Mining Constrained Regions of Interest: An optimization approach

Alexandre Dubray Guillaume Derval Siegfried Nijssen Pierre Schaus

Table of contents

Introduction

Motivations

- The amount of spatiotemporal data is exploding (smartphone applications, sport devices, fleet management, etc)
- There is a need to process more efficiently these data
- Rewrite the raw trajectories (GPS points) as sequence of *Regions of Interest* (ROI)
- Multiple applications:
    - Trajectory pattern mining
    - Next location prediction
    - Urban management
    - ...

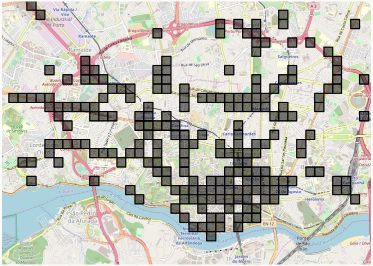
Preparation of the data

1. Divide the map with a $N \times M$ grid of square cells.
2. Assign a density value to each cell. A cell is dense if its density is above a threshold
3. Express the ROI as an aggregation of dense cells

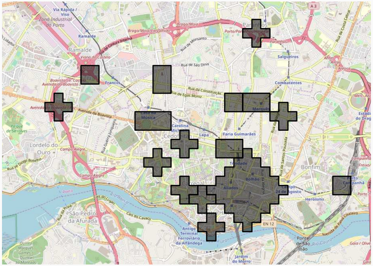
Multiple possibility for the density

- Number of trajectories that cross the cell (with and without interpolation)
- Number of trajectories that stayed at least 10 minutes in a cell
- etc.

Example of ROIs



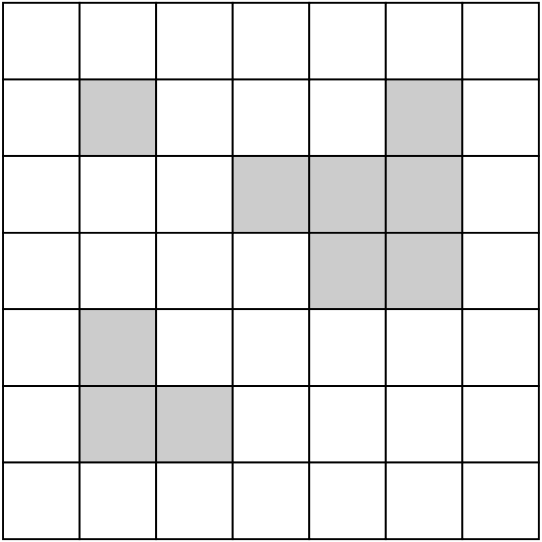
(a) Initial set of dense cells



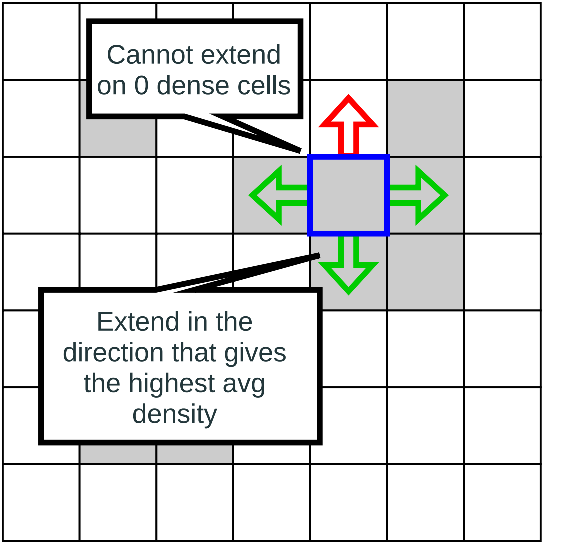
(b) Solution found by our method

PopularRegion

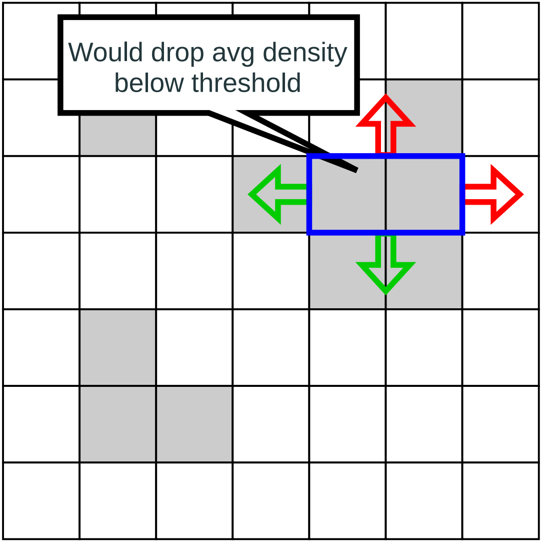
Execution of the algorithm



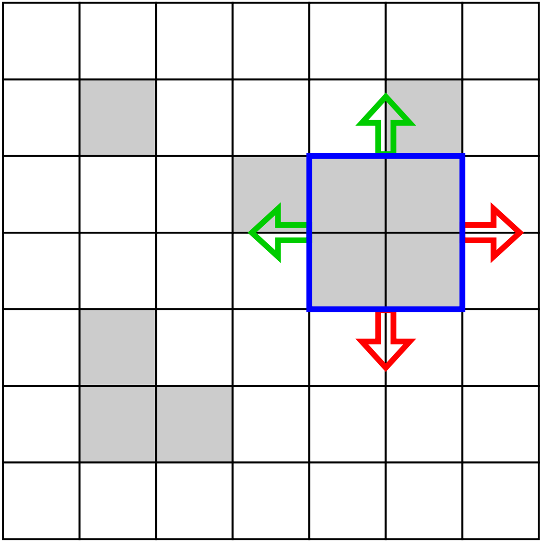
Execution of the algorithm



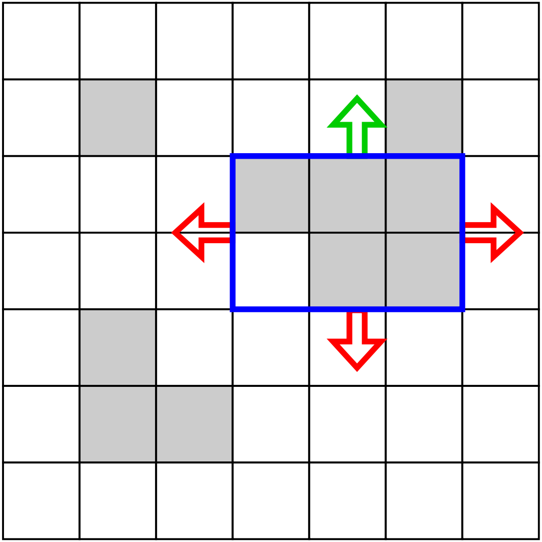
Execution of the algorithm



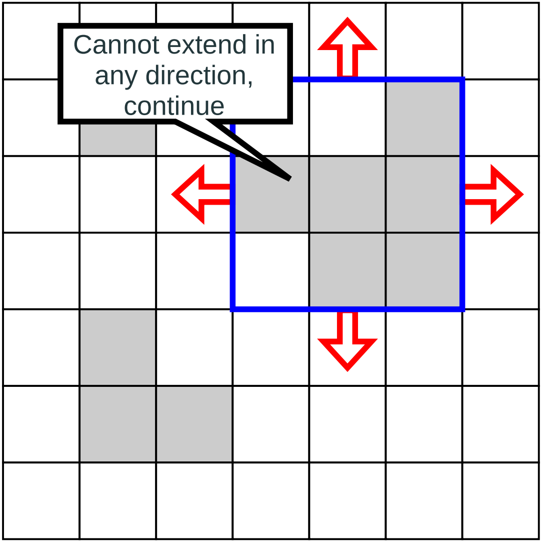
Execution of the algorithm



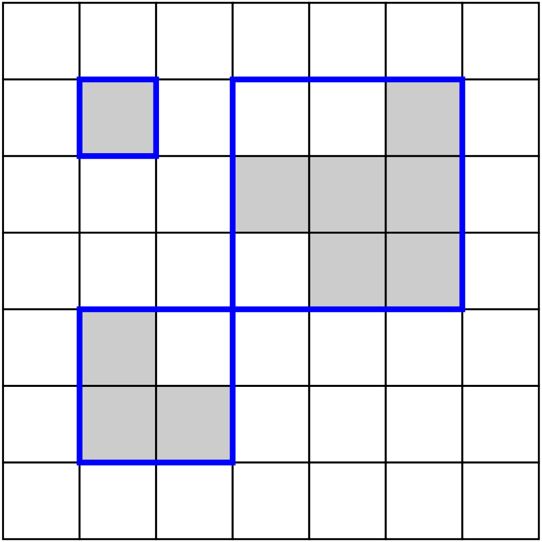
Execution of the algorithm



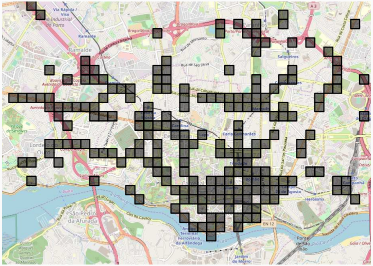
Execution of the algorithm



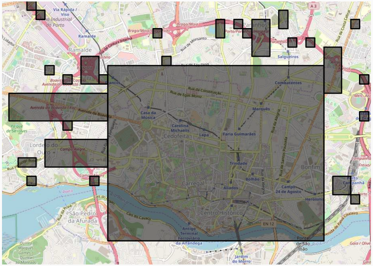
Execution of the algorithm



Result of the algorithm



(a) Initial set of dense cells



(b) Solution with 5% min average density

Advantages and disadvantages

- Scalable
- Intuitive and good results for most configurations

But...

- No formalization of the output
- Only rectangular regions
- Does not easily accept background knowledge

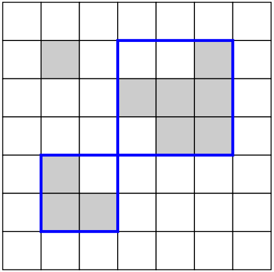
Our method

Outline

1. Generate a set of candidate ROI
    - Can have any shape
    - Impose *intra-ROI* constraints
2. Select from the candidates K final ROIs
    - Found by an optimization problem
    - Impose *inter-ROI* constraints

ROIs as an encoder

- The ROIs encode the dense status of the cells
- Example of encoding with two rectangles
    - 1 dense cells is not covered
    - 4 non-dense cells are covered
    - The encoding makes 5 errors
- We prefer encoding with less errors



Formalization of the problem (1)

Some notations:

- Let $\mathcal{G}$ be the grid, $\mathcal{G}^*$ the set of dense cells, $\mathcal{S}$ a set of candidates and θ the minimum density threshold
- d_i (resp. u_i) is the number of dense (resp. non-dense) cells covered by the candidate $R_i \in \mathcal{S}$
- K is the number of ROIs we want to find

A first optimization model

$$
\begin{aligned}
&\text{minimize } |\mathcal{G}^*| + \sum_{R_i \in \mathcal{S}} x_i \cdot (u_i - d_i) \\
&\text{subject to} \\
&\qquad \textstyle\sum_{R_i \in \mathcal{S}} x_i \leq K \\
&\textstyle\sum_{R_i \in \mathcal{S} | c \in R_i} x_i \leq 1 \qquad \forall c \in \mathcal{G} \\
&\qquad\qquad x_i \in \{0, 1\} \quad \forall R_i \in \mathcal{S}
\end{aligned}
$$

Formalization of the problem (2)

In practice how to set the K? Use the Minimum Description Length Principle!

- Let $Sol \subseteq \mathcal{S}$ be a valid solution
- Length of the errors:

$$L(\mathcal{G} \mid Sol) = 2|\mathcal{G}^*| + \sum_{R_i \in Sol} 2(u_i - d_i)$$

- Length of the model:

$$L(Sol) = \sum_{R_i \in Sol} size(R_i)$$

- Minimum Description Length principle tells that the best solution is:

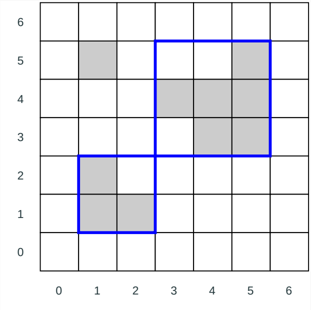
$$\arg\min_{Sol \in \mathcal{S}} L(\mathcal{G} \mid Sol) + L(Sol)$$

The final optimization model

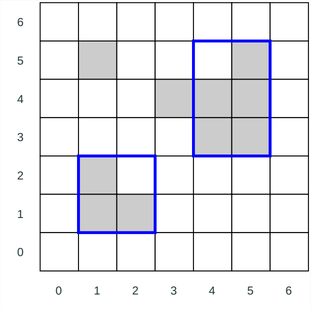
$$
\begin{aligned}
&\text{minimize} \sum_{R_i \in \mathcal{S}} x_i \cdot (2(u_i - d_i) + size(R_i)) \\
&\text{subject to} \\
&\textstyle\sum_{R_i \in \mathcal{S} | c \in R_i} x_i \leq 1 \qquad \forall c \in \mathcal{G} \\
&\qquad\qquad x_i \in \{0, 1\} \quad \forall R_i \in \mathcal{S}
\end{aligned}
$$

Example

- $L(\mathcal{S}) = 4 + 4 = 8$
- $L(\mathcal{G} \mid \mathcal{S}) = 2 \cdot (1 + 4) = 10$
- Total length of this model is $8 + 10 = 18$

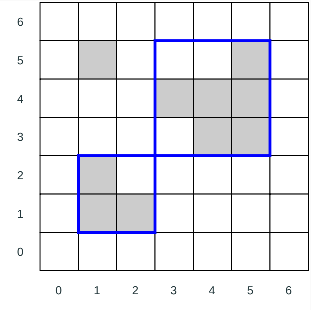


- $L(\mathcal{S}) = 4 + 4 = 8$
- $L(\mathcal{G} \mid \mathcal{S}) = 2 \cdot (2 + 2) = 8$
- Total length of this model is $8 + 8 = 16$

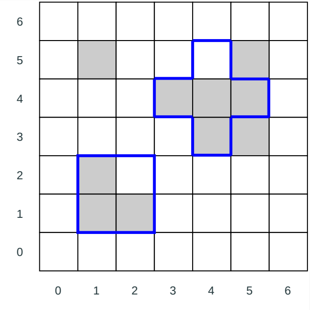


Example with circles

- $L(\mathcal{S}) = 4 + 4 = 8$
- $L(\mathcal{G} \mid \mathcal{S}) = 2 \cdot (1 + 4) = 10$
- Total length of this model is $8 + 10 = 18$



- $L(\mathcal{S}) = 4 + 3 = 7$
- $L(\mathcal{G} \mid \mathcal{S}) = 2 \cdot (2 + 3) = 10$
- Total length of this model is $7 + 10 = 17$



The full method

1. Generate the set of candidates $\mathcal{S}$ (e.g. enumerate all distinct rectangle on the grid)
    - Candidate can have any shape
    - Compute their contribution to the description length
    - Apply *intra-ROI* constraints to filter the candidate set
2. Solve the optimization model
    - Apply *inter-ROI* constraints
    - Model these constraint with linear constraints
3. The result of the optimization (choice of candidates) is the set of ROIs we return

Experiments

Setup

- Two versions of our method
  - With only rectangular regions
  - With rectangular and circular regions
- Showing results on Kaggle taxis dataset ($\approx$1.6 million trajectories)
- Comparing with PopularRegion and OPTICS (when clustering the dense cells)

Execution time

Minimum density threshold	2%			5%		
Grid side size	100	150	200	100	150	200
Number of dense cells ($\lvert\mathcal{G}^*\rvert$)	571	597	537	230	178	137
Number of ILP candidates	23 814	7 779	3 399	2 880	1 232	434
ILP optimization time (s)	4.328	0.464	0.109	0.113	0.044	0.029
PopularRegion run time (s)	0.003	0.005	0.006	0.002	0.003	0.004
OPTICS run time (s)	0.209	0.222	0.200	0.084	0.065	0.051

Description Length

- For high density threshold, number of errors becomes similar
- ILP-based methods produce smaller models
- Overall the Description Length is inferior for ILP-based methods

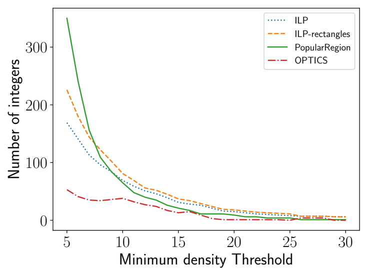


Figure 3: Encoding of the errors

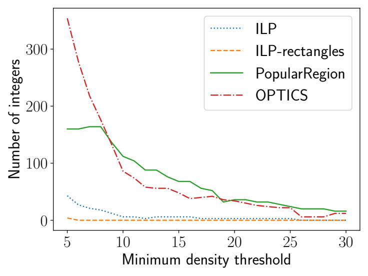


Figure 4: Encoding fo the models

Robustness to noise

- Start from a 100×100 grid
- Move every element of the trajectories to a neighboring cell with a probability p
- Choose the new cell randomly in a square of size 10 around the initial cell
- Recompute solution and compare to initial solution (with min density threshold 5%)

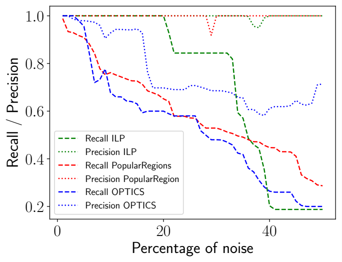


Figure 5: Recall and precision

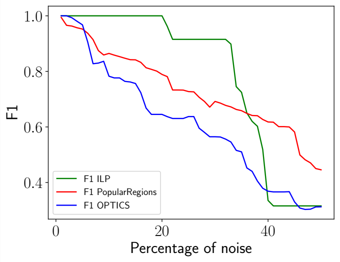


Figure 6: F1-measure

Conclusion

- We propose an optimization model to find K ROIs from trajectory data
- Using the MDL principle we can automatically set the K
- Our method is more flexible than specific method since it accepts a wide range of constraints
- the solutions find by our method are more robust and better generalize the dense cells distribution
- The runtime of the ILP becomes reasonnable as long as there is not too much candidates