# Trajectory Bayesian Indexing : The Airport Ground Traffic Case

Cynthia Delauney, Nicolas Baskiotis and Vincent Guigue

IEEE 19th International Conference on Intelligent Transportation Systems
Rio de Janeiro, Brazil

November $2^{nd}$, 2016

# Context: spatio-temporal series analysis

**Trace** = set of measures (id, time, location, *contextual info*)

**Issues :**

- Clustering/categorization [Jiang et al. 08]
- Anomaly detection [Bu et al. 09]
- **Indexing** [Guttman et al. 84, Chakka et al. 03, Zheng et al. 11]

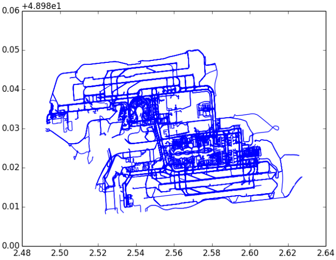
**Challenges :**

- Variable size
- Noise(s)
- Data amount

# Main goal: light & rich indexing

**Use cases:**

- **Query** What is close to a given situation?
- **Analysis** What are the common features shared by close trajectories?
- **Predict** Does the current trajectory become closer to a risk situation?

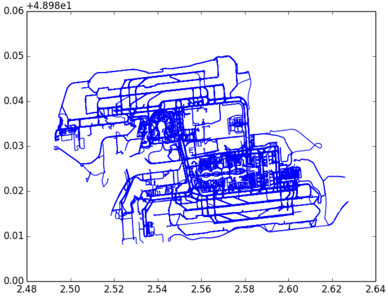



Paris international airport
*Roissy-Charles-De-Gaulle*

- Which trajectory **representation**?
- Which **metrics** between trajectories?

# Raw Data

## Whole dataset:

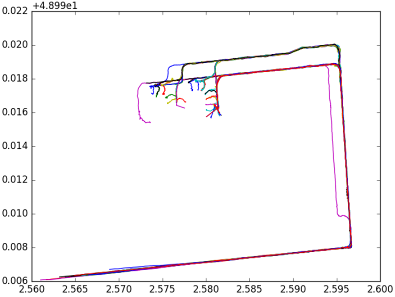



$1$ year $\sim 130\,000$ trajectories

$\sim 350$ Gb (with a rich context)

$|T_k| \sim 1000$ in average

## Trajectory samples:

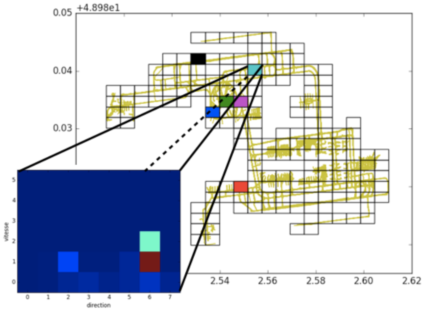



$$T_k = \{\mathbf{c}, (t_1, \ell_1, \ldots, t_{|T_k|}, \ell_{|T_k|})\}$$

$$t \in \mathbb{R}, \ell \in \mathbb{R}^2$$

$c$ : context, $t_i$ : time, $\ell_i$ : location

# Discretization & Bag of Words




$S \times 6$ velocites $\times 8$ directions

$\Rightarrow$ Fixed dimensions $Z$

$S = 30 \times 30 \Rightarrow Z = 43200$

Word definition:

$$w_i = (\ell, v, d) \in \mathbb{N}^3$$

location, velocity, direction

$$\downarrow$$

$$T_k = \{\mathbf{c}, \mathbf{w}\}, \qquad \mathbf{w} \in \mathbb{N}^Z$$

$$\downarrow$$

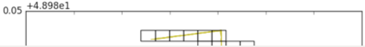
Frequency normalization:

$$w_i \Rightarrow w_i^f = \frac{w_i}{\sum_j w_j} \in \mathbb{R}_+$$

$$\downarrow$$

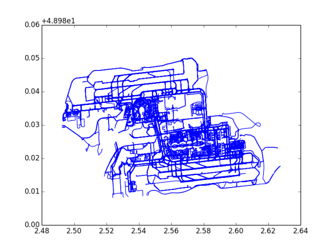
$$T_k = \{\mathbf{c}, \mathbf{w}^f\}$$

# Discretization & Bag of Words

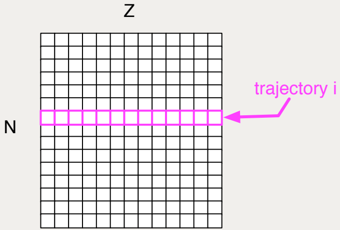



3

n




$\Longrightarrow$




$+$

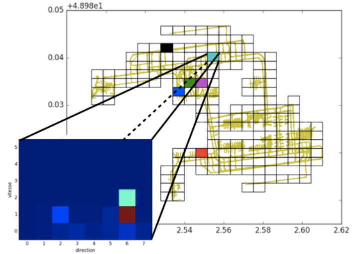
$S$

$\Rightarrow$

$$T_k = \{\mathbf{c}, \mathbf{w}^J\}$$

$$S = 30 \times 30 \Rightarrow Z = 43200$$

# Naive Bayes Modeling




$Z = 43200$

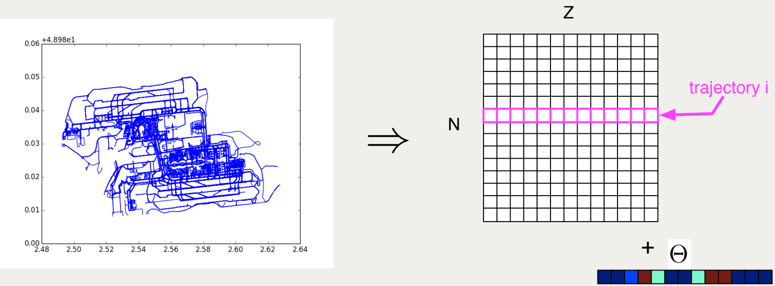
Multinomial model:

$$\Theta = \begin{bmatrix} \vdots \\ \theta_i = p(w_i|\ell) \\ \vdots \end{bmatrix} \in \mathbb{R}^Z$$

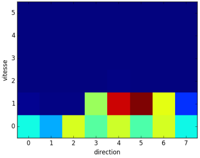
$$p(w_i|\ell) = \frac{\sum_k w_i^{(k)}}{\sum_k \sum_{\{j|\ell \in w_j\}} w_j^{(k)}}$$

# Naive Bayes Modeling

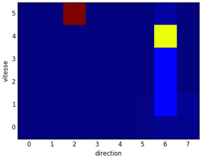
Multinomial model:



# Entropy issue : a normalization is requiered




Parking (green)
High entropy

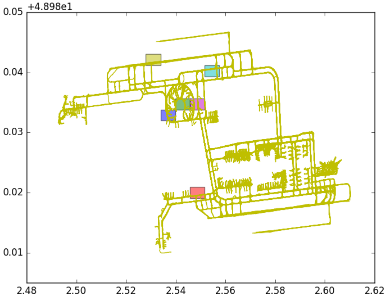


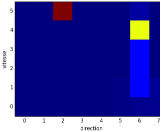

Runway (yellow)
Low entropy


Local normalization procedure :

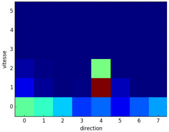
$$\underbrace{\theta_i = p(w_i|\ell)}_{\text{likelihood}} \Rightarrow \underbrace{\theta_i = \frac{p(w_i|\ell)}{\max_{\{i|\ell\in w_i^k\}} p(w_i|\ell)}}_{\text{locally norm. likelihood}}$$

# Local behavior descriptions

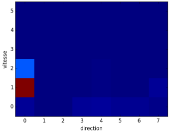


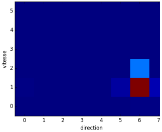

Yellow



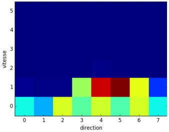

Blue



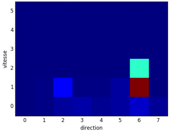

Magenta




Cyan



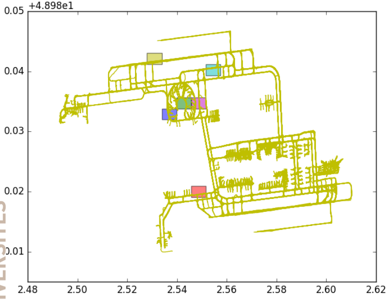

Green



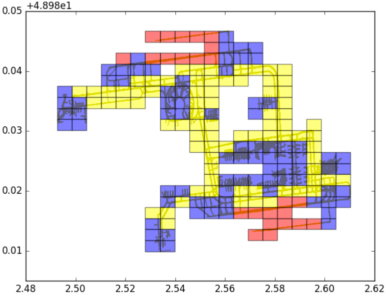

Red

# Local behavior descriptions

Spatial caracterization:



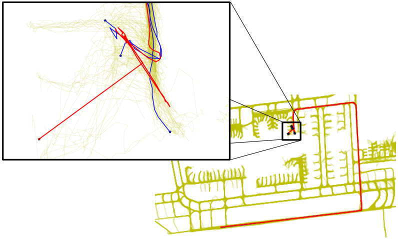

$\Rightarrow$



# Query examples

**Simple framework:**

- Query : 1 trajectory
- Answers : $k(= 3)$ Nearest Neighbors (Euclidian distance)

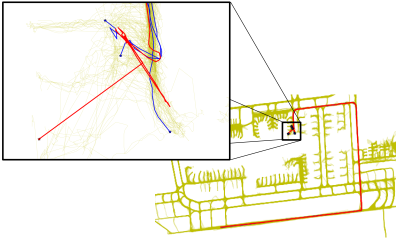


*Query in the original representation space*

# Query examples

**Simple framework:**

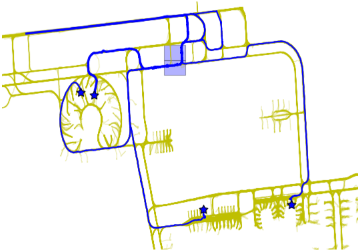
- Query : 1 trajectory
- Answers : $k(=3)$ Nearest Neighbors (Euclidian distance)



*Query in the original representation space*

***Smart* query:**

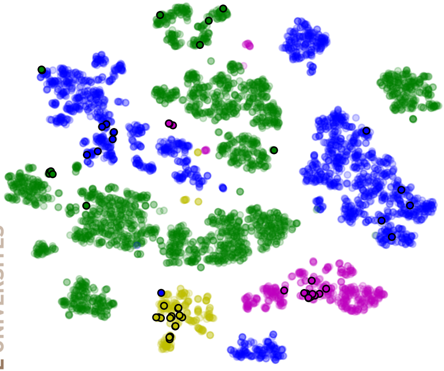
- Query = region $\ell$ (all velocit./dir.)
- Sorted answers:
  4 Lowest likelihood



*Query in representation space + likelihood*

# Consistency of the representations

## 1 dot = 1 (take-off) trajectory



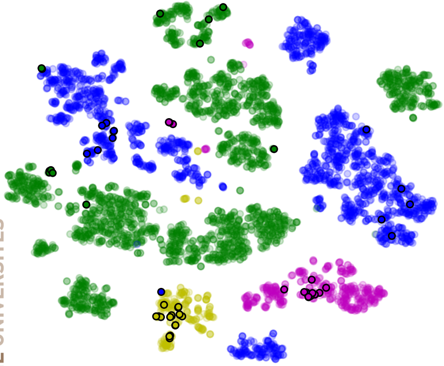
T-SNE projection (2D)

- Unsupervised learning... difficult to evaluate
- Colors = airport configurations
  - 4 runways
  - East or west direction

⇒ **Clear latent space division**

# Consistency of the representations

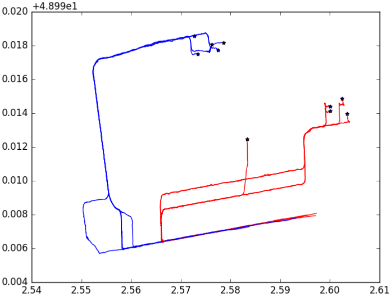
1 dot = 1 (take-off) trajectory



T-SNE projection (2D)

Fine analysis of the magenta cluster:

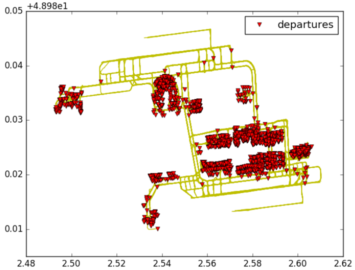
- left sub-cluster
- right sub-cluster

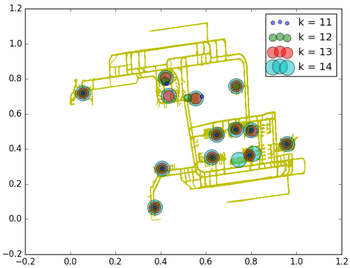


# [Parallel exp.] Finding late trajectories

Protocol :

1. Clustering of the parkings



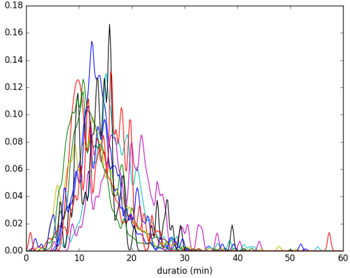

⇒ 10 clusters

# [Parallel exp.] Finding late trajectories

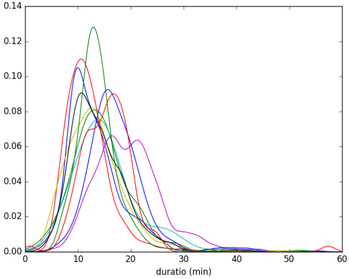
Protocol :

1. Clustering of the parkings
2. Taxiing duration pdf estimate

Raw estimate




Smoothed estimate (Parzen)

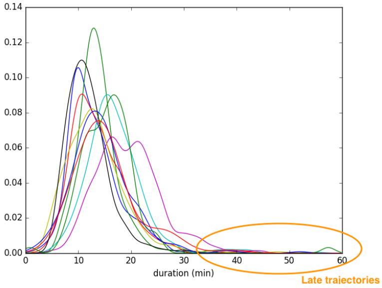


# [PARALLEL EXP.] FINDING LATE TRAJECTORIES

Protocol :

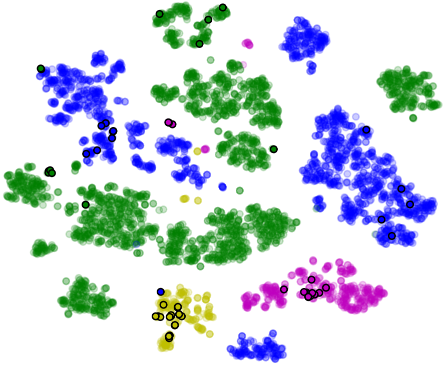
1. Clustering of the parkings
2. Taxiing duration pdf estimate
3. Late = last percentile

Smoothed estimate (Parzen) + last percentiles of each cluster



# Lateness topology

## Circled dot = late trajectory



T-SNE projection (2D)

We detect some regularities in late trajectories

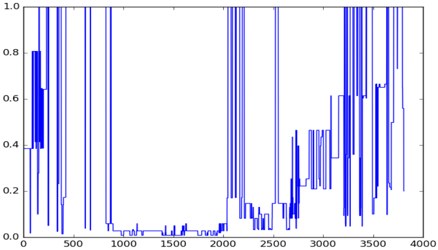
Outliers (often) correspond to late trajectories

# Single trajectory likelihood

(Re-)introducing **time** in the analysis:
Trajectory = series of words ⇒ series of likelihoods

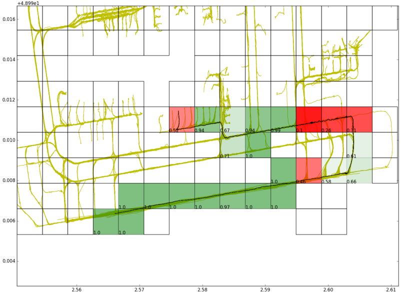
$$T = \{w_{t_1}, \dots, w_{t_{|T|}}\} \Rightarrow \{\mathcal{L}(w_{t_1}), \dots, \mathcal{L}(w_{t_{|T|}})\}$$

Likelihood course of a late trajectory:

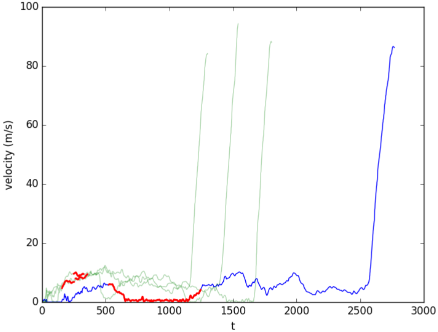


# Single trajectory likelihood

## Spatial mapping

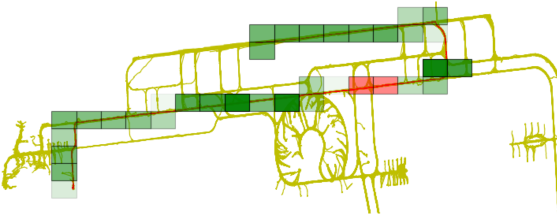



## Velocity mapping



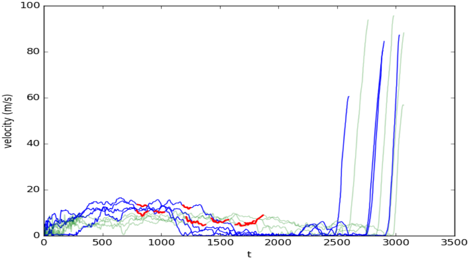

The plane had an abnormal **low** velocity in 3 spatial tiles of the grid

# Smart query



Finding trajectories with:

anomaly in the region $\ell$
& velocity $>$ ML velocity



# Conclusion & Perspectives

## Conclusion

- **Very light** way to index trajectories
- Consistent
- (Local) **likelihood**
- Many possible coding (presence, frequency, tf-idf...)

*inspired from text indexing*

## Perspectives

- Indexing ⇒ categorization with **continuous modeling** (neural network)
- Identifying **precursory events** of abnormal situations
- Trajectory ⇒ **Situation** (multiple vehicles)

*bigram?*

Many thanks to AWACS FUI Grant partners: SafetyLine, IFSTTAR, ADP