route in the pilot site area | hh/mm/ss | Quantitative |
| | Average fuel consumption (local area) | Mean fuel consumption of the vehicle in the pilot site area | l/tkm | Quantitative |
| | Average $CO_2$ emissions (local area) | Mean $CO_2$ emissions of the vehicle in the pilot site area | $gCO_2$/tkm | Quantitative |
| | Average fuel savings (local area) | Mean of fuel savings of a CO-GISTICS vehicle per route respect to the situation before CO-GISTICS | % | Quantitative |
| Driver-specific metrics | Perceived system usefulness and consequences | Perception of the potential usefulness and of the benefits deriving by the use of a CO-GISTICS system | Six-point rating scale | Qualitative and Quantitative |
| | Driving behaviour | Questionnaire able to assess variations during the normal driving (perception of frequency of lapses, errors, violations, etc.) | Five-point rating scale | Qualitative and Quantitative |
| | Customer satisfaction | Customer satisfaction related to a CO-GISTICS service | Five-point rating scale | Qualitative |

# Quantitative KPIs Formulas

| KPI | Formula | Reference |
|---|---|---|
| Total distance per route | $Distance\,(km) = \sum_{t=1}^{t=n-1} Haversine[pos(t), pos(t+1)]$ | [16] |
| Total duration per route | $Duration(minutes) = Duration(t, t+1)$ | [4] |
| Average speed of a truck in a route | $Average\ Speed\,(km/h) = \frac{\sum_{t=1}^{t=n-1} Haversine[pos(t), pos(t+1)]}{Duration(t,t+1)}$ | [4] |
| Instantaneous speed | $Speed\,(km/h) = \frac{Haversine[pos(t), pos(t+1)]}{Duration(t,t+1)} for\ t\ in 1, \ldots, t-1.$ | [4] |
| Truck speed standard deviation in a route | $Speed\ Standard\ deviation(\%) = \sqrt{\frac{\sum_{t=1}^{t=n-1} [Speed(t) - AverageSpeed]^2}{n\ GPS points\ in\ a\ trip}}$ | [12] |
| Instantaneous acceleration | $Acceleration(m/s^2) = \frac{Speed(t,t+1) - Speed(t-1,t)}{Duration(t,t+1)} for\ t\ in 1, \ldots, t-2.$ | [4] |
| Average fuel consumption of a vehicle per route | $Av.\ fuel\ consumption\ per\ route\ (MJ) = \frac{\sum_{t=1}^{T} Total\ energy consumption_{(t)}}{T}$ | [14] |
| Average $CO_2$ emissions of a vehicle per route | $Av.\ CO_2\ emissions\ per\ route\ (kg\ CO2) = \frac{\sum_{t=1}^{T} CO2\ emissions\ per\ route_{(t)}}{T}$ | [4];[9] |
| Average fuel savings | $Av.\ Fuel\ Savings\ (\%) = \frac{Average\ Fuel\ Consumption_{operation}(t)}{Average\ Fuel\ Consumption_{baseline}} * 100$ | [3] |

[No Title]

Co-funded by the Erasmus+ Programme of the European Union

# Freight transport patterns extraction using Floating Car Data, case study in Thessaloniki

# *Dataset*

- ***GPS pulses***
- ***383 truck vehicles** in total*
- *duration: **2 months***
- *average of **4 pulses per minute***
- *location: **prefecture of Thessaloniki***

*After **data cleansing**, analysis is conducted on 90 vehicles, for ~ 4000 routes*

# KPIs for Trips

- *mean distance per trip* → 35 km
- *mean # of trips per day* → 1,9 trips/day
- *mean trip duration* → 3 hours and 15 minutes
- *mean loading time* → 52 minutes
- *mean time of vehicle immobility per trip* → 82 min
- *mean # of stops per trip* → 4 stops
- *mean percentage of time moving* → 57%
- *mean # of stops over 5/10 minutes* → 2,2 stops/ 1,7 stops

# KPIs for Trips Entering the City Center

- *mode time of entrance* → 9:20:00
- *mode time of exit* → 15:40:00
- *mean duration spent in city center* → ~ 1 ώρα
- *mean trip time* → 4:30 hours
- *mean # of stops in city center* → 2,2 stations
- *mean # of stops per trip* → 4 stops
- *mean trip distance* → 40 km
- *percentage of vehicles in this category* → 20%

ADA
ADVANCED DATA ANALYTICS IN BUSINESS

Co-funded by the Erasmus+ Programme of the European Union

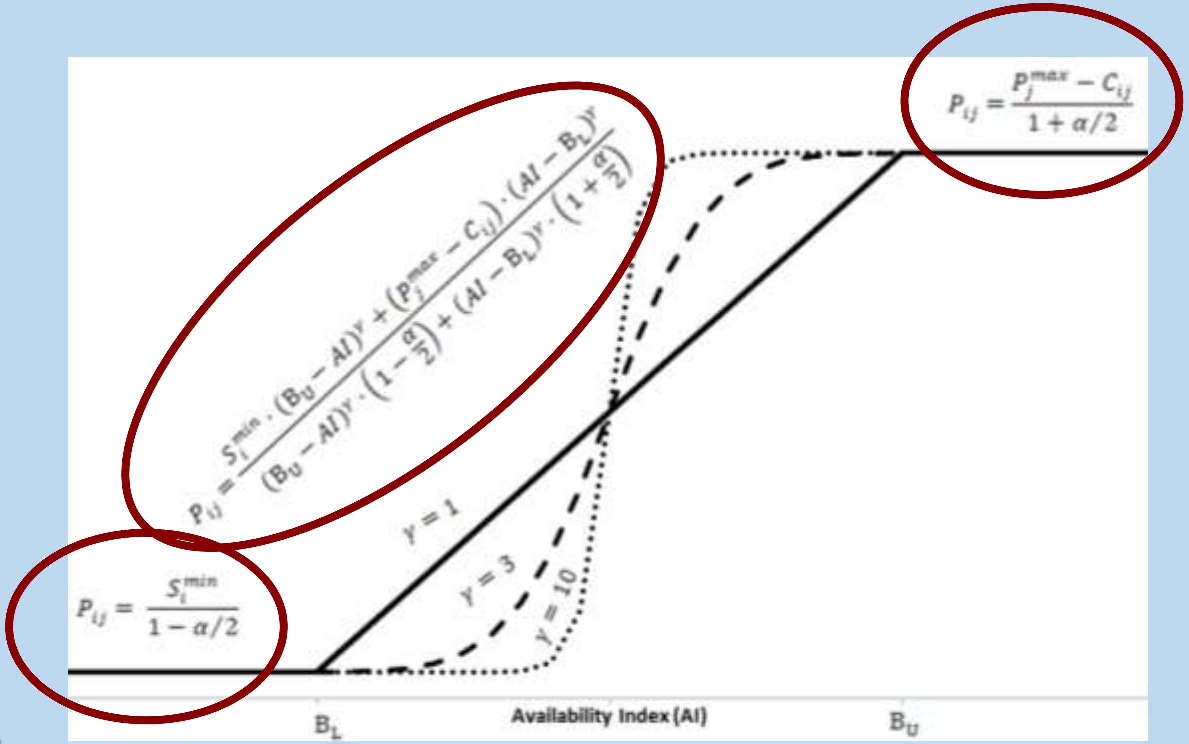
# DEVELOPMENT OF A “FAIR” MARKETPLACE FOR ON-DEMAND CAPACITY MATCHING

# Introduction and scope

- *Independent e-market place (MP)*
- *Matching manufacturers to warehouses*
  - ✓ *splitting the benefit under a "fairness rule", according to market status*
- *Parameters considered:*
  - ✓ *min sell price defined by warehouse*
  - ✓ *max buy price defined by manufacturer*
  - ✓ *MPs commission fee rate*
  - ✓ *transaction cost per match*
  - ✓ *Availability index "AI" (demand to total supply ratio)*
  - ✓ *Manufacturer-warehouse distance*

Co-funded by the
Erasmus+ Programme
of the European Union

# The Marketplace

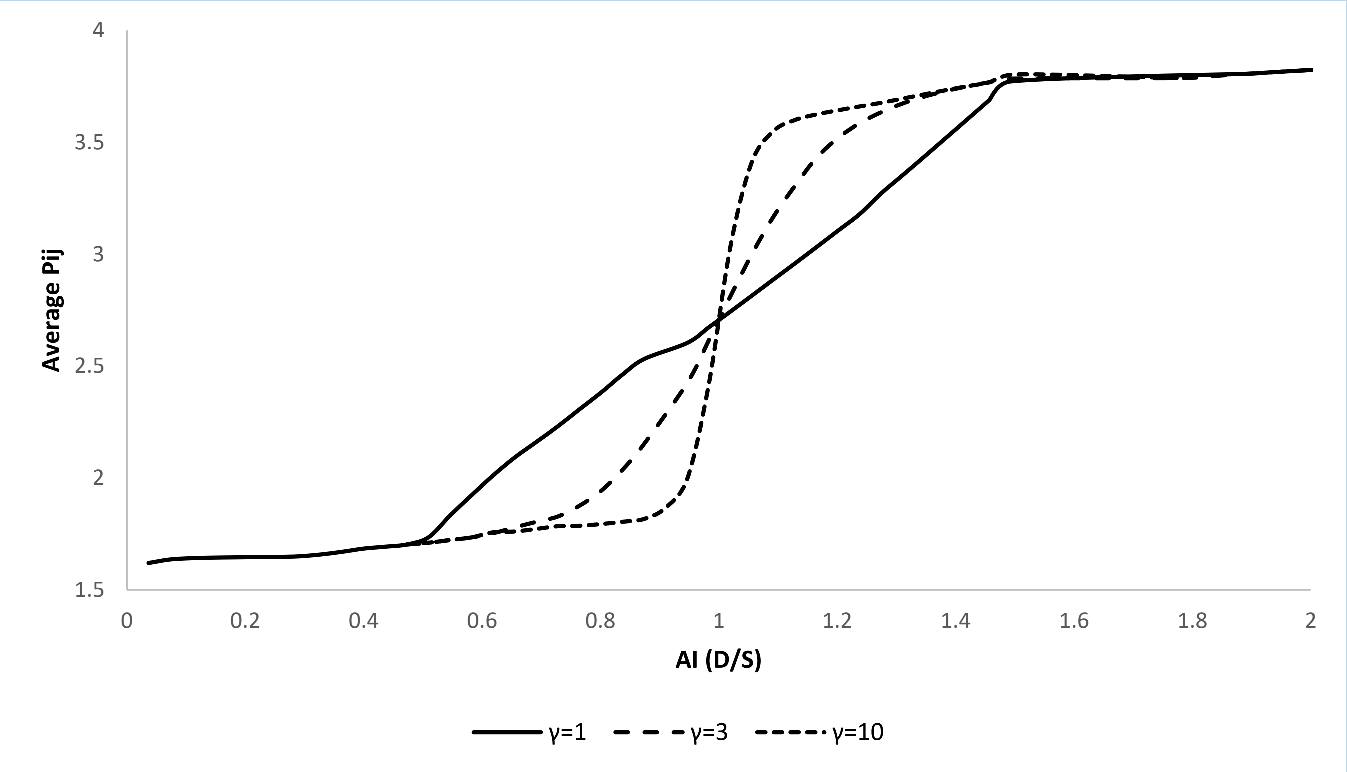



- *price decided according to AI and adjusting factor γ*
- *demand is too high compared to the available space .*
- *all the economic benefit should be given to the warehouse*

- *too much excess warehouse capacity.*
- *all the economic benefit of the transaction is given to the manufacturer*
- *no economic benefit should given to the warehouse*

# Case-study

- *Real-world case and data*
- *20 manufacturers, 13 warehouse operators*
- *All manufacturers were matched / 6 of 13 warehouses utilized*
- *2 warehouses covered 76% of manufacturers demand (sound result, as those are amongst the cheapest, largest and most suitably located warehouses of the dataset)*



# Risks estimation in the transport of dangerous goods for supporting policy making

# Introduction and scope

*Methodological framework for the estimation of hazardous levels during Fright transportation*

- *Qualitative and Quantitative Indicators*

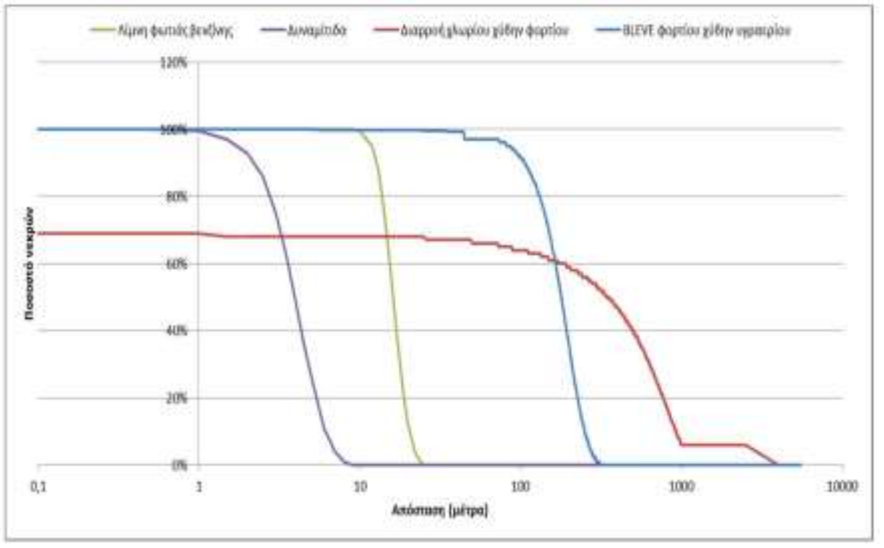
# Data Sources

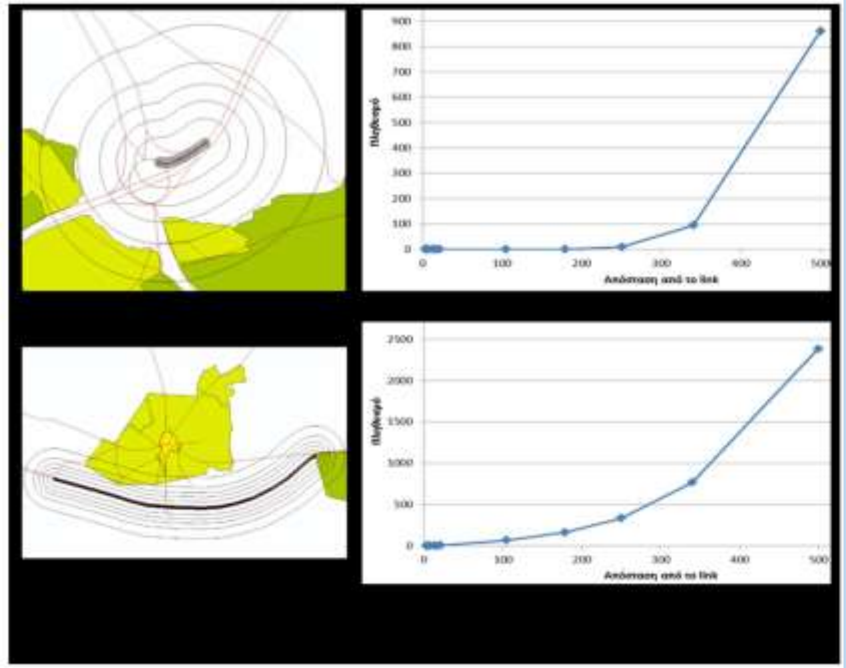
- *Traffic model data for the assessment of accident Occurrence rate*
- *Urban development data for the calculation of population density and impact on accident occurrence*
- *Freight flow data to determine the likelihood of involvement of dangerous-goods vehicles in accidents*
- *Accident data (previous years)*

# Criteria for risk assessment

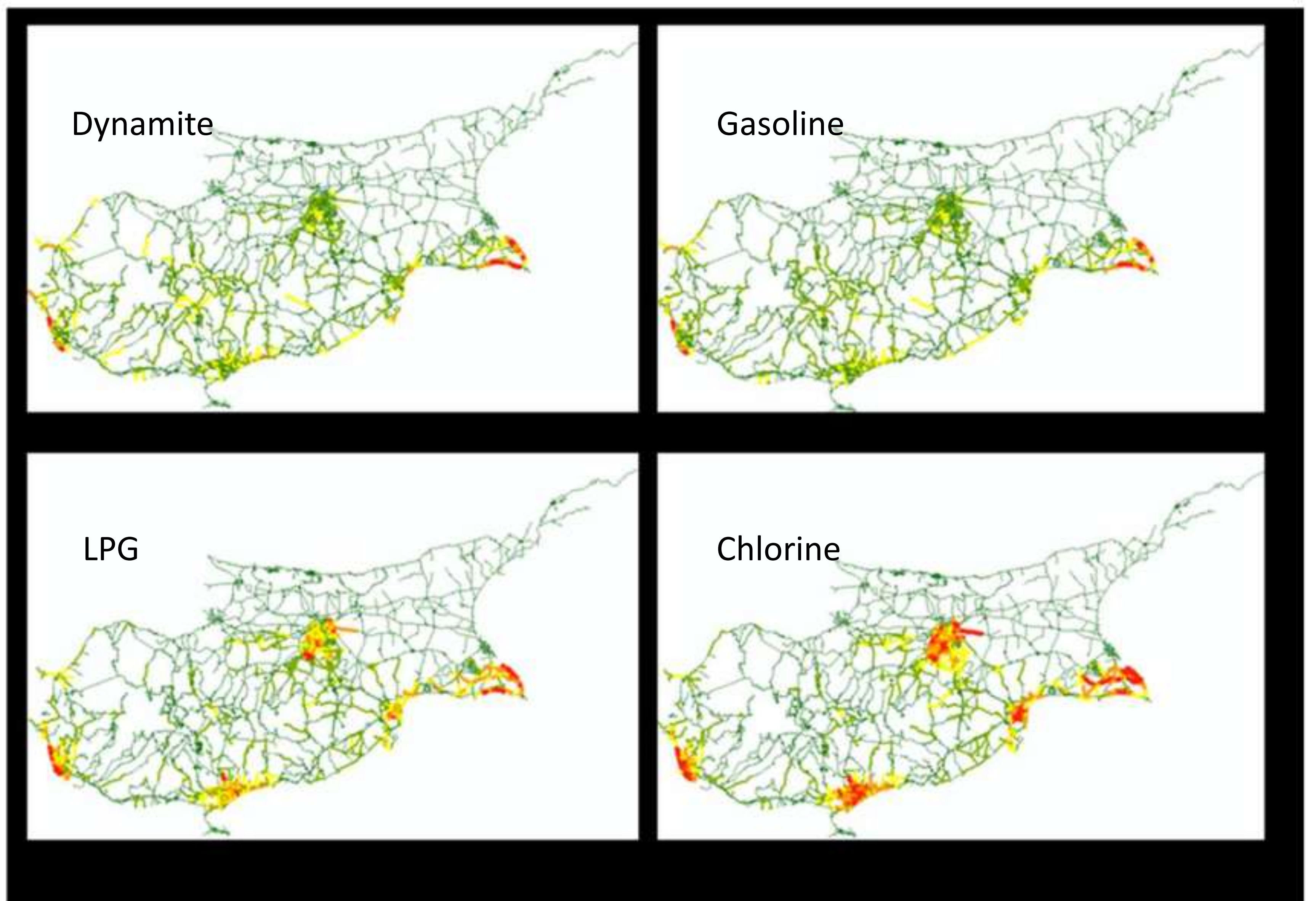
| Factors | Level | Material Type | | | |
|---|---|---|---|---|---|
| | | Explosive Material | Liquid Fuel | Gas Fuel | Chlorine |
| Rate of accidents (Accidents/ year) | Low | < 0,5 | | | |
| | Medium | <1,5 | | | |
| | High | >=1,5 | | | |
| Traffic situation (flow/ capacity) | Low | < 0,25 | | | |
| | Medium | <0,7 | | | |
| | High | >= 0,7 | | | |
| Transits of DG vehicles (annual number of DG transits | Low | <12 | < 2.700 | < 2.700 | <12 |
| | Medium | N/A | < 8.100 | < 8.100 | N/A |
| | High | N/A | >= 8.100 | >= 8.100 | N/A |
| Fatal effects (number of deaths) | Low | < 1 | | | |
| | Medium | < 3 | | | |
| | High | >= 3 | | | |

# *Dispersion rates*



# Results



# CERTH-HIT Portal

*Applications and open datasets published/hosted in our portal:*

- *www.trafficthess.imet.gr*
- *www.trafficpaths.imet.gr*
- *www.trafficthessreports.imet.gr*
- *www.opendata.imet.gr*
- *www.thessmd.imet.gr*

# THANK YOU!