# Benchmarking Driving Efficiency using Data Science Techniques applied on Large-Scale Smartphone Data



**Dimitris Tselentis**

Civil - Transportation Engineer
Ph.D. Candidate – Researcher

www.nrso.ntua.gr/dtsel/
dtsel@central.ntua.gr

# Scope

- **Methodological approach** for driving safety efficiency benchmarking:
    - trip
    - driver
    - multi-criteria analysis

- Safety **efficiency index**
    - travel characteristics
    - driving behaviour metrics
    - smartphone devices

- **Smartphone** devices
    - large-scale data
    - naturalistic driving conditions

# Research questions

- How well can driving safety **efficiency** be **benchmarked**? Can data science techniques and large-scale data provide sufficient answers?

- What are the **temporal evolution** characteristics of driving efficiency? What do the drivers' groups formed represent?

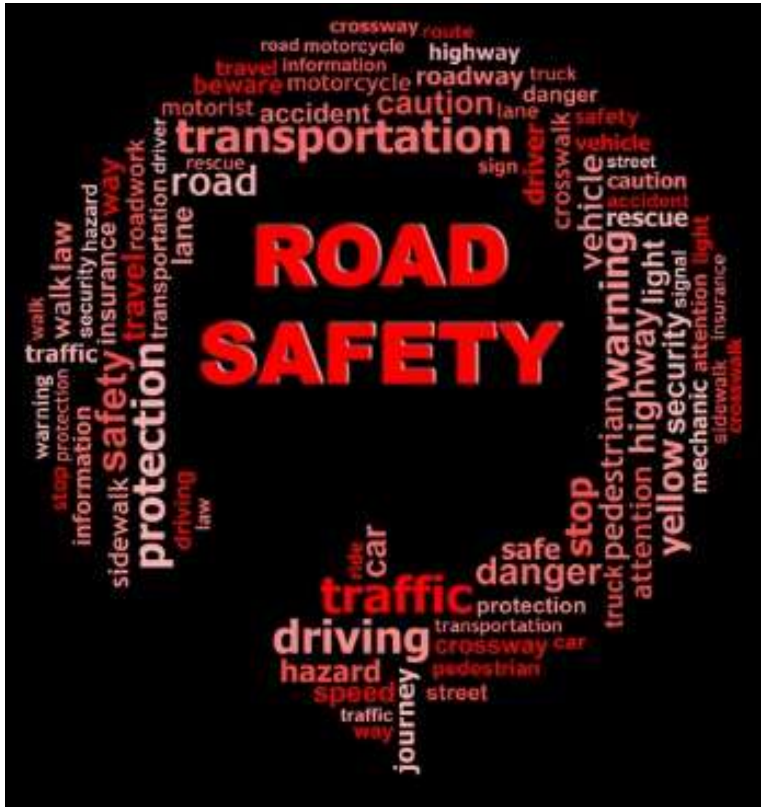
- What is the required amount of **driving data** that should be collected for each driver?

- How can the **least efficient** trips of a database be identified?

# State-of-the-art (1/4)

- Methods for measuring **safety efficiency** in transportation
    - low number of parameters
    - correlation between factors
    - quantify overall driving efficiency
        - multi-criteria analysis

- **Linear programming** for efficiency measurement
    - Data Envelopment Analysis (DEA)
        - Banks, Companies, Hospitals, Staff etc.
        - Transport (Systems, Traffic safety etc.)
    - Driving behaviour research



# State-of-the-art (2/4)

- **Elimination of limiting barriers** existed so far
    - Mobile phone technology
    - High cost of
        - in-vehicle data recording systems (e.g. OBD)
        - data plans
        - cloud computing
    - Low penetration rate of smartphones
    - Inability to manage and exploit Big Data

- Current **technological advances**
    - collect and exploit data through mobile phones
    - easier and more accurately

# State-of-the-art (3/4)

- **Driving data** collection
  - Naturalistic driving experiments
  - Driving simulator experiments
  - In-depth accident investigation

- Driving metrics - Adequate **amount**
  - assessment of each driver
  - deficient amount of data => uncertain or unreasonable results
  - excessive amount of data => significantly increase required processing time

# State-of-the-art (4/4)

- **UBI** schemes
  - Pay-As-You-Drive (PAYD)
  - Pay-How-You-Drive (PHYD)
  - Pay-at-the-pump (PATP)

- **Travel** behaviour characteristics
  - Total distance
  - Road network type
  - Risky hours driving
  - Trip frequency
  - Vehicle type
  - Weather conditions

- **Driving** behaviour characteristics
  - Speeding
  - Harsh braking/ acceleration/ cornering
  - Seatbelt use
  - Mobile phone use

# Knowledge gap (1/2)

- **Benchmarking** driving safety efficiency
    - microscopic driving data
    - travel and driving behaviour

- **Large-scale** data from naturalistic driving experiments

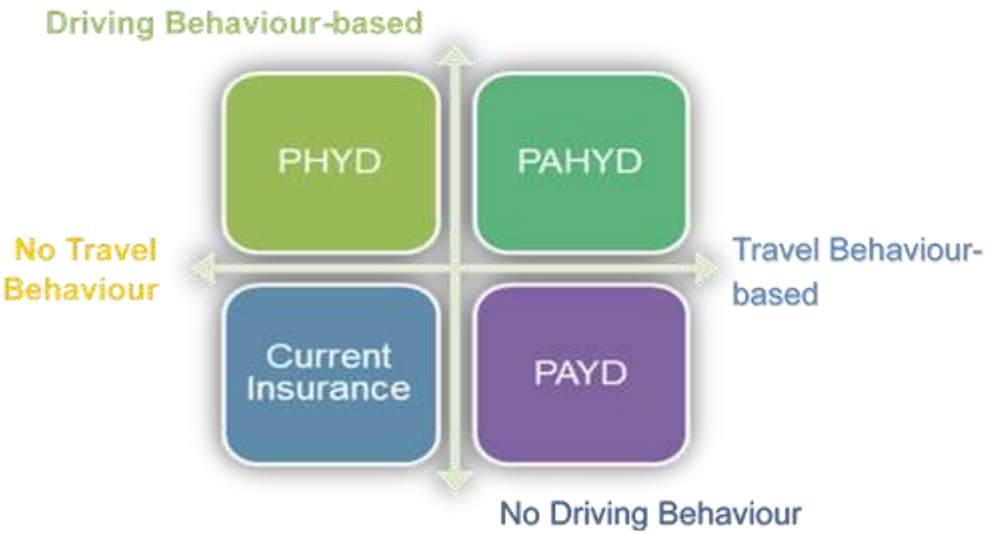
- **Human factors** recorded from smartphone
    - Harsh acceleration/ braking events
    - Time of mobile usage
    - Time of driving over the speed limits

# Knowledge gap (2/2)

- Amount of **driving data** to be recorded

- **Usage-based** insurance (UBI) schemes
    - travel and driving behaviour
    - reduce annual mileage
    - improve their driving behaviour

- **Risk factor**
    - risk's increase rate
    - driving behaviour
    - mileage

# Methodological approach

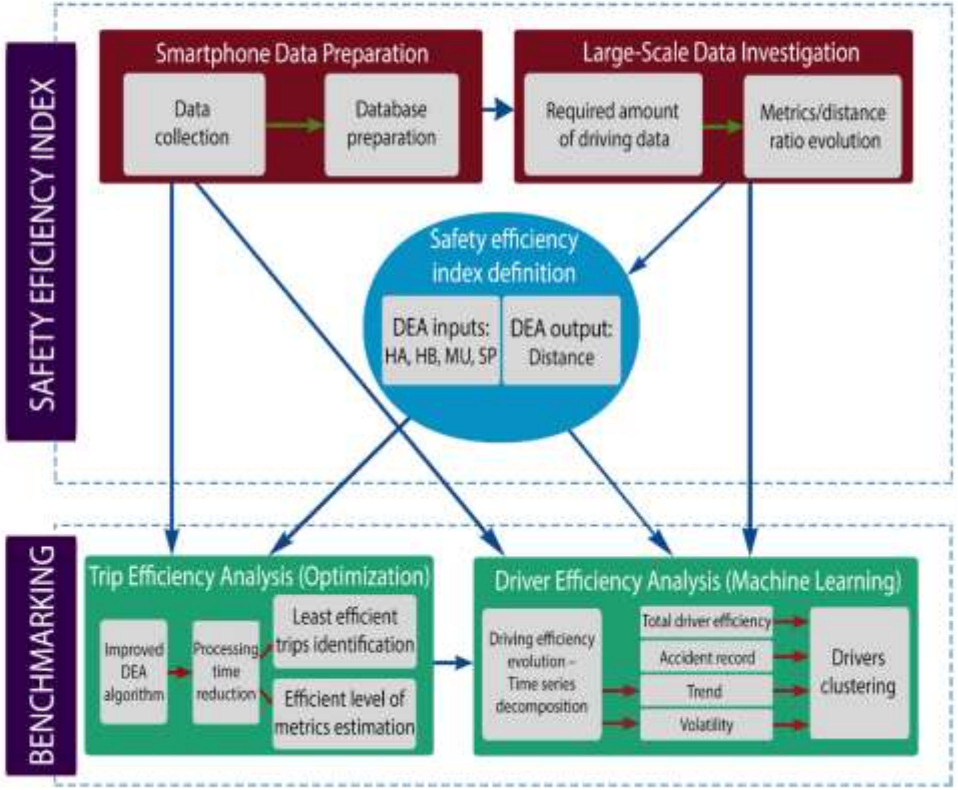
- **Smartphone data** collection
  - data preparation

- Large-scale **data investigation**
  - investigation of metrics-distance ratio evolution
  - adequate driving data

- **Safety** efficiency **index** estimation

- **Trip efficiency** analysis
  - identification of the least efficient trips
  - estimation of the efficient level of metrics

- **Driver efficiency** analysis
  - time-series decomposition
  - drivers clustering

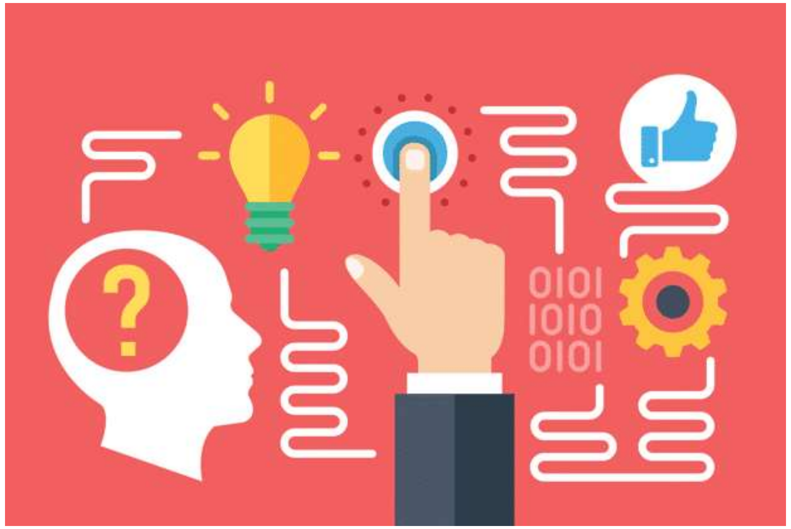


# Data envelopment analysis (1/2)

- **Optimization** technique
  - Performance measurement using **DEA**
    - Companies, banks, hospitals, staff etc.

- **Significant** computation cost
  - Exact solution
  - Reduced Basis Entry
  - Convex Hull
  - Reduce processing time

- **Decision-Making-Unit** (DMU)
  - Factory, company etc.
  - Trips, drivers
    - variables are continuous and quantitative
    - a driver should reduce the frequency of his driving characteristics for a given mileage




# Data envelopment analysis (2/2)

- **Efficiency index** Driving_efficiency$_B$

- **Input-oriented** DEA
  - minimize inputs (number of HA, HB events etc.) per driving distance

- **Constant-returns-to-scale** (CRS) problem
  - the sum of all inputs changes proportionally to the sum of driving output (distance)

- **Efficient level** of driving characteristics for a trip/ driver
  - inputs
  - outputs

$$\min(\text{Driving\_Efficiency}_B)$$

Subject to the following constraints:

$$\text{Driving\_Efficiency}_B * x_o - X * \lambda \geq 0$$

$$Y * \lambda \geq y_o$$

$$\lambda_i \geq 0 \forall \lambda_i \in \lambda$$

$$Metric_i = \sum_{j=1}^{m} \lambda_j * Metric_j$$

$$distance_{urban} = distance_i \,/\, \text{Driving\_Efficiency}_i$$

# Efficiency index parameters

**Risk exposure** indicators:

- Total **distance** travelled

**Driving behaviour** indicators**:**

- **Harsh events**
    - Number of harsh braking (longitudinal acceleration) (HA)
    - Number of harsh acceleration (longitudinal acceleration) (HB)

- **Speeding** (SP)
    - seconds driving over the speed limit

- **Mobile phone** use distraction (MU)
    - seconds using the mobile phone


**Road types**:

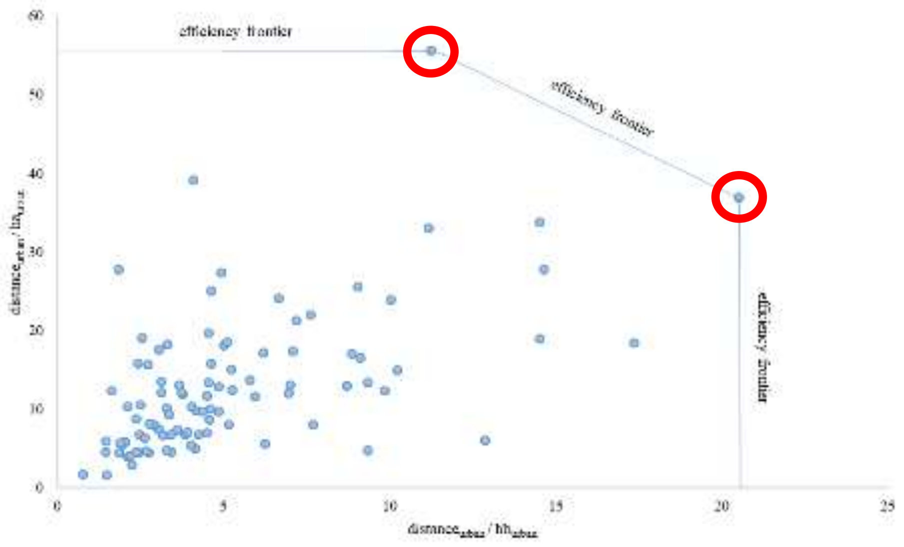
- Urban road network (signalized or not network, speed limit ≤ 50 km/h)
- Urban express road network (speed limit 50 – 90 km/h)
- Highways (speed limit ≥ 90 km/h)

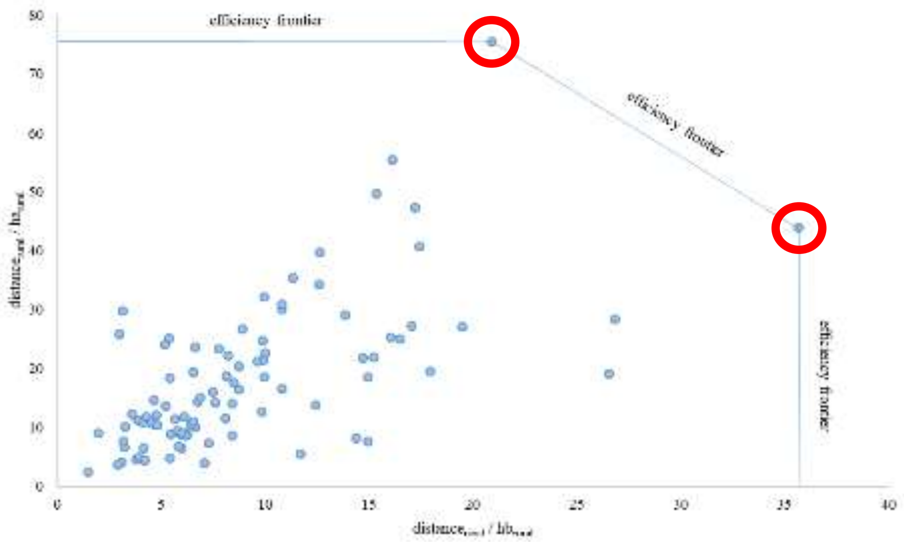
# Driving efficiency analysis using DEA

- DEA results **2-D illustration**
  - Urban
  - Rural

- DEA **inputs**
  - Number of harsh acceleration events
  - Number of harsh braking events

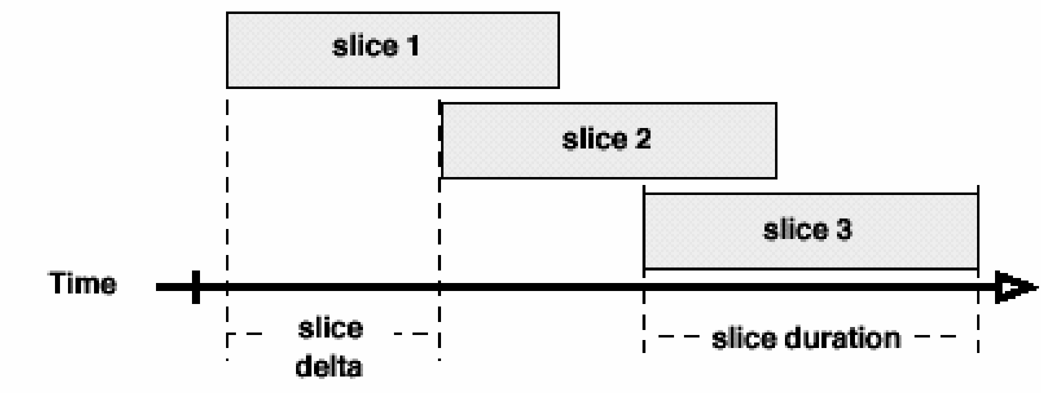
- DEA **outputs**
  - Trip distance

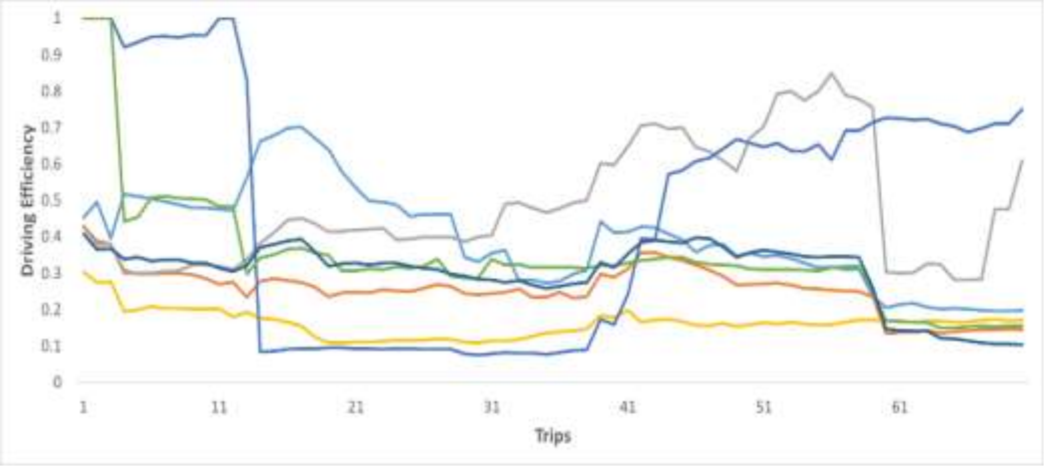


# Temporal evolution of efficiency

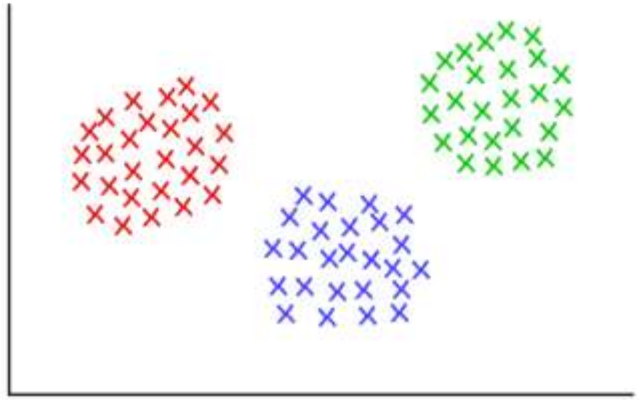
- **Aggregated** driving efficiency **benchmarking**
- Driving efficiency **benchmarking** in a **sliding window**
    - Driving behaviour changes
    - Time-series analysis
    - Time-series decomposition
        - Stationarity
        - Trend
        - Volatility
    - Road type



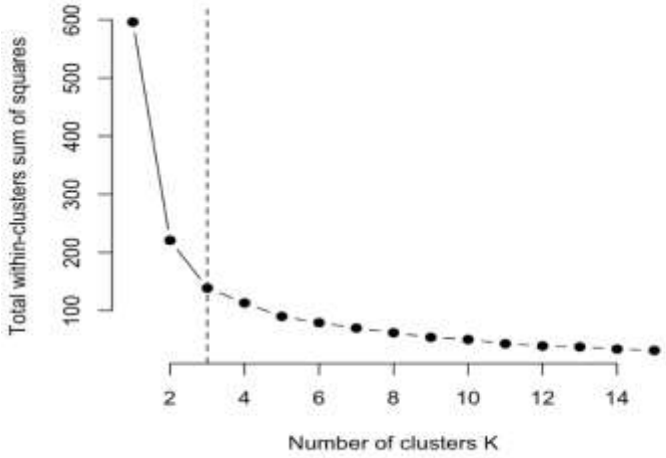



# Driver clustering

- **Clustering**
  - identify driving profile and their characteristics
  - k-means algorithm
  - elbow method



- Clustering **attributes**
  - total efficiency
  - trend
  - volatility
  - with no accident history (data_sample_1)
  - with accident history (data_sample_2)
  - results comparison



# Smartphone data collection

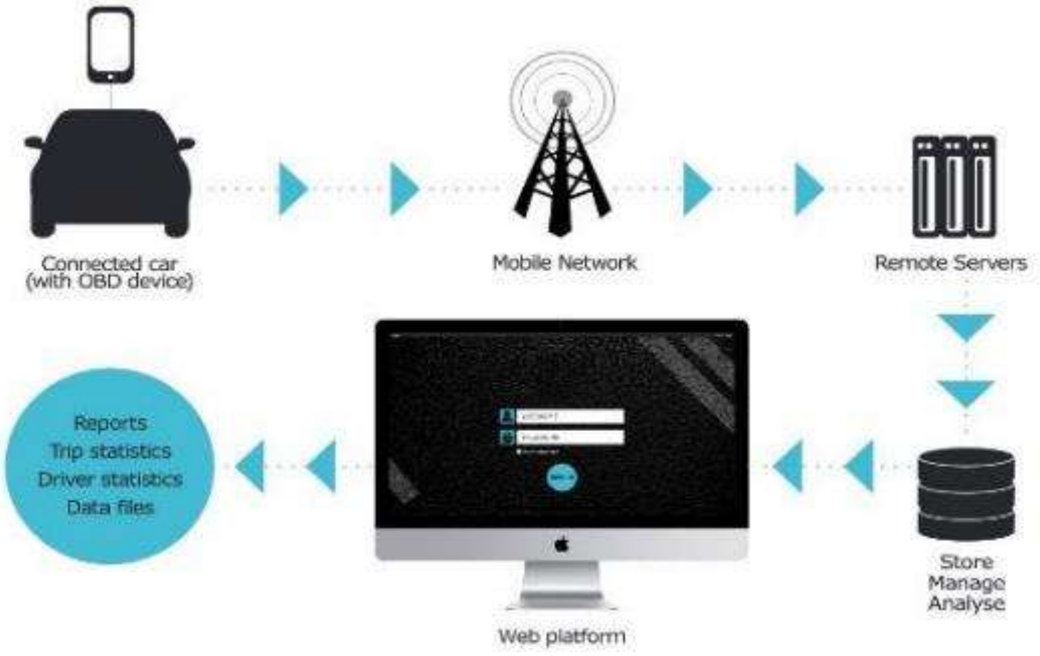
- A mobile **application** to record user's driving behaviour (automatic start / stop)

- A variety of **APIs** is used to read mobile phone sensor data

- Data is **transmitted** from the mobile App to the central database

- Data are **stored** in a sophisticated database where they are managed and processed

- **Indicators** are designed using
  - machine learning algorithms
  - big data mining techniques




Source: OSeven Telematics

# Questionnaire data

- **Driving experience**
  - Years of driving experience
  - Mileage
- **Vehicle**
  - Age of the vehicle
  - Engine capacity
  - Ownership
  - Fuel consumption
- **Driving behaviour**
  - Accident history
  - Self-assessment
- **Demographics**
  - Age
  - Gender
  - Education

# Data preparation

- Data are **anonymized**
    - user-agnostic approach
    - identify driving behaviors and patterns
    - causality between behaviour and other factors
    - large-scale samples
    - no information on demographics or accident record

- **Python** programming language
    - filter aggregate data
    - retain only necessary information
    - aggregate data
    - data analysis

# Data sample

| | Sampling time investigation | Trip efficiency analysis | Driver efficiency analysis | | | |
|---|---|---|---|---|---|---|
| | | | data_sample_1 | | data_sample_2 | |
| | | | Urban | Rural | Urban | Rural |
| **Number of drivers** | 171 | 88 | 100 | 100 | 43 | 39 |
| **Number of trips** | 49,722 | 10,088 | 23,000 | 15,000 | 9,890 | 5,850 |

# Data sample - Driver analysis

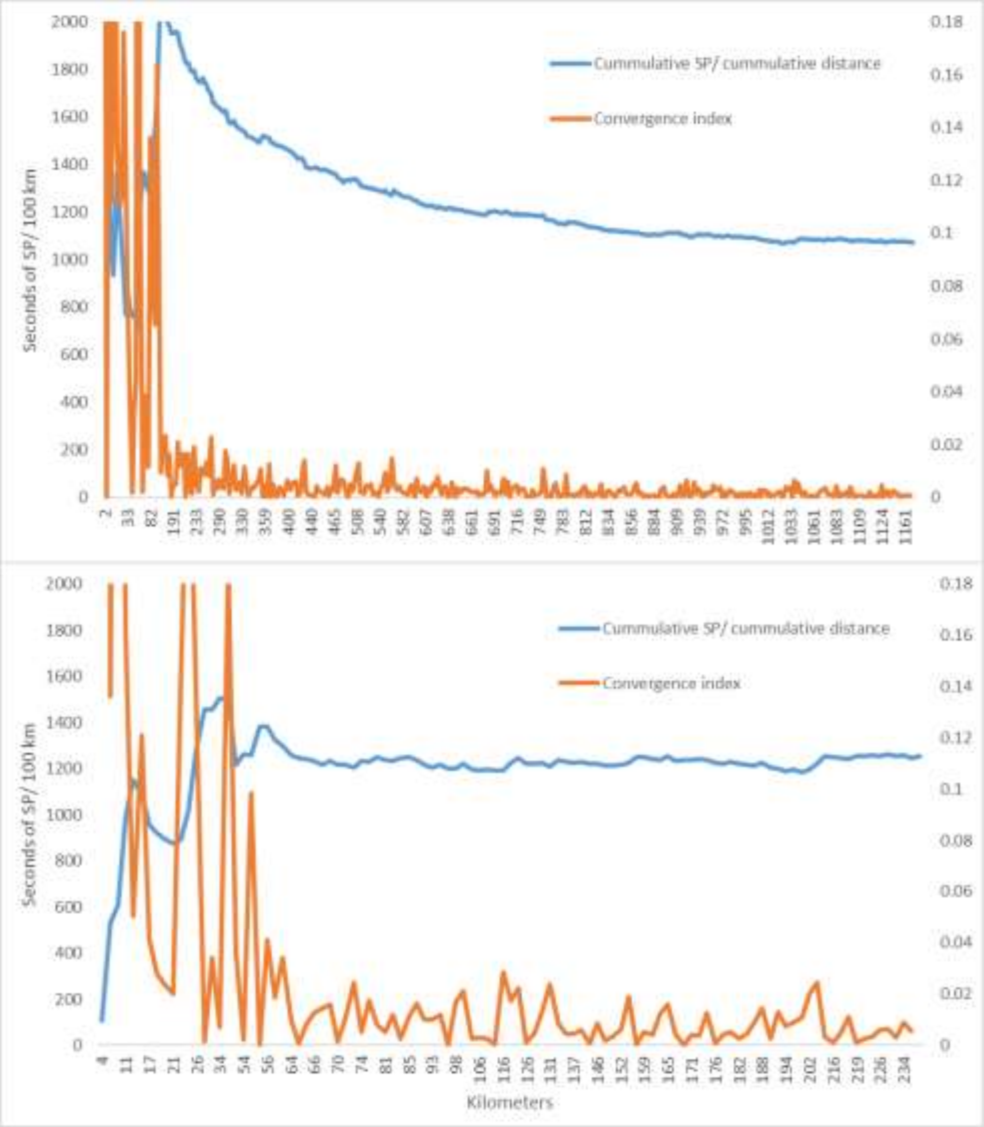
- **Criteria**
  - At least 50 more trips
    - than the minimum number required
    - to create the time series
  - Positive mileage on both road types
  - Positive sum of input attributes (HA, HB etc.)
  - Maximum number of drivers

- **Urban**
  - 100 drivers & 23,000 trips – w/o questionnaire data
  - 43 drivers & 9890 trips – w questionnaire data

- **Rural**
  - 100 drivers & 15,000 trips – w/o questionnaire data
  - 43 drivers & 5850 trips – w questionnaire data

# Data investigation (1/7)

- **Convergence index**
  - Moving window of 40 trips
  - at least 200 km
  - ≤ 5%



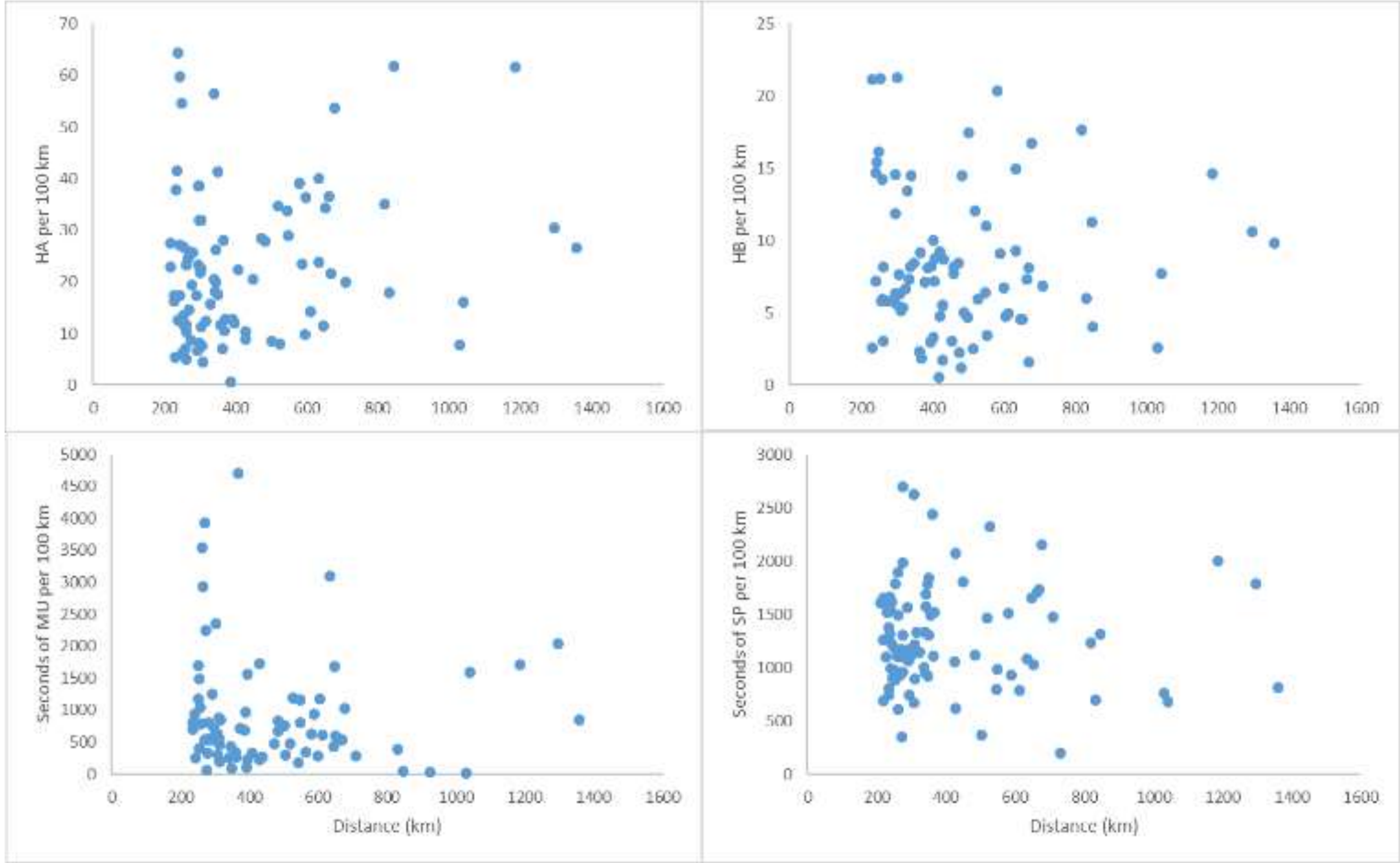
# Data investigation (2/7)

## Urban RN

- Weak **positive correlation** between HA and required distance for convergence

- Required monitoring distance is **higher** for **more** aggressive drivers

- **No** apparent **trend** for the rest of the metrics



# Data investigation (3/7)

## Urban RN

- **Same** sampling **periods** are required for drivers of different percentile value range
- **Maximum** median **distance** value
  - all metrics should have converged to their cumulative average
    - the driving sample acquired is adequate
    - the input/ output ratio is relatively constant to perform DEA analysis
- **519** km
  - 75 trips

| Metric | Percentile range | Metric descriptive statistics | | | | Distance to convergence | | |
|---|---|---|---|---|---|---|---|---|
| | | Average | St. Dev | Min | Max | Average | Median | St. Dev |
| HA | 0% – 25% | 8.18 | 2.61 | - | 11.61 | 389 | 309 | 177 |
| | 25% – 50% | 15.92 | 2.8 | 11.61 | 20.54 | 399 | 336 | 208 |
| | 50% – 75% | 25.01 | 2.48 | 20.54 | 30.14 | 415 | 322 | 246 |
| | 75% – 100% | 43.23 | 10.92 | 30.14 | - | 526 | 519 | 293 |
| HB | 0% – 25% | 3.05 | 1.26 | - | 4.95 | 515 | 478 | 179 |
| | 25% – 50% | 6.17 | 0.69 | 4.95 | 7.32 | 408 | 335 | 155 |
| | 50% – 75% | 8.6 | 0.83 | 7.32 | 10.61 | 558 | 431 | 300 |
| | 75% – 100% | 15.65 | 3.12 | 10.61 | - | 469 | 341 | 248 |
| MU | 0% – 25% | 204 | 101 | - | 332 | 559 | 407 | 395 |
| | 25% – 50% | 495 | 78 | 332 | 606 | 440 | 347 | 174 |
| | 50% – 75% | 799 | 111 | 606 | 1041 | 443 | 381 | 247 |
| | 75% – 100% | 2063 | 994 | 1041 | - | 493 | 366 | 310 |
| SP | 0% – 25% | 727 | 194 | - | 947 | 489 | 339 | 308 |
| | 25% – 50% | 1081 | 67 | 947 | 1198 | 343 | 293 | 123 |
| | 50% – 75% | 1402 | 119 | 1198 | 1594 | 378 | 312 | 188 |
| | 75% – 100% | 1919 | 318 | 1594 | - | 455 | 348 | 287 |

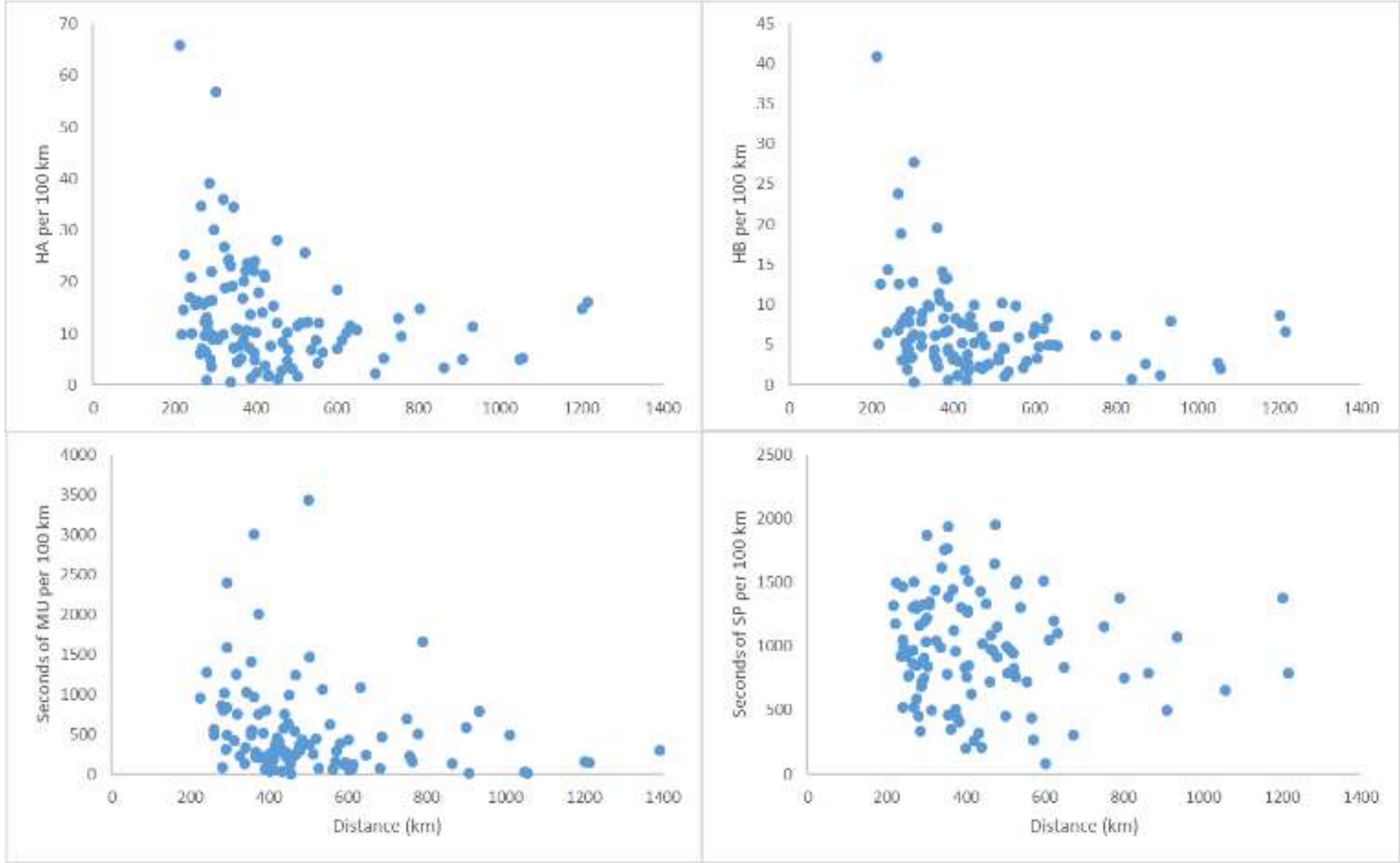
# Data investigation (4/7)

Urban express RN

- Weak **negative correlation** between HA, HB, mobile usage and the required distance for convergence

- Required monitoring distance is **higher** for **less** aggressive/ risky/ distracted drivers

- **No** apparent **trend** for speeding



# Data investigation (5/7)

Urban express RN

- **Different** sampling **periods** are required for drivers of different percentile value range
  - less volatile than in urban road

- **Maximum** median **distance** value

- **579** km
  - 81 trips

| Metric | Percentile range | Metric descriptive statistics | | | | Distance to convergence | | |
|---|---|---|---|---|---|---|---|---|
| | | Average | St. Dev | Min | Max | Average | Median | St. Dev |
| HA | 0% – 25% | 3.69 | 1.7 | - | 6.42 | 507 | 431 | 219 |
| | 25% – 50% | 8.8 | 1.33 | 6.42 | 10.79 | 405 | 376 | 134 |
| | 50% – 75% | 13.58 | 1.98 | 10.79 | 17.02 | 499 | 417 | 270 |
| | 75% – 100% | 27.27 | 11.04 | 17.02 | - | 349 | 336 | 85 |
| HB | 0% – 25% | 2.05 | 0.84 | - | 3.15 | 52 | 439 | 221 |
| | 25% – 50% | 4.35 | 0.67 | 3.15 | 5.6 | 429 | 403 | 126 |
| | 50% – 75% | 6.92 | 0.69 | 5.6 | 8.28 | 489 | 440 | 221 |
| | 75% – 100% | 13.54 | 7.16 | 8.28 | - | 389 | 360 | 187 |
| MU | 0% – 25% | 85 | 48 | - | 157 | 630 | 579 | 261 |
| | 25% – 50% | 263 | 50 | 157 | 371 | 485 | 442 | 209 |
| | 50% – 75% | 511 | 92 | 371 | 747 | 501 | 448 | 192 |
| | 75% – 100% | 1334 | 684 | 747 | - | 411 | 362 | 166 |
| SP | 0% – 25% | 454 | 170 | - | 745 | 448 | 399 | 189 |
| | 25% – 50% | 851 | 74 | 745 | 970 | 444 | 397 | 221 |
| | 50% – 75% | 1142 | 112 | 970 | 1315 | 419 | 388 | 172 |
| | 75% – 100% | 1526 | 181 | 1315 | - | 425 | 356 | 195 |

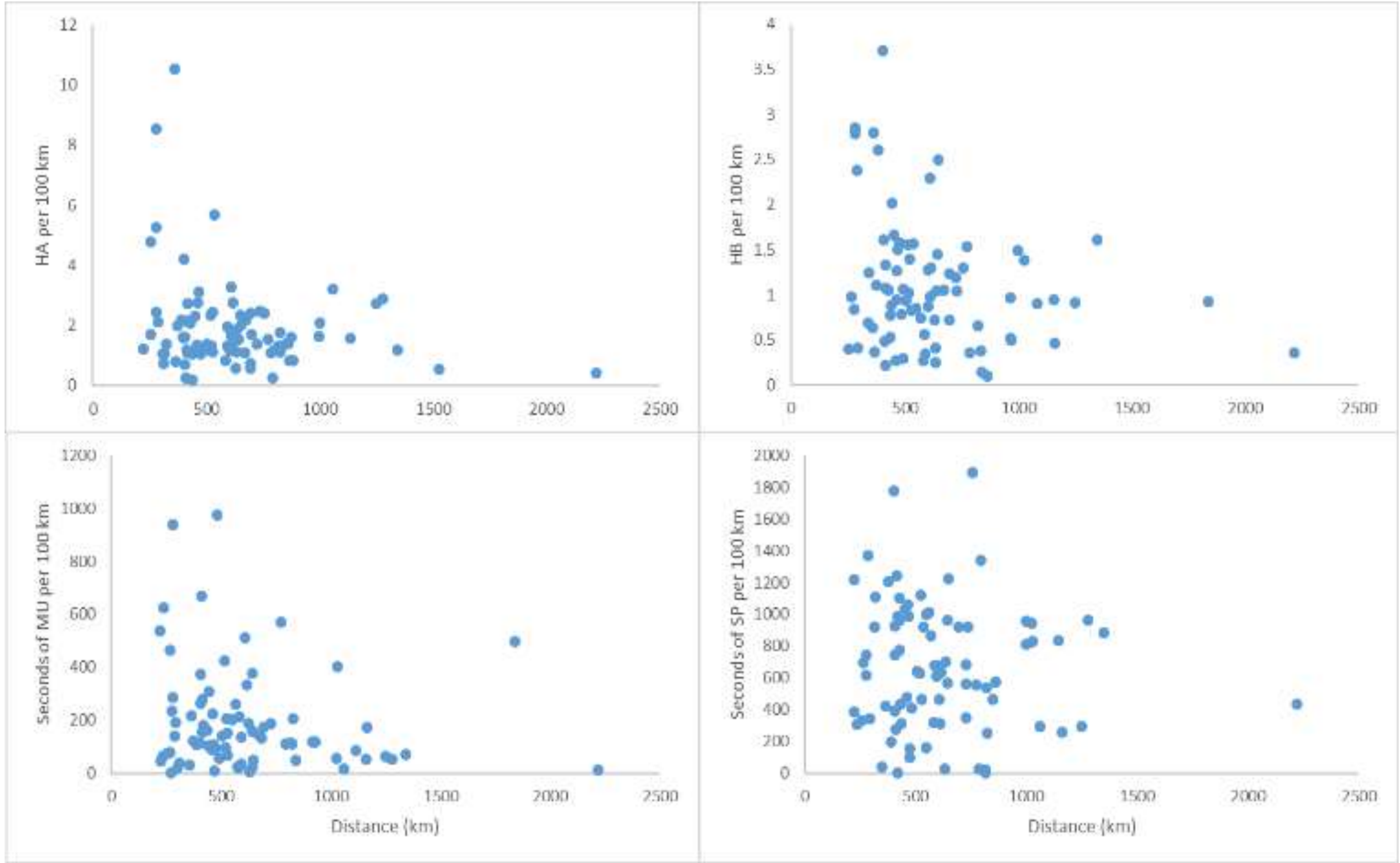
# Data investigation (6/7)

<u>Highway</u>

- Weak **negative correlation** between HA, HB, mobile usage and the required distance for convergence

- Required monitoring distance is **higher** for **less** aggressive/ risky/ distracted drivers

- **No** apparent **trend** for speeding



# Data investigation (7/7)

Highway

- **Different** sampling **periods** are required for drivers of different percentile value range

- **Maximum** median **distance** value

- **611** km
    - is not investigated

| Metric | Percentile range | Metric descriptive statistics | | | | Distance to convergence | | |
|---|---|---|---|---|---|---|---|---|
| | | Average | St. Dev | Min | Max | Average | Median | St. Dev |
| **HA** | 0% – 25% | 0.74 | 0.29 | - | 1.1 | 678 | 585 | 440 |
| | 25% – 50% | 1.26 | 0.12 | 1.1 | 1.54 | 639 | 605 | 234 |
| | 50% – 75% | 1.87 | 0.24 | 1.54 | 2.3 | 607 | 611 | 242 |
| | 75% – 100% | 3.77 | 2.12 | 2.3 | - | 593 | 529 | 289 |
| **HB** | 0% – 25% | 0.36 | 0.12 | - | 0.56 | 705 | 592 | 413 |
| | 25% – 50% | 0.83 | 0.1 | 0.56 | 0.97 | 700 | 562 | 376 |
| | 50% – 75% | 1.15 | 0.13 | 0.97 | 1.38 | 572 | 606 | 174 |
| | 75% – 100% | 2.05 | 0.64 | 1.38 | - | 542 | 472 | 251 |
| **MU** | 0% – 25% | 35 | 20 | - | 66 | 722 | 592 | 468 |
| | 25% – 50% | 101 | 18 | 66 | 135 | 623 | 498 | 285 |
| | 50% – 75% | 174 | 26 | 135 | 223 | 564 | 540 | 196 |
| | 75% – 100% | 455 | 206 | 223 | - | 533 | 445 | 349 |
| **SP** | 0% – 25% | 193 | 124 | - | 346 | 610 | 550 | 285 |
| | 25% – 50% | 505 | 91 | 346 | 641 | 649 | 592 | 392 |
| | 50% – 75% | 807 | 100 | 641 | 950 | 666 | 601 | 295 |
| | 75% – 100% | 1168 | 249 | 950 | - | 544 | 464 | 241 |

# Trip efficiency analysis (1/3)

- **Convex Hull** technique outperforms

- Significant **time reduction**
  - As the database becomes larger
  - Convex Hull DEA – 5 minutes
  - RBE DEA – 12.6 days
  - Standard DEA – 40.7 days

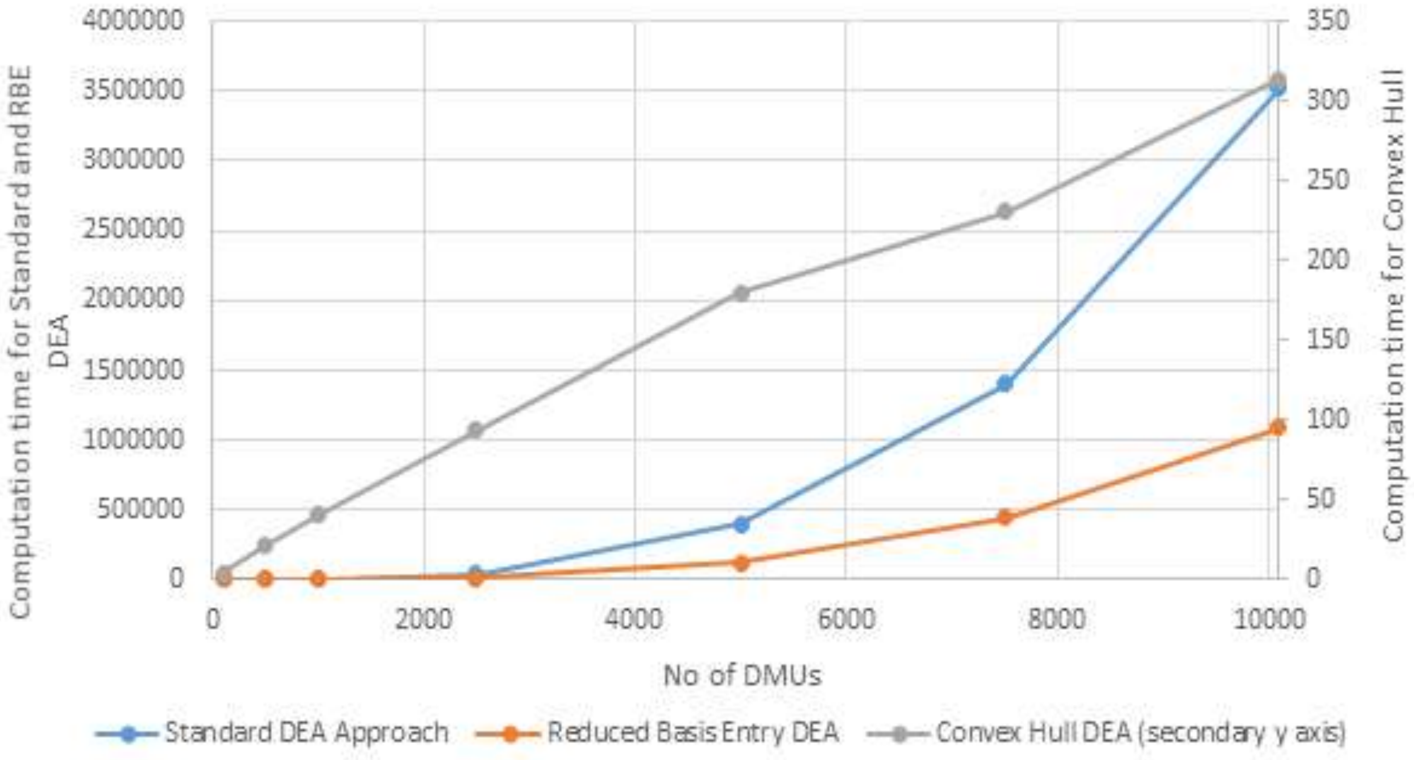
| | Computation time (sec) | | | CH DEA % computation time improvement over | | RBE DEA % computation time improvement over Standard DEA |
|---|---|---|---|---|---|---|
| **No of DMUs** | Standard DEA Approach | RBE DEA | Convex Hull DEA | Standard DEA Approach | RBE DEA | |
| **100** | 11 | 6 | 4 | 63.64% | 33.33% | 45.45% |
| **500** | 477 | 169 | 21 | 95.60% | 87.57% | 64.57% |
| **1000** | 3250 | 1121 | 41 | 98.74% | 96.34% | 65.51% |
| **2500** | 44435 | 15570 | 94 | 99.79% | 99.40% | 64.96% |
| **5000** | 398485 | 123986 | 180 | 99.95% | 99.85% | 68.89% |
| **7500** | 1400909 | 444498 | 231 | 99.98% | 99.95% | 68.27% |
| **10088** | 3519372 | 1089731 | 314 | 99.99% | 99.97% | 69.04% |

*** Inputs = [$ha_{urban}$, $ha_{rural}$, $ha_{highway}$], Outputs = [$distance_{urban}$, $distance_{rural}$, $distance_{highway}$]**

# Trip efficiency analysis (2/3)

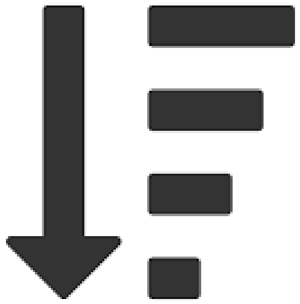
- Convex Hull DEA
    - linearly increased

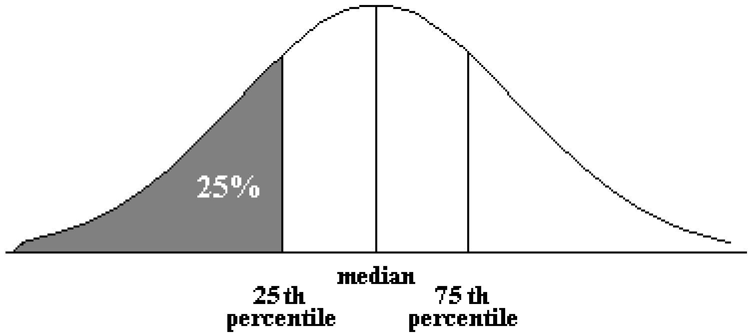
- Standard and RBE DEA
    - Exponentially increased



# Trip efficiency analysis (3/3)

- **Least efficient** trips
- **Efficiency** estimation
- **Sort**
    - larger to smaller
- **Percentile**
    - 5%, 10%, 25% etc.



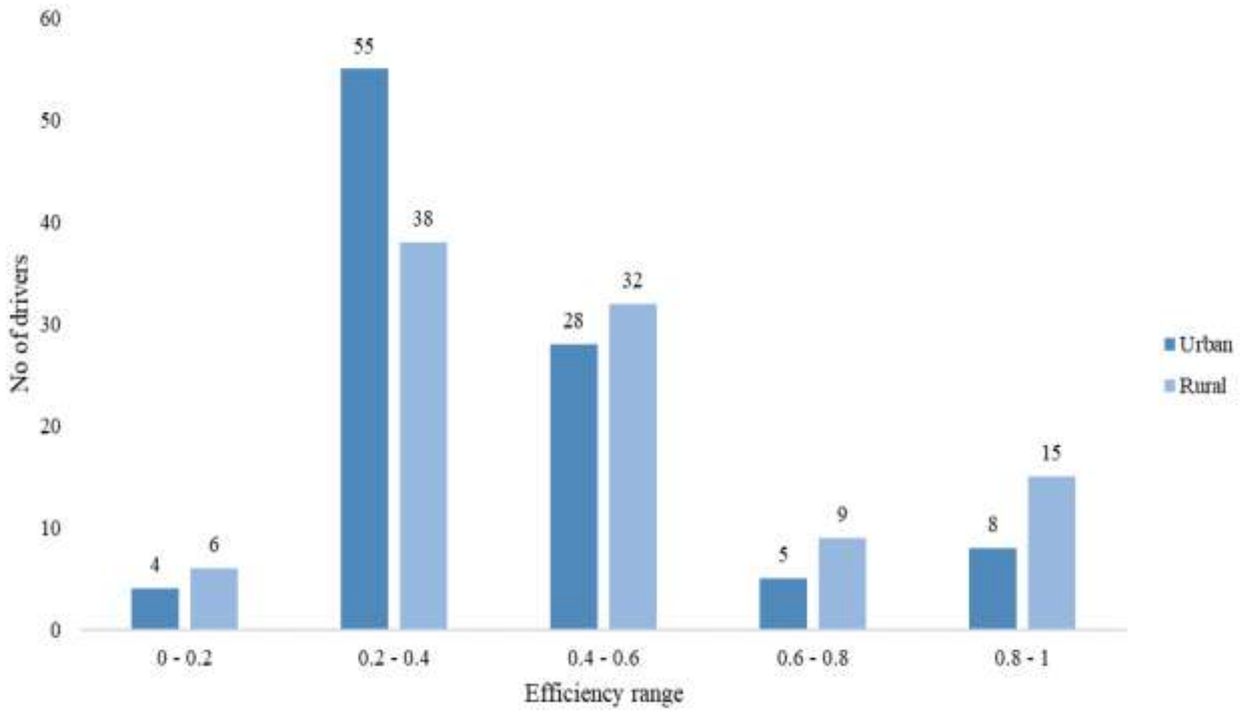


# Driver efficiency analysis - Aggregated efficiency (1/13)

- data_sample_1

- **Similar distribution** in both road types

- Drivers gathered in **low-medium** efficiency ranges
    - DEA provides a relative index

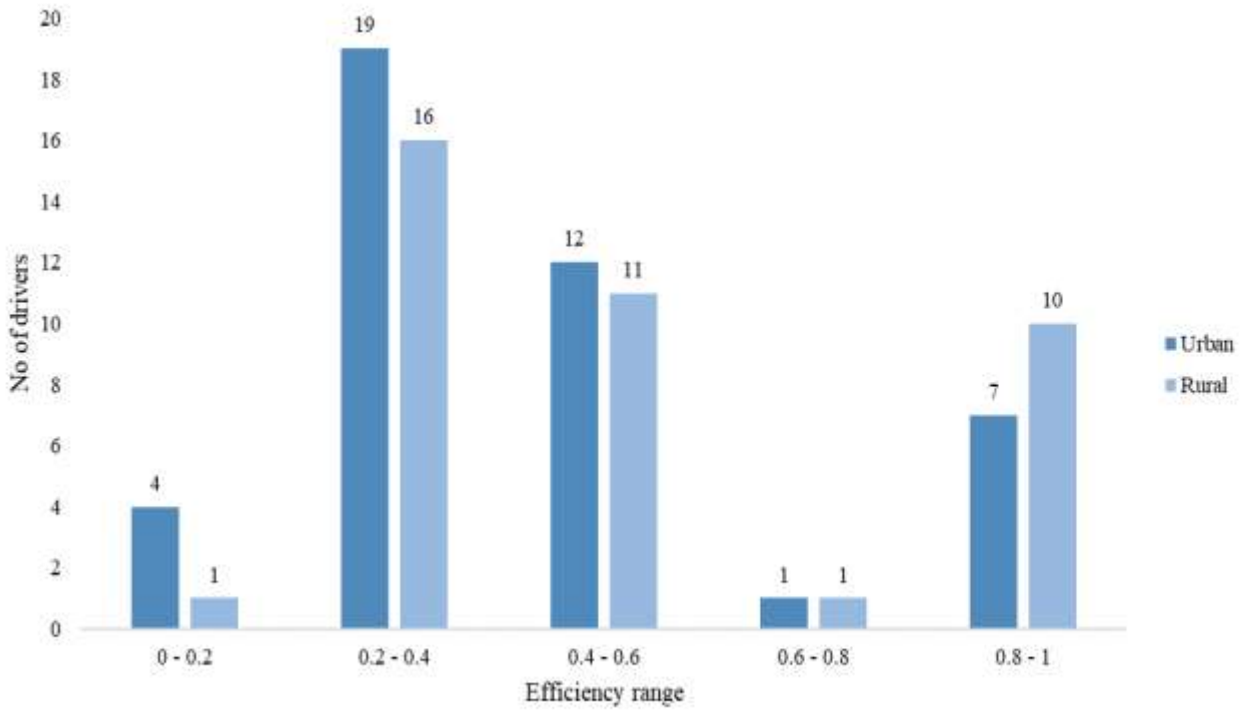


# Driver efficiency analysis - Aggregated efficiency (2/13)

- data_sample_2

- **Similar distribution** in both road types

- Drivers gathered in **low-medium** efficiency ranges
    - DEA provides a relative index



# Driver efficiency analysis - Aggregated efficiency (3/13)

- **Aggregated** driving behaviour
  - Percentiles of driving efficiency
  - Three efficiency classes
    - 1: Non-efficient drivers
    - 2: Weakly efficient drivers
    - 3: Efficient drivers
  - Median driving characteristics

- **Urban RN**
  - class 1 -> class 2
    - number of harsh acceleration/ braking events
  - class 2 -> class 3
    - time of speeding

- **Urban express** RN
  - class 1 -> class 2
    - time of speeding
  - class 2 -> class 3
    - number of harsh acceleration events
    - time of mobile usage

| Sample type | Road type | No of drivers | Driving characteristics (/100km) | Efficiency classes | | |
|---|---|---|---|---|---|---|
| | | | | Class 1: 0 - 25 % percentile | Class 2: 25 - 75 % percentile | Class 3: 75 - 100 % percentile |
| data_sample_1 | Urban | 100 | efficiency | 0.22 | 0.36 | 0.61 |
| | | | ha | 21.49 | 11.82 | 8.82 |
| | | | hb | 9.64 | 5.31 | 3.68 |
| | | | mu (sec) | 316 | 205 | 141 |
| | | | sp (sec) | 1243 | 878 | 355 |
| | Rural | 100 | efficiency | 0.24 | 0.42 | 0.90 |
| | | | ha | 34.11 | 24.06 | 11.30 |
| | | | hb | 14.92 | 9.16 | 5.42 |
| | | | mu (sec) | 529 | 419 | 165 |
| | | | sp (sec) | 1564 | 1004 | 708 |
| data_sample_2 | Urban | 43 | efficiency | 0.21 | 0.38 | 1.00 |
| | | | ha | 39.26 | 21.71 | 9.98 |
| | | | hb | 16.38 | 8.07 | 4.19 |
| | | | mu (sec) | 751 | 553 | 100 |
| | | | sp (sec) | 1892 | 965 | 477 |
| | Rural | 39 | efficiency | 0.28 | 0.44 | 1.00 |
| | | | ha | 23.04 | 11.86 | 7.49 |
| | | | hb | 9.28 | 5.21 | 3.16 |
| | | | mu (sec) | 316 | 305 | 160 |
| | | | sp (sec) | 1423 | 939 | 378 |

# Driver efficiency analysis - Aggregated efficiency (4/13)

- **Mobile usage**
    - Slight differences between classes 1 and 2
        - approximately the same usage for drivers of all classes
        - DEA sensitivity to outliers
- **Higher number** of harsh events in express Urban roads than in Urban roads
- Number of HA events = 2 * Number of HB events

| Sample type | Road type | No of drivers | Driving characteristics (/100km) | Efficiency classes | | |
|---|---|---|---|---|---|---|
| | | | | Class 1: 0 - 25 % percentile | Class 2: 25 - 75 % percentile | Class 3: 75 - 100 % percentile |
| data_sample_1 | Urban | 100 | efficiency | 0.22 | 0.36 | 0.61 |
| | | | ha | 21.49 | 11.82 | 8.82 |
| | | | hb | 9.64 | 5.31 | 3.68 |
| | | | mu (sec) | 316 | 205 | 141 |
| | | | sp (sec) | 1243 | 878 | 355 |
| | Rural | 100 | Efficiency | 0.24 | 0.42 | 0.90 |
| | | | ha | 34.11 | 24.06 | 11.30 |
| | | | hb | 14.92 | 9.16 | 5.42 |
| | | | mu (sec) | 529 | 419 | 165 |
| | | | sp (sec) | 1564 | 1004 | 708 |
| data_sample_2 | Urban | 43 | efficiency | 0.21 | 0.38 | 1.00 |
| | | | ha | 39.26 | 21.71 | 9.98 |
| | | | hb | 16.38 | 8.07 | 4.19 |
| | | | mu (sec) | 751 | 553 | 100 |
| | | | sp (sec) | 1892 | 965 | 477 |
| | Rural | 39 | efficiency | 0.28 | 0.44 | 1.00 |
| | | | ha | 23.04 | 11.86 | 7.49 |
| | | | hb | 9.28 | 5.21 | 3.16 |
| | | | mu (sec) | 316 | 305 | 160 |
| | | | sp (sec) | 1423 | 939 | 378 |

# Driver efficiency analysis - Efficient level (5/13)

| | Real level of metrics | | | | | | Lamdas of peers: Driver No | | | | Efficient level of metrics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Driver No** | $distance_{urban}$ | $ha_{urban}$ | $hb_{urban}$ | $speeding_{urban}$ | $mobile_{urban}$ | Theta | 12 | 34 | 40 | 42 | $distance_{urban}$ | $ha_{urban}$ | $hb_{urban}$ | $speeding_{urban}$ | $mobile_{urban}$ |
| **1** | 1868 | 326 | 134 | 21712 | 9954 | 0.581 | - | 0.52 | 0.14 | 0.06 | 3214.3 | 159.8 | 77.9 | 12617.9 | 5784.7 |
| **2** | 2456 | 574 | 85 | 27049 | 13974 | 0.696 | - | 0.19 | 0.40 | 0.20 | 3526.5 | 154.8 | 59.2 | 18838.2 | 9732.1 |
| **3** | 1634 | 709 | 509 | 15888 | 42817 | 0.391 | 0.79 | - | 0.20 | - | 4182.6 | 277.0 | 78.1 | 6206.9 | 10434.2 |
| **4** | 2219 | 233 | 181 | 27052 | 12421 | 0.637 | - | 0.23 | 0.31 | 0.18 | 3481.0 | 148.5 | 59.1 | 17244.6 | 7917.9 |
| **5** | 4223 | 1088 | 309 | 37825 | 29581 | 0.613 | 0.30 | 1.16 | 0.47 | - | 6887.3 | 417.6 | 189.5 | 23192.7 | 18137.9 |
| **6** | 2773 | 652 | 251 | 25829 | 25887 | 0.529 | 0.60 | 0.51 | 0.30 | - | 5245.3 | 344.7 | 128.8 | 13654.8 | 13685.4 |
| **7** | 2086 | 265 | 149 | 20880 | 14036 | 0.619 | - | 0.52 | 0.28 | 0.04 | 3371.9 | 163.9 | 79.2 | 12917.2 | 8683.2 |
| **8** | 1789 | 460 | 323 | 17185 | 5960 | 0.656 | - | 0.75 | - | 0.01 | 2728.0 | 184.3 | 98.1 | 11269.8 | 3908.5 |
| **9** | 1630 | 184 | 82 | 30801 | 8955 | 0.528 | - | - | 0.21 | 0.22 | 3085.0 | 97.2 | 30.8 | 14784.6 | 4731.5 |
| **10** | 808 | 266 | 95 | 9985 | 6562 | 0.443 | 0.11 | 0.24 | 0.04 | - | 1824.6 | 97.2 | 42.1 | 4421.6 | 2905.8 |
| **11** | 3012 | 913 | 152 | 21604 | 43562 | 0.585 | 0.72 | - | 0.83 | - | 5149.6 | 308.3 | 88.9 | 12636.1 | 23107.5 |
| **12** | 1462 | 329 | 92 | 5074 | 7781 | 1.000 | 1.00 | - | - | - | 1462.0 | 329.0 | 92.0 | 5074.0 | 7781.0 |

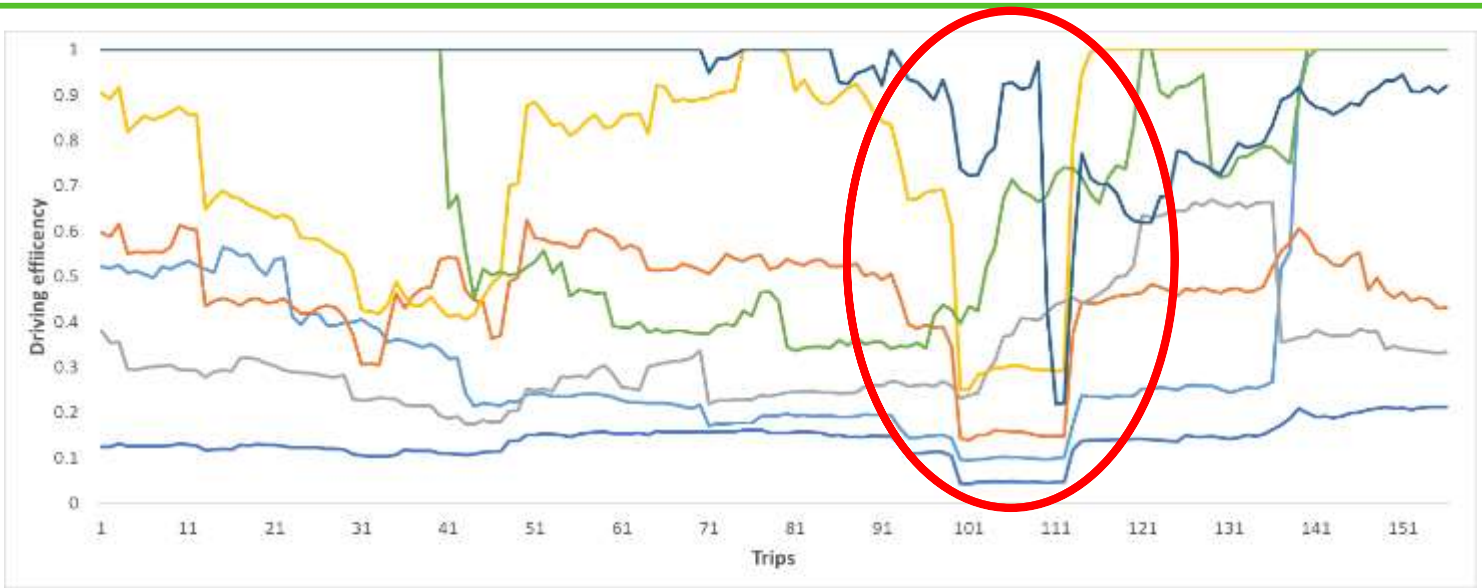
*** Inputs = ['$ha_{urban}$', '$hb_{urban}$', '$speeding_{urban}$', '$mobile_{urban}$'], Output = ['$distance_{urban}$']**

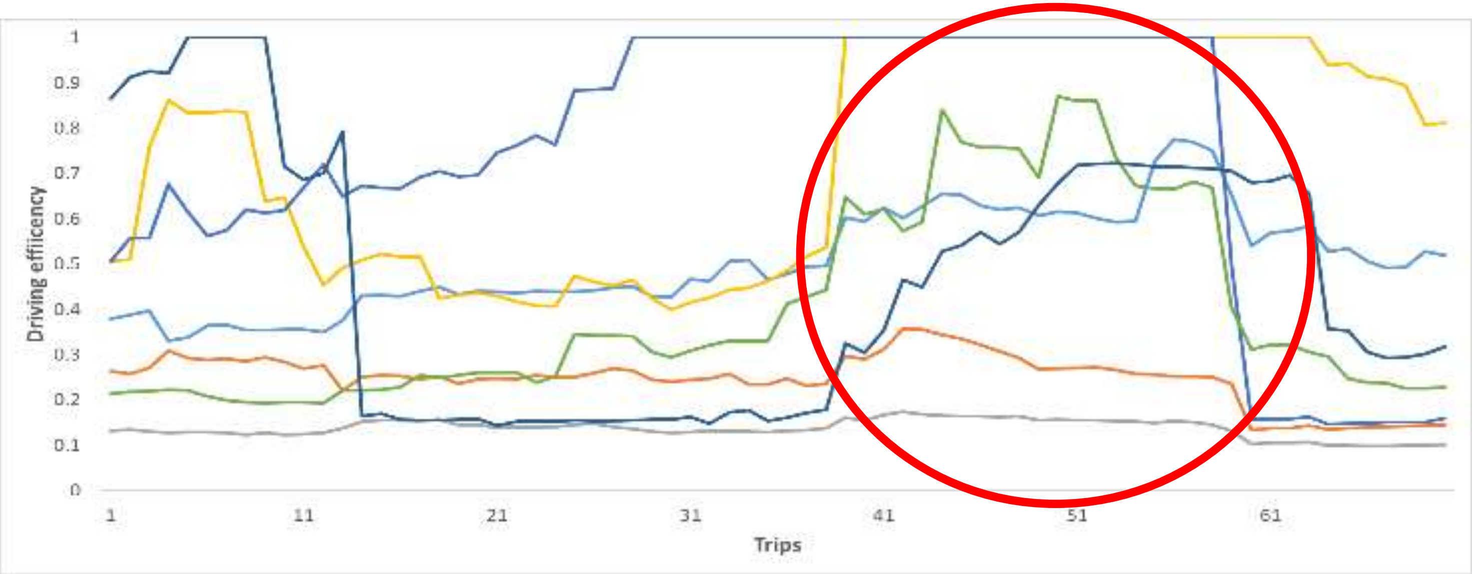
# Driver efficiency analysis - Temporal evolution (6/13)

- **Temporal evolution**
    - time window
        - Urban: 75
        - Express Urban: 81

- **Less efficient** drivers are less volatile

- Common local **minimum** and **maximum** points
    - outlier existence
    - efficiency is benchmarked



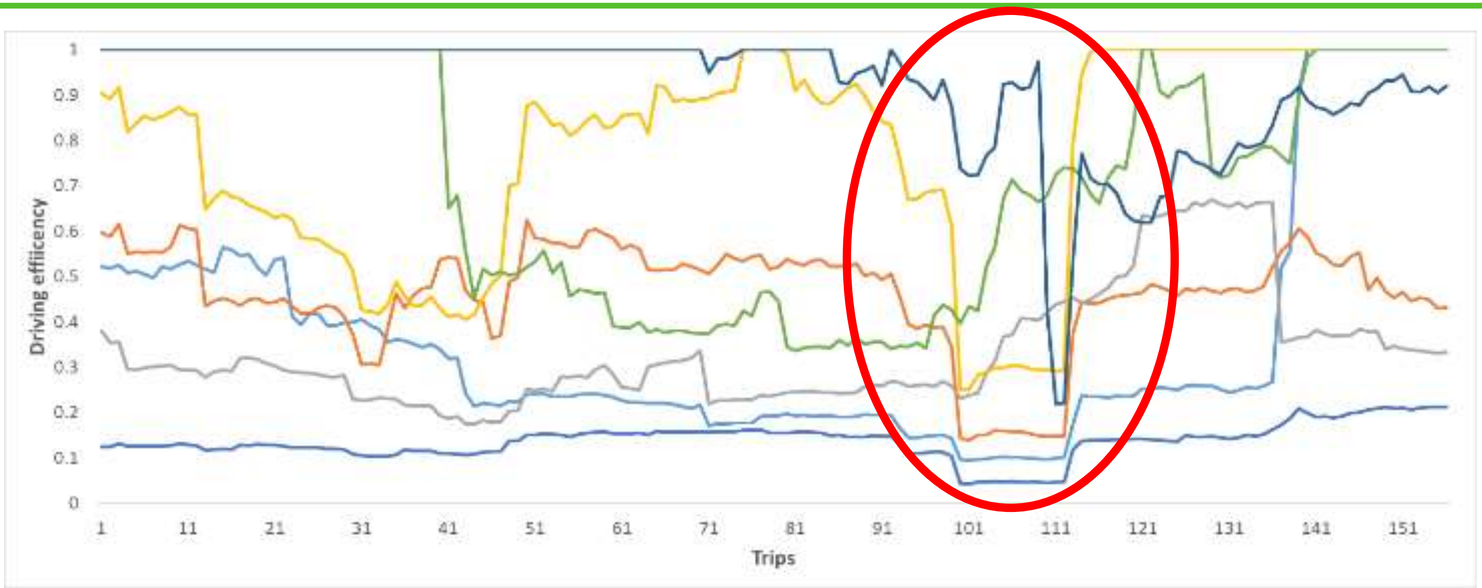
# Driver efficiency analysis - Temporal evolution (7/13)

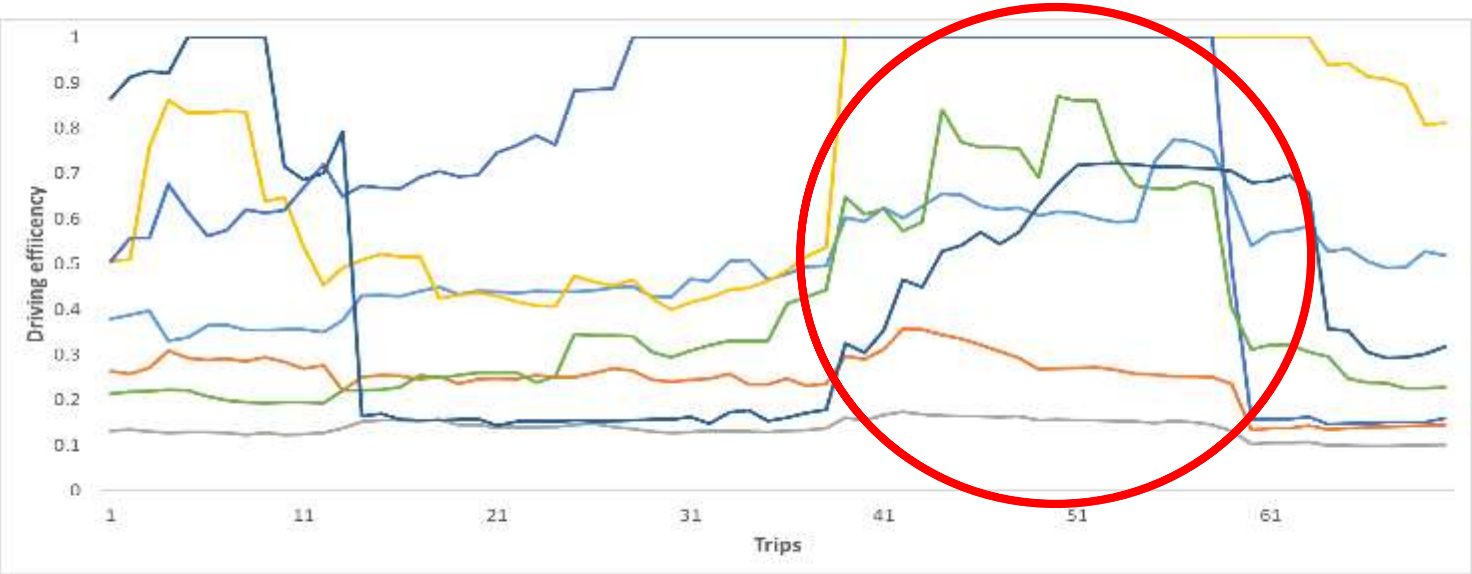
- **Time-series**

- **Stationarity**
    - no stationary urban road users
    - insignificant number of rural road users

- **Trend**
    - average approximately the same in both road types
    - higher range in rural road type

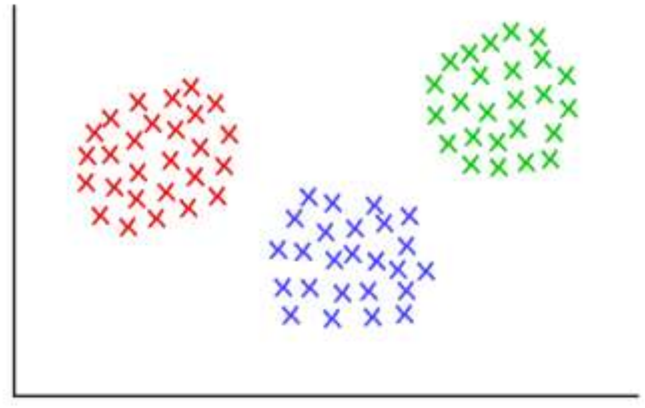
- **Volatility**
    - average approximately the same in both road types
    - higher range in rural road type



# Driver efficiency analysis - Clustering (8/13)

- **Three clusters** for both samples
    - cluster 1: typical driver
    - cluster 2: unstable driver
    - cluster 3: cautious driver



# Driver efficiency analysis - Clustering (9/13)

- **Typical** drivers:
    - high number of drivers
    - low total efficiency
    - very low positive trend
    - medium to high volatility
    - low accident frequency

| Sample type | Road type | Cluster | Trend ($*10^{-3}$) | Volatility | Efficiency | Accidents/ 10 years of driving experience |
|---|---|---|---|---|---|---|
| data_sample_1 | Urban | 1 (typical) | very low positive | medium - high | low | low - medium |
| | | 2 (unstable) | medium positive | medium - high | medium | low |
| | | 3 (cautious) | medium negative | low - medium | medium - high | low |
| | Rural | 1 (typical) | low positive | medium | low | low - medium |
| | | 2 (unstable) | high negative | high | medium - high | medium - high |
| | | 3 (cautious) | high positive | medium - high | high | low |
| data_sample_2 | Urban | 1 (typical) | very low positive | medium | low | low |
| | | 2 (unstable) | low - medium positive | medium | low | high |
| | | 3 (cautious) | medium negative | low | high | low |
| | Rural | 1 (typical) | barely no trend | medium - high | low | low |
| | | 2 (unstable) | low negative | medium | low | high |
| | | 3 (cautious) | high positive | medium - high | high | low |

# Driver efficiency analysis - Clustering (10/13)

- **Unstable** drivers:
  - efficiency
    - medium to high total for data_sample_1
    - low total for data_sample_2
  - trend
    - medium positive for data_sample_1
    - medium negative for data_sample_2
  - medium to high volatility
    - high for data_sample_1
    - medium for data_sample_2
  - medium to high accident frequency

| Sample type | Road type | Cluster | Trend (*$10^{-3}$) | Volatility | Efficiency | Accidents/ 10 years of driving experience |
|---|---|---|---|---|---|---|
| data_sample_1 | Urban | 1 (typical) | very low positive | medium - high | low | low - medium |
| | | 2 (unstable) | medium positive | medium - high | medium | low |
| | | 3 (cautious) | medium negative | low - medium | medium - high | low |
| | Rural | 1 (typical) | low positive | medium | low | low - medium |
| | | 2 (unstable) | high negative | high | medium - high | medium - high |
| | | 3 (cautious) | high positive | medium - high | high | low |
| data_sample_2 | Urban | 1 (typical) | very low positive | medium | low | low |
| | | 2 (unstable) | low - medium positive | medium | low | high |
| | | 3 (cautious) | medium negative | low | high | low |
| | Rural | 1 (typical) | barely no trend | medium - high | low | low |
| | | 2 (unstable) | low negative | medium | low | high |
| | | 3 (cautious) | high positive | medium - high | high | low |

# Driver efficiency analysis - Clustering (11/13)

- **Cautious** drivers:
  - high total efficiency
  - trend
    - medium negative for urban road
    - high positive for express urban road
  - medium volatility
    - low to medium for urban road
    - medium to high for express urban road
  - low accident frequency

| Sample type | Road type | Cluster | Trend ($*10^{-3}$) | Volatility | Efficiency | Accidents/ 10 years of driving experience |
|---|---|---|---|---|---|---|
| data_sample_1 | Urban | 1 (typical) | very low positive | medium - high | low | low - medium |
| | | 2 (unstable) | medium positive | medium - high | medium | low |
| | | 3 (cautious) | medium negative | low - medium | medium - high | low |
| | Rural | 1 (typical) | low positive | medium | low | low - medium |
| | | 2 (unstable) | high negative | high | medium - high | medium - high |
| | | 3 (cautious) | high positive | medium - high | high | low |
| data_sample_2 | Urban | 1 (typical) | very low positive | medium | low | low |
| | | 2 (unstable) | low - medium positive | medium | low | high |
| | | 3 (cautious) | medium negative | low | high | low |
| | Rural | 1 (typical) | barely no trend | medium - high | low | low |
| | | 2 (unstable) | low negative | medium | low | high |
| | | 3 (cautious) | high positive | medium - high | high | low |

# Driver efficiency analysis - Clustering (12/13)

- Attributes' values are **reducing** while shifting to a class of **higher efficiency**

- **Typical** drivers
  - metrics' values in urban roads are twice as in express urban roads

- **Cautious** drivers
  - significantly lower level of driving characteristics for all metrics

| Sample type | Road type | No of drivers | Driving characteristics | Cluster 1 (typical) | Cluster 2 (unstable) | Cluster 3 (cautious) |
|---|---|---|---|---|---|---|
| data_sample_1 | Urban | 100 | efficiency | 0.33 | 0.61 | 0.81 |
| | | | ha | 26.75 | 17.20 | 6.89 |
| | | | hb | 10.05 | 7.30 | 3.44 |
| | | | mu | 499 | 165 | 60 |
| | | | sp | 1095 | 619 | 1240 |
| | Rural | 100 | efficiency | 0.36 | 0.69 | 0.88 |
| | | | ha | 14.97 | 6.36 | 9.65 |
| | | | hb | 7.59 | 3.09 | 3.61 |
| | | | mu | 234 | 285 | 86 |
| | | | sp | 988 | 923 | 347 |
| data_sample_2 | Urban | 43 | efficiency | 0.37 | 0.33 | 1.00 |
| | | | ha | 22.05 | 41.26 | 14.13 |
| | | | hb | 8.10 | 15.39 | 5.28 |
| | | | mu | 653 | 481 | 82 |
| | | | sp | 1349 | 1140 | 436 |
| | Rural | 39 | efficiency | 0.37 | 0.39 | 1.00 |
| | | | ha | 12.35 | 19.24 | 5.74 |
| | | | hb | 6.66 | 6.84 | 2.46 |
| | | | mu | 316 | 380 | 149 |
| | | | sp | 1125 | 1149 | 415 |

# Driver efficiency analysis - Clustering (13/13)

- **Unstable** drivers of data_sample_2
  - high number of accidents
  - low driving efficiency
  - highest number of harsh acceleration and braking events
  - highest mobile use and speeding in express urban road

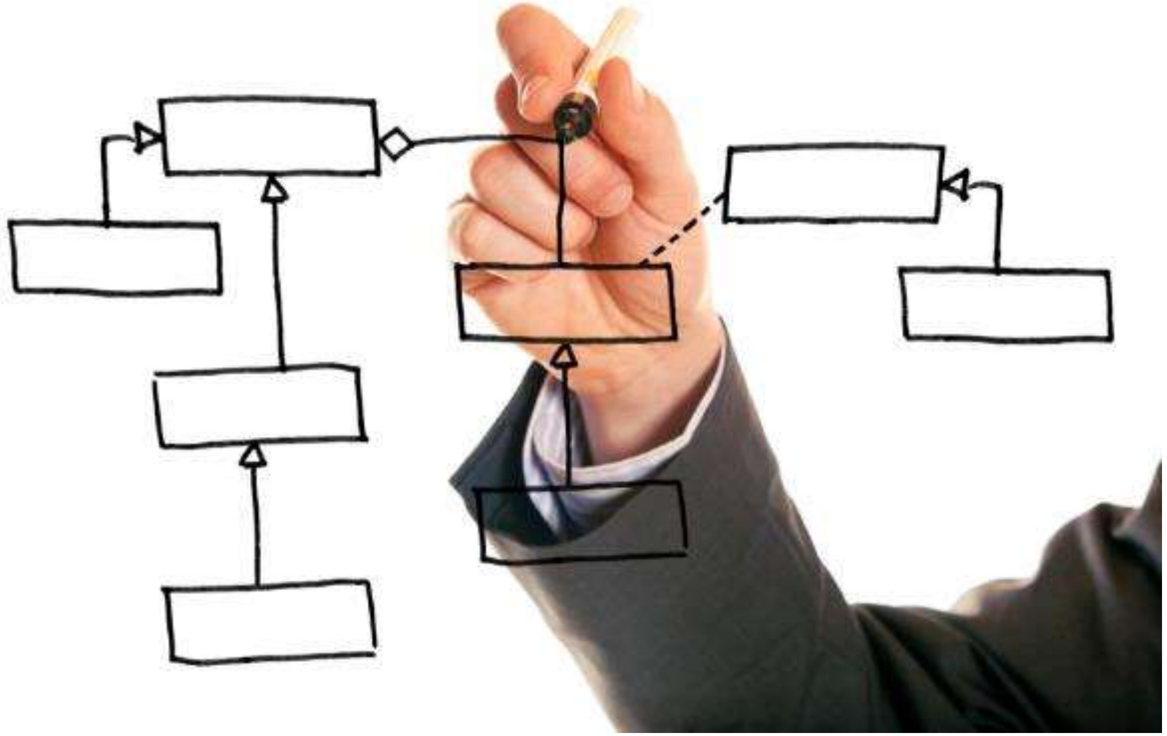
- **Non-steady** drivers with **poor** driving behaviour => most **risky** drivers

| Sample type | Road type | No of drivers | Driving characteristics | Cluster 1 (typical) | Cluster 2 (unstable) | Cluster 3 (cautious) |
|---|---|---|---|---|---|---|
| data_sample_1 | Urban | 100 | efficiency | 0.33 | 0.61 | 0.81 |
| | | | ha | 26.75 | 17.20 | 6.89 |
| | | | hb | 10.05 | 7.30 | 3.44 |
| | | | mu | 499 | 165 | 60 |
| | | | sp | 1095 | 619 | 1240 |
| | Rural | 100 | efficiency | 0.36 | 0.69 | 0.88 |
| | | | ha | 14.97 | 6.36 | 9.65 |
| | | | hb | 7.59 | 3.09 | 3.61 |
| | | | mu | 234 | 285 | 86 |
| | | | sp | 988 | 923 | 347 |
| data_sample_2 | Urban | 43 | efficiency | 0.37 | 0.33 | 1.00 |
| | | | ha | 22.05 | 41.26 | 14.13 |
| | | | hb | 8.10 | 15.39 | 5.28 |
| | | | mu | 653 | 481 | 82 |
| | | | sp | 1349 | 1140 | 436 |
| | Rural | 39 | efficiency | 0.37 | 0.39 | 1.00 |
| | | | ha | 12.35 | 19.24 | 5.74 |
| | | | hb | 6.66 | 6.84 | 2.46 |
| | | | mu | 316 | 380 | 149 |
| | | | sp | 1125 | 1149 | 415 |

# Conclusions (1/5)

- **Innovative methodological approach** for driving efficiency benchmarking as well as for estimating the efficient level of metrics of each driver.

- The **integration** of DEA with the **convex hull** algorithmic approach yielded significantly better results than the rest of the approaches tested



# Conclusions (2/5)

- The required **sampling** mileage is identified and is different for each:
  - road type
  - driving metric
  - driving aggressiveness/ risk/ distraction level

- **Not a single critical metric** to determine the required driving data amount


- **More** aggressive/ risky drivers need **less** monitoring in express urban road and highways

# Conclusions (3/5)

- **Insignificant difference** in mobile usage between drivers of medium and low efficiency classes

- The number of harsh **acceleration** events is **twice** as much as the number of harsh **braking** events

- The **shift** between efficiency classes is mainly affected by **different** driving **metrics** in urban and express urban road types


# Conclusions (4/5)

- Average **volatility** is approximately the **same** in both road types

- Average **trend** is approximately the **same** between the two road types

- **Stationarity** is similar for all drivers and road types

# Conclusions (5/5)

- There are **three** drivers' **clusters** the: a) typical, b) unstable and c) cautious drivers

- Drivers should be **continuously monitored** and re-evaluated to capture temporal shifts

- Prior **information** on driving **accident** data affects only the form of the most **unstable** drivers

# Contributions

- Develops a **methodological framework** for driving safety efficiency evaluation on trip and driver basis based on data science techniques


- **Quantifies** the **driving data** that should be collected when evaluating driving behaviour in terms of safety

- Provides insights on the main **driving** behaviour **profiles** that exist and their characteristics

- Makes use of an **innovative** smartphone **data collection** system

# Impact

- Personal and general **feedback** to drivers on
  - their overall driving efficiency and its evolution
  - an inefficient trip is performed
  - driving characteristics that should be improved
  - each road type

- **Reduce** individual driving **risk**

- Develop **insurance** pricing schemes
  - charge premiums based on driving efficiency


# Future challenges (1/2)

- Exploit a **larger** driving **sample**
  - highway road type investigation
  - safety efficiency estimation is more representative
  - DEA is less sensitive to outliers
  - metrics to distance ratio evolution
  - relationship between the aggressiveness of a driver and the necessary monitoring distance
  - known accident data record
  - investigate more on the computation time optimization


- Type of **approach**
  - macroscopic
    - shifts of drivers over long time periods
  - microscopic
    - study the spatiotemporal driving characteristics of each trip
  - combination


- **Prediction** of drivers' overall efficiency or cluster
  - provide recommendation on the level of metrics that should be reached

# Future challenges (2/2)

- **DEA limitations** – zero sum input attributes

- **Increase** the number of **attributes**
    - headways
    - lane changing
    - eye movement
    - drowsiness

- **Data recording** through
    - cameras
    - eye-tracking devices
    - radars
    - LiDARs
    - on-board-diagnostic devices (OBD)

- Study **dynamic evolution** of driving efficiency to quantify
    - how rapidly driving profiles change
    - how much driving profiles change



Thank you

# Benchmarking Driving Efficiency using Data Science Techniques applied on Large-Scale Smartphone Data


**Dimitris Tselentis**

Civil - Transportation Engineer
Ph.D. Candidate – Researcher

www.nrso.ntua.gr/dtsel/
dtsel@central.ntua.gr