SAC 2020 - 35th ACM/SIGAPP Symposium On Applied Computing

Sampling strategies to create moving regions from real world observations

Authors

Rogério Luís C. Costa, Enrico Miranda, Paulo Dias and José Moreira

Acknowledgments This work is partially funded by National Funds through the FCT - Foundation for Science and Technology, in the context of the projects UID/CEC/00127/2013 and POCI-01-0145-FEDER-032636.

IEETA
Institute of Electronics and
Informatics Engineering of Aveiro

University of Aveiro

Spatiotemporal Data Representation Models

- **Discrete model**
  - Snapshot time + corresponding observation

1	2019-10-04 15:34:00	POLYGON((417 140,395 151,349 ...
2	2019-10-07 04:26:00	POLYGON((470 136,433 140,429 ...
3	2019-10-13 05:15:00	POLYGON((896 1,893 32,850 34,7...
4	2019-10-15 21:51:00	POLYGON((1234 50,1205 76,1208...

- Spatial and temporal resolutions may lead to accuracy, storage and performance problems
- How to represent objects' properties at in-between observations?

Spatiotemporal Data Representation Models

- **Continuous model**
  - Abstract Data Types: *moving points*, *moving lines*, *moving regions*, …
  - Sliced representation[1] - *moving object* is represented by:

    (i) geometry + (ii) function (to represent the object's evolution)

Region interpolation problem: With object representations at two timestamps, a *function* is used to interpolate the representations

[1]Forlizzi, L., Güting, R. H., Nardelli, E., & Schneider, M. (2000). A data model and data structures for moving objects databases. *Proceedings of the 2000 ACM SIGMOD* 319–330.

*Image adapted from: Moreira, J., Duarte, J., & Dias, P. *Modeling and Representing Real-World Spatio-Temporal Data in Databases*.

The Problem

University of Aveiro

- **How to use the continuous model to represent real world data?**



How to represent burnt area evolution obtained in aerial bush-fires captured by drones?

Dataset (Basic) Preparation Workflow

University of Aveiro

Data acquisition

- Approximately 15 minutes
- 25 frames per second
- More than 22,500 observations

Object segmentation

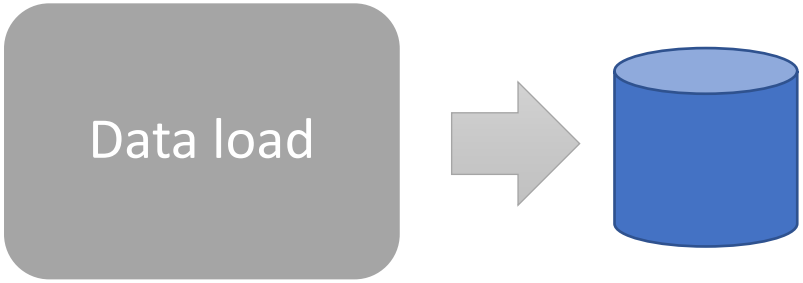
- Extract frames
- Identify the Object of Intereset (i.e. burnt area) at each frame
- Create WKT representations of burnt area contours

Data load

Create the moving region representation

How to represent the objects?

University of Aveiro



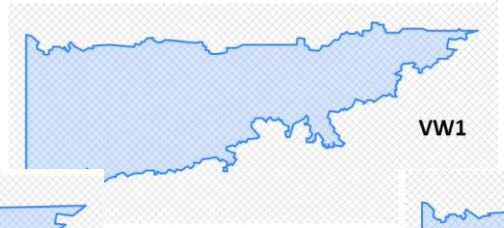
Load polygons' representations (WKT) into DBMS

Average number of points per geometry: 2,000

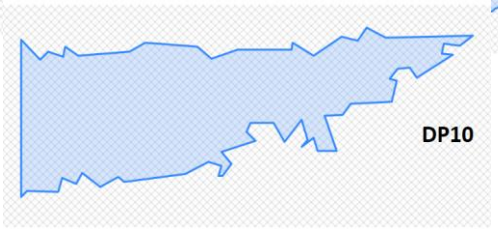
More than 43 million points to represent all observations

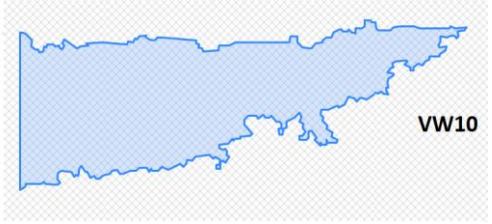
Can we use simplification algorithms without impacting in the interval specification or in interpolation results? Which are the algorithm and simplification level to use?

DP: Simplified with Douglas–Peucker algorithm
DP1: tolerance 1
DP10 : tolerance 10



VW: Simplified with Visvalingam-Whyatt algorithm
VW1: tolerance 1
VW10 : tolerance 10





How to represent the objects?

University of Aveiro

We built four datasets - DP1, DP10, VW1 and VW10 – using distinct simplification algorithms and compared such data with the original data representation in terms of….

Number of points

	DP1	DP10	VW1	VW10
min (% of original)	37	3	79	21
max (% of original)	41	6	79	24
avg (% of original)	40	4	79	22

Jaccard Index (JI)

	DP1	DP10	VW1	VW10
min	0.992	0.941	0.998	0.987
max	0.998	0.983	0.999	0.996
avg	0.996	0.967	0.999	0.993

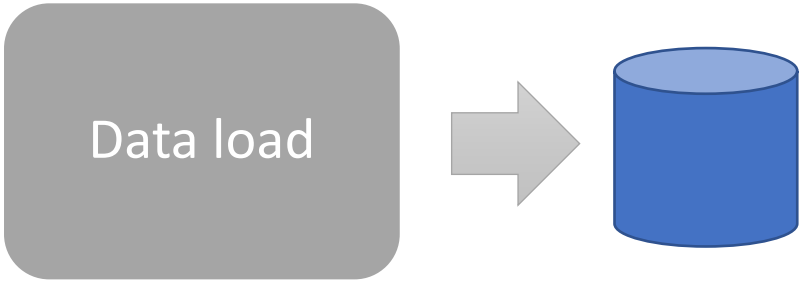
Hausdorff Distance (HD)

	DP1	DP10	VW1	VW10
min	1	8	1	3
max	1	10	49	49
avg	1	9	3	7

- DP10 leads to the largest simplification (simplified data has just 4% of original number of points), but to the lowest JI (with avg value of 0.967)
- VW simplification leads to local deformation (highset HD), but also to high values of JI (over 0.99)

Which data to load?

University of Aveiro



Load polygons' representations (WKT) into DBMS

More than 22,500 observations

Can we use just some of the observations and still get a reasonable representation?
How to select which observations to use?

Which data to load?

University of Aveiro

To select the key observations

- We propose that the *dissimilarity distance* between two selected representations should not be *too large*
  - Otherwise, may have poor quality interpolations

In our algorithm

- We compare geometries: P[i] and P[j]
- Compute the dissimilarity distance
- If geometries are too distant,

  then we select P[i], P[j - 1]

Input:
observations: set of observations
Distance: dissimilarity function
α: tolerance threshold
numObservations: number of observations
Output:
time slices identified in the set of observations
Method:
$i \leftarrow 1$
$j \leftarrow i + 1$
while $j \leq numObservations$ **do**
  $d_{i,j} \leftarrow Distance(observations(i).obs, observations(j).obs)$
  if $d_{i,j} > \alpha$ **then**
    $observations(i).beginInterval \leftarrow true$
    $observations(j-1).endInterval \leftarrow true$
    $i \leftarrow j$
  end if
  $j \leftarrow j + 1$
end while
$obs(i).beginInterval \leftarrow true$
$obs(j-1).endInterval \leftarrow true$

Which data to load?

University of Aveiro

Dissimilarity distance-based example

- Some representations on the evolution of an object
  - Consider centroids are aligned
- Using Jaccard Distance as the distance metric

Is the Jaccard Distance between these two geometries greater than α?

No / Yes

Experimental evaluation

University of Aveiro

For each dataset: original, DP1, DP10, VW1 and VW10, we

Selected key observations

- Using dissimilarity distance-based observation selection
- Evaluating several values for the distance limit threshold α

Created simulated representations

- Using PySpatioTemporalGeom[1], we created simulated representation of not selected observations

Compared simulated data with original one

- Using Jaccard Index to evaluate geometry similarity

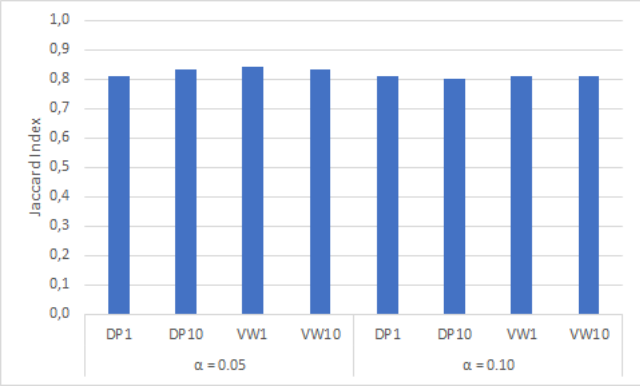
[1] Mckennney, M., & Frye, R. (2015). Generating moving regions from snapshots of complex regions. *ACM Transactions on Spatial Algorithms and Systems*, *1*(1), 1–30.

Experimental evaluation

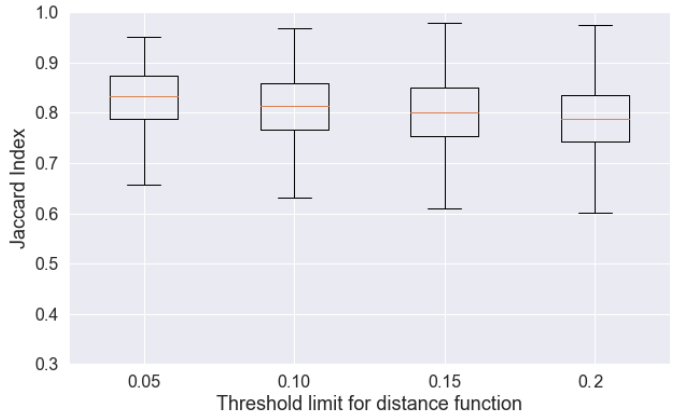
University of Aveiro

Number of selected observations for dataset and per value of α

Average Jaccard Index – Simulated geometries vs original ones – Evaluating simplification methods



Jaccard Index – Simulated geometries vs original ones – Evaluating several values of α (VW10 dataset)

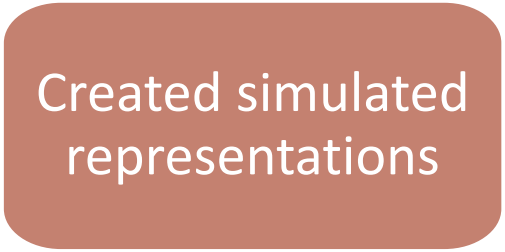


Experimental evaluation

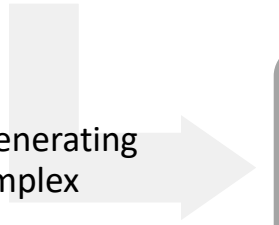
University of Aveiro

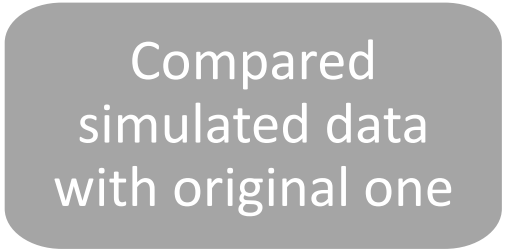
- Alternative approach for observation selection: **Fixed-size - Equidistant sampling**
- For each dataset: original, DP1, DP10, VW1 and VW10, we

- Using fized-size strategy
- Evaluating several size values



- Using PySpatioTemporalGeom[1], we created simulated representation of not selected observations





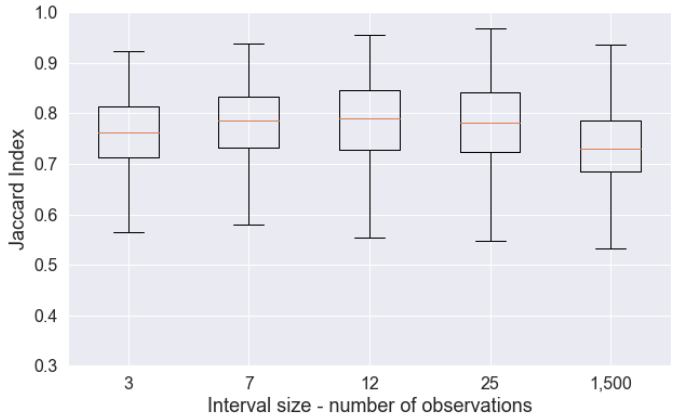
- Using Jaccard Index to evaluate geometry similarity

[1] Mckennney, M., & Frye, R. (2015). Generating moving regions from snapshots of complex regions. *ACM Transactions on Spatial Algorithms and Systems*, *1*(1), 1–30.

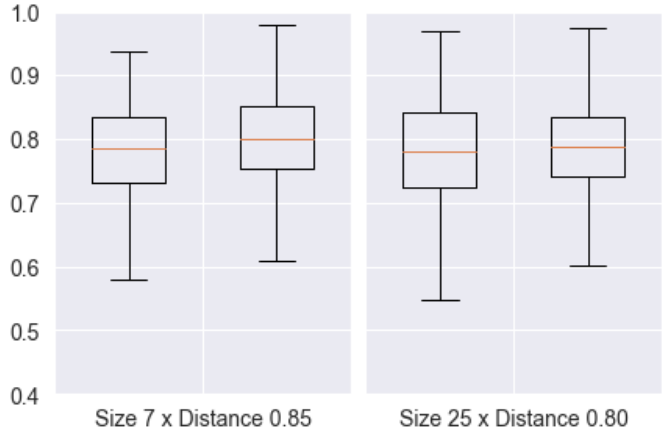
Experimental evaluation

University of Aveiro

Jaccard Index – Simulated geometries vs original ones –
VW10 dataset and several values of α (uniform sampling)



Jaccard Index – Simulated geometries vs original ones -
Uniform Sampling and Dissimilarity-based (VW10 dataset)

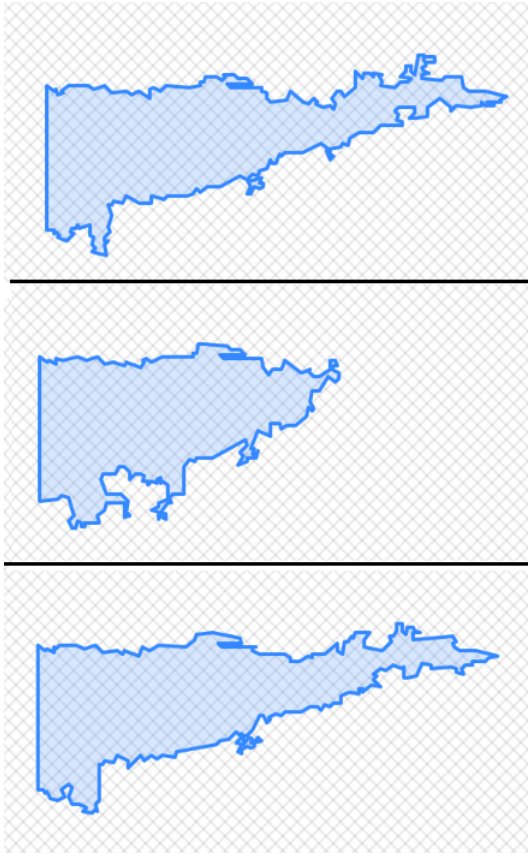


Experimental evaluation

University of Aveiro

- Distance-based strategy created “single frame intervals“
  - Some geometries are too different from its previous and following ones – i.e. outliers!?

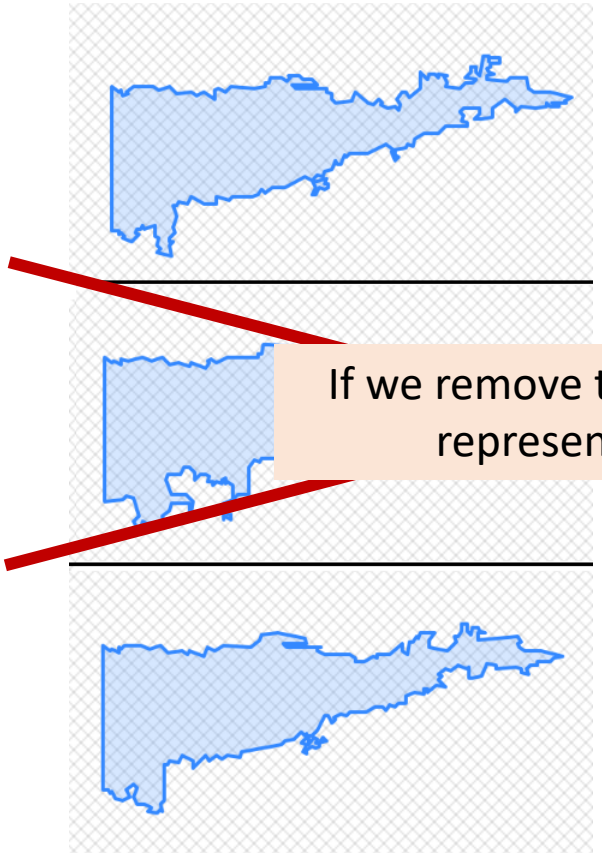
Innacurate object detection during fragmentation – smoke influence



Experimental evaluation

University of Aveiro

- Distance-based strategy created “single frame intervals“



Obs 788 (last observation of an interval starting in obs 779)

If we remove the innacurate representation….

urate

on

Obs 790 (first observation of an interval that ends in 811)

Distance-based strategy would merge intervals - creating a just one from observations 779 to 811

Conclusions

In this work, we

- Present a real-world dataset based on images captured by drones
- Evaluate the use of geometry simplification algorithms
- Propose a new algorithm to select the key observations to be used
- Evaluate the use of proposed algorithm over datasets with distinct levels of simplification
- Compare the proposed observation selection strategy to another one
- Use an interpolation function to simulate geometries and compare them with the real ones

Conclusions

- Interpolation generates in-between representations that are similar to real ones
- Dissimilarity distance-based strategy
  - Achieves good compression rates (e.g. 8% of original data when $\alpha = 0.2$)
  - Leads to better interpolation quality than uniform sampling
  - Helps on the identification of possibly innacurate data
- Very large simplification (e.g. DP10 dataset) influences on the number of intervals when using low values of α (e.g. 0.05)

Future Work

- Study how to detect outliers and noise data for a better continuous representation
- Apply obtained parameters as part of studies on fire propagation
  - I.e., to simulate burnt area evolution when there is no real data

Thank you!

Project website
http://most.web.ua.pt/