

DE VINCI RESEARCH CENTER

ESILV
DE VINCI ENGINEERING SCHOOL

DE VINCI HIGHER EDUCATION

FROM FLOWS TO GRAPHS: DATA-DRIVEN INSIGHTS ON LATENT OVERTOURISM WITH FREQUENT PATTERN MINING

Hugo Alatrista-Salas, Gaël Chareyron, Sonia Djebali, Imen Ouled-Dlala, **Nicolas Travers**
ESILV, DVRC Laboratory, Paris, France

DOING @ ADBIS'25
Tampere, Finland

OVERTOURISM

Overtourism describes destinations where hosts or guests, locals or visitors feel that there are too many visitors and that the quality of life in the area, or the quality of the experience has deteriorated unacceptably

[Butler R., Overtourism, 2019]

https://responsibletourismpartnership.org/overtourism/

Latent overtourism refers to destinations showing early warning signs before tipping points

[Adie, Falk & Savioli, Current issues in Tourism, 2020]



https://www.solimarinternational.com/what-is-overtourism-and-how-can-we-prevent-it/

ISSUE

Overtourism

- Phenomenon with huge impact on society (GDP)
- Well-know POIs which witness overtourism
- *Can we measure the impact of overtourism on the neighbourhood ?*

Tourists' evolution (prescriptions) within destinations can witness this impact

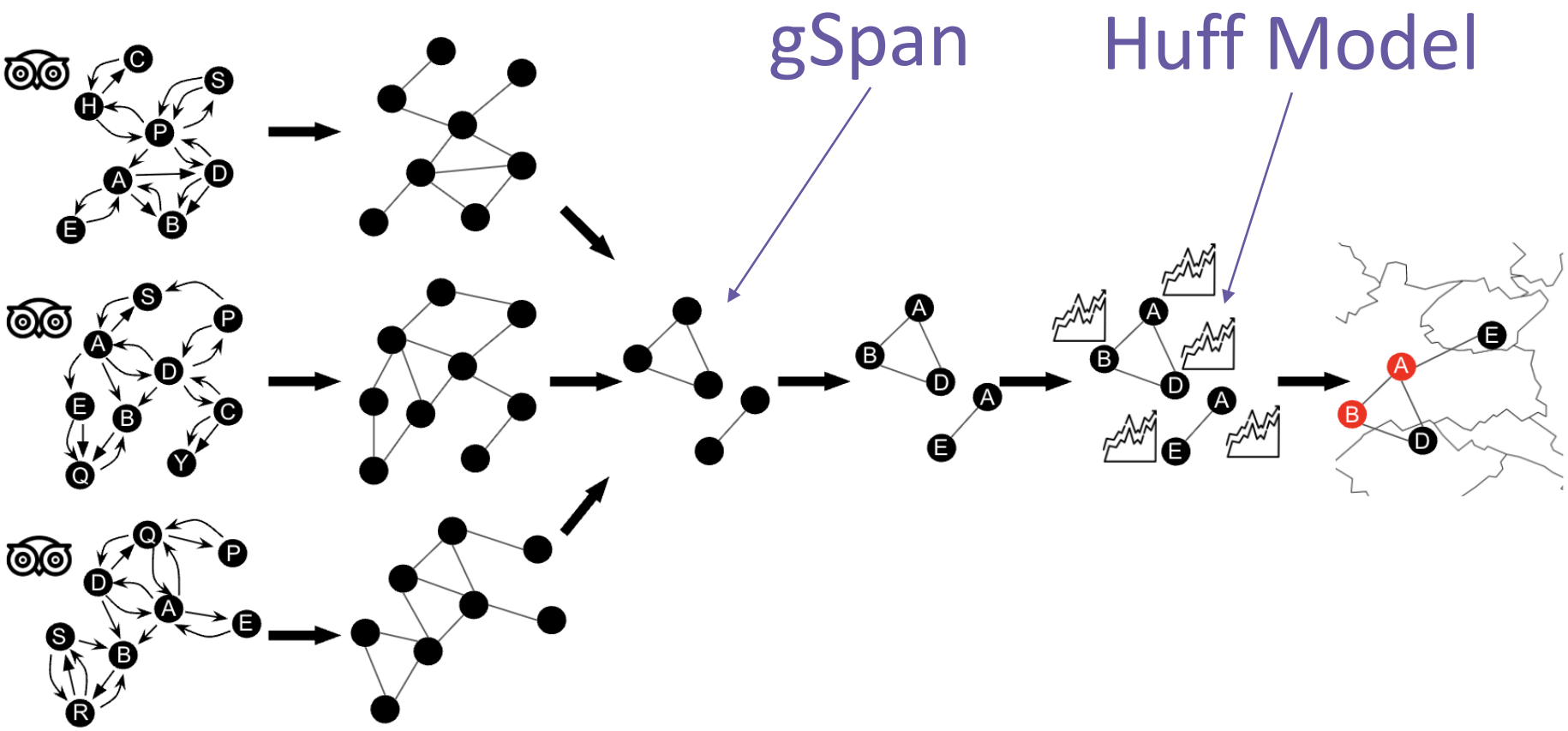
- Requires data flows
- Time and space evolution

➡ **Sequential Pattern Mining** can be a way to extract this behaviour

RELATED WORK

- **Graph-Based Techniques for Analyzing Tourism**
  - Geo-tweets to graphs, provenance, mobile positioning
  - Markov chains, sequence clustering, mobility patterns
  - ⟶ Frequent sub-pattern mining: *gSpan [Yann & Hann, ICDM'22] (DFS-based)*
- **Trajectory Mining**
  - SPADE + density-based clustering => sequential patterns
  - spatial gridding + K-Means clustering => spatiotemporal behavioural patterns
  - ➡️ scalability issues & causality challenges
- **Analysing Overtourism**
  - Qualitative and Survey-Based (social carrying capacity, social exchange theory, overnights, flights, GDP)
  - Textual Data Analysis (extract insights from social media sources and assess the opinions)
  - Quantitative and Machine Learning Approaches (classification, PCA, clustering – city level)
  - Spatiotemporal and Sensor-Based Analysis
    (Taxi & traffic congestion, Mahalanobis distance & TCN, tourism density, GIS – city centers)
  - ➡️ Scalability & dynamicity issues, no border effects

METHODOLOGY



MINING PROCESS

Circulation Multidigraph.

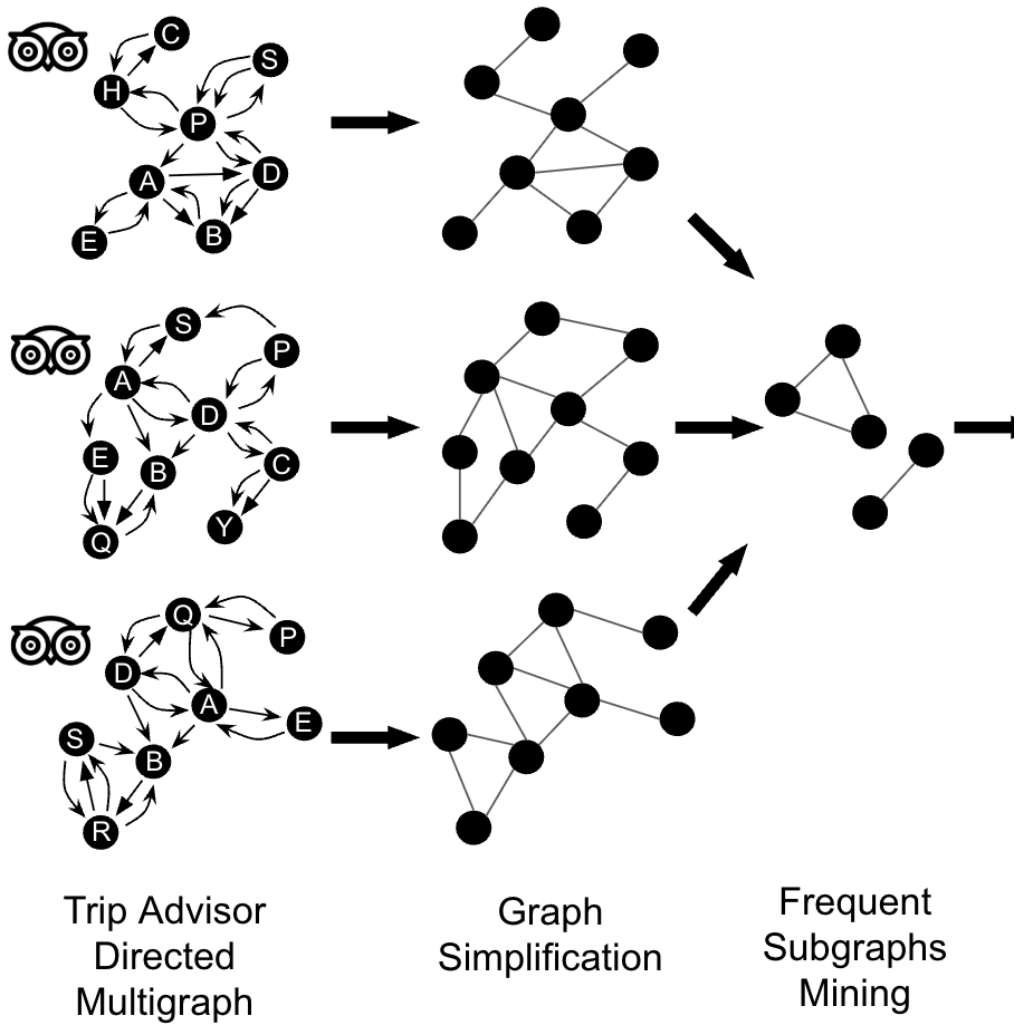
A directed multigraph $G = (V, E, \mu, L_V, L_E)$

- **μ**: ordered pair of vertices $(u, v) \in E$ (allow duplicates)
- **L_V, L_E**: label properties (lat/long, time, reviews, nationality, rating, etc.)

Subgraph Pattern Mining.

Subgraph *g* Isomorphism.

- Applying: *gSpan*
- With support(*g*) ≥ σ



gSPAN

Unlink *Apriori*-based methods, it doesn't generate candidates

DFS code to represent unique graphs

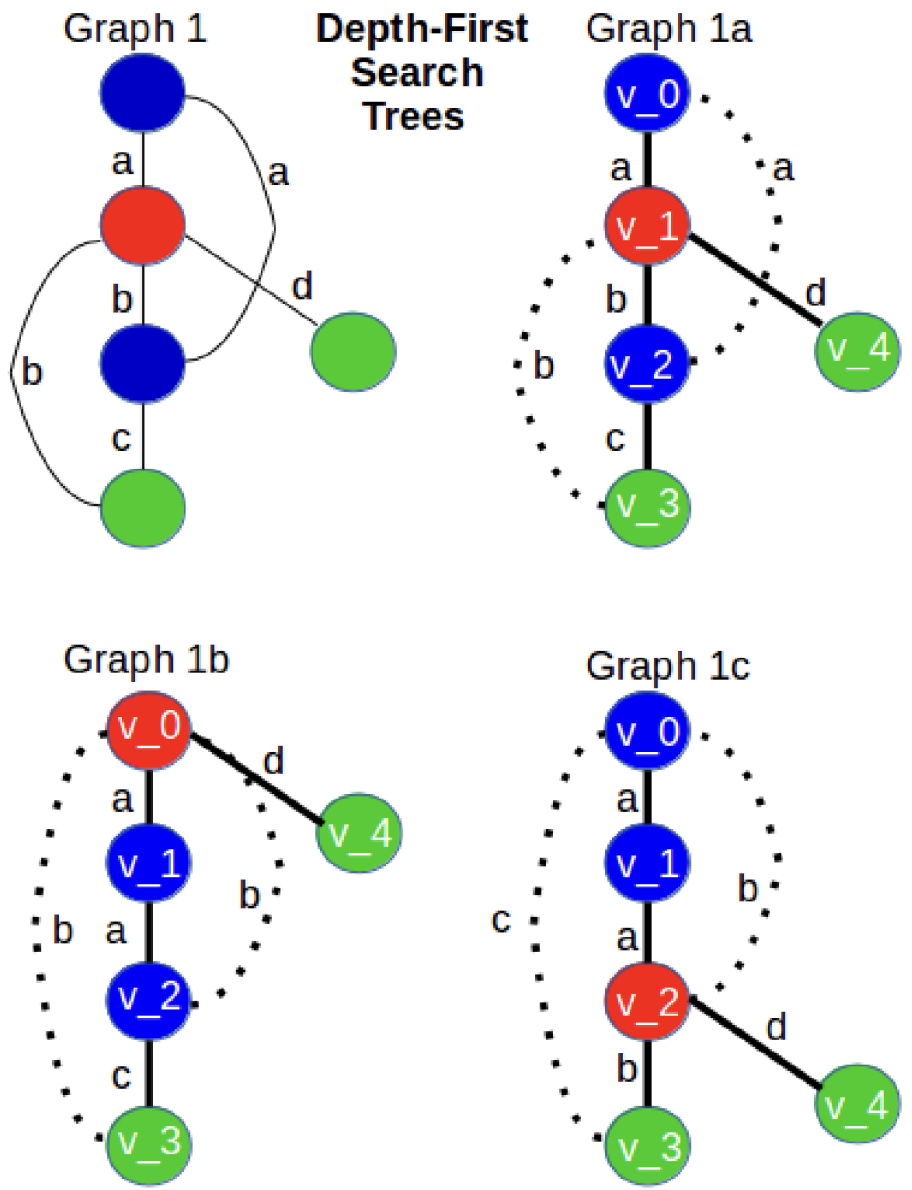
Uses a lexicographic order and minimal DFS code as the canonical form

Use the rightmost extension to efficiently develop patterns

Pruning the research space thanks to the anti-monotony of the support

→ *Reduce the search space to frequent patterns only*

[Yann & Hann, ICDM'22]



HUFF MODEL

$$p_{i,j} = \frac{A_j^{\alpha} / D_{ij}^{\beta}}{\sum_{j \neq i} \frac{A_j^{\alpha}}{D_{ij}^{\beta}}}$$

[D.L. Huff, Journal of Marketing 1964]

Gravity-based spatial interaction model

Store attractiveness vs. travel cost

where:

- $i, j \in \mathcal{V}_g$: vertices are among frequent patterns,
- $p_{i,j}$: probability that individual *i* visit site *j*
- A_j: attractiveness of site *j* (represented by a node's feature: $A_j \in \mathcal{L}_{\mathcal{V}_j}$)
- D_{ij}: distance (Haversine) between site *i* and site *j*
- α, β: sensitivity parameters to A_j and D_{ij}.

➡ *Study locations' attractiveness evolution within frequent patterns*

EXPERIMENTS



Tripadvisor data

- 2012 – 2021 (96 months)
- Cities' area: suburban areas (OECD)
  - Focus on Tampere, Lille, Barcelona

	Locations /nodes	Reviews /comments	Multi-digraph edges	Graph simplification (edges)	# frequent patterns (gSpan)
Tampere	688	21k	7.3k	51	4
Lille	2,531	209k	77k	514	74
Barcelona	18,514	1.05M	696k	11,767	1,010

PATTERNS ANALYSIS

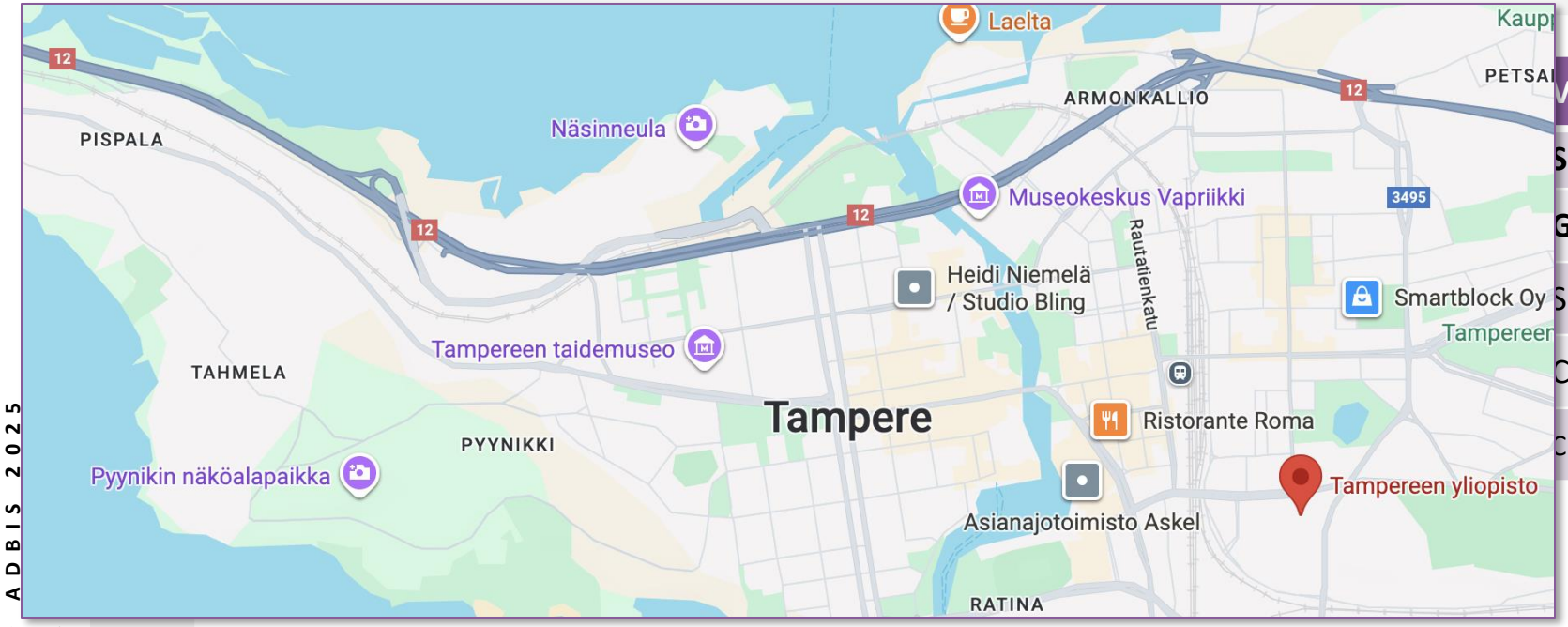
Pattern ID	Vertices	Support (#months)
0	[269814, 269817]	85
1	[269817, 269820]	72
2	[269814, 269817, 269820]	71
3	[269814, 269820]	70
4	[269814, 269820, 269817]	66

Lille's Most Frequent Subgraphs

Pattern ID	Vertices	Support (#months)
0	[190166, 190624]	95
1	[190162, 190166]	92
2	[190624, 190629]	91
3	[190166, 190624, 190629]	91
4	[190166, 1059712]	91

Barcelona's Most Frequent Subgraphs

LIKE,SHARE, FOLLOW ! @poledevinci | DEVINCI.FR



Vertices	Fq
Sagrada Famìlia	926
Guëll Parc	866
Sainte-Marie-de-la-Mer Church	483
Casa Mìla	431
Camp Nou	397

Barcelona

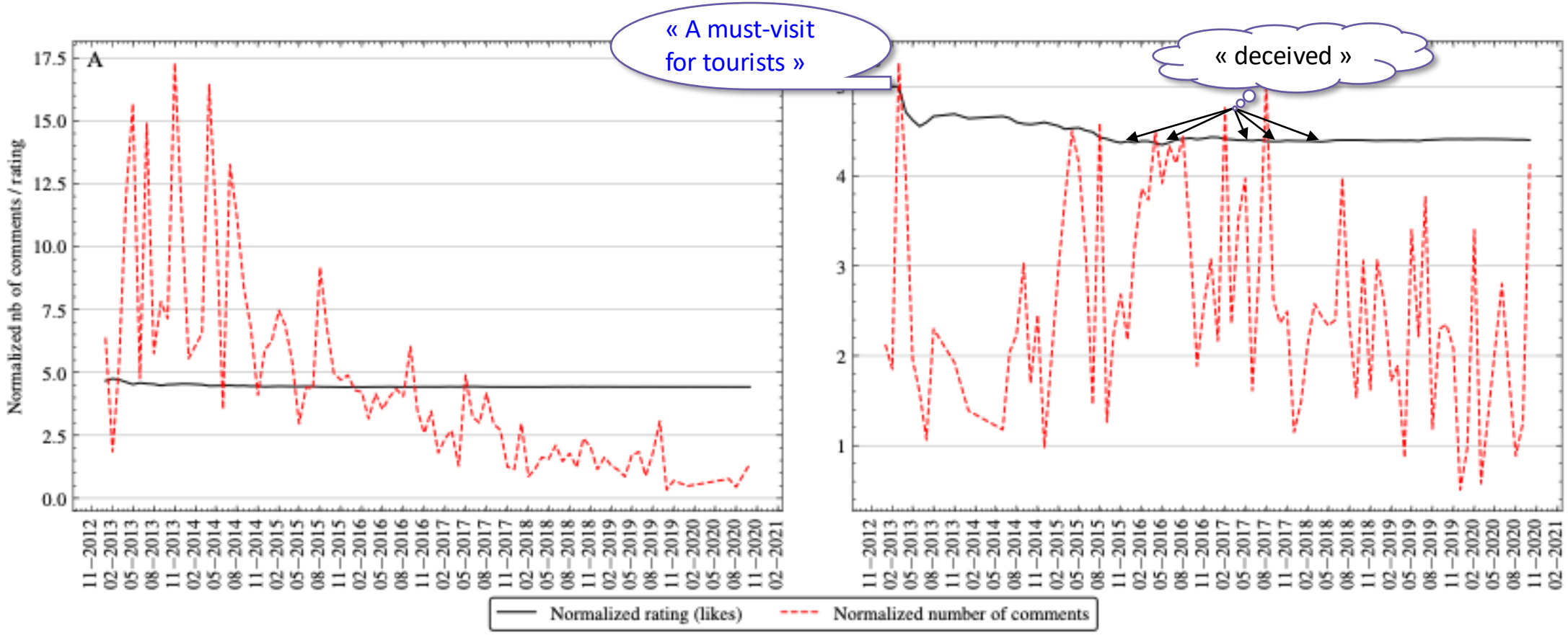
ID	Vertices	Fq
269820	**Palais des Beaux-Arts de Lille**	45
269814	Grand Place	44
269817	Vieille Bourse	40
793722	La Chicoree	25
247498	Lille Métropole Musée d'Art Moderne	13

Lille

ID	Vertices	Fq
661108	**Park and Pyynikki observation tower**	2
777510	Pizzeria Napoli	2
604970	Tampere Cathedral	1
941703	**Pyynikin Nakotorninkahvila**	1

Tampere

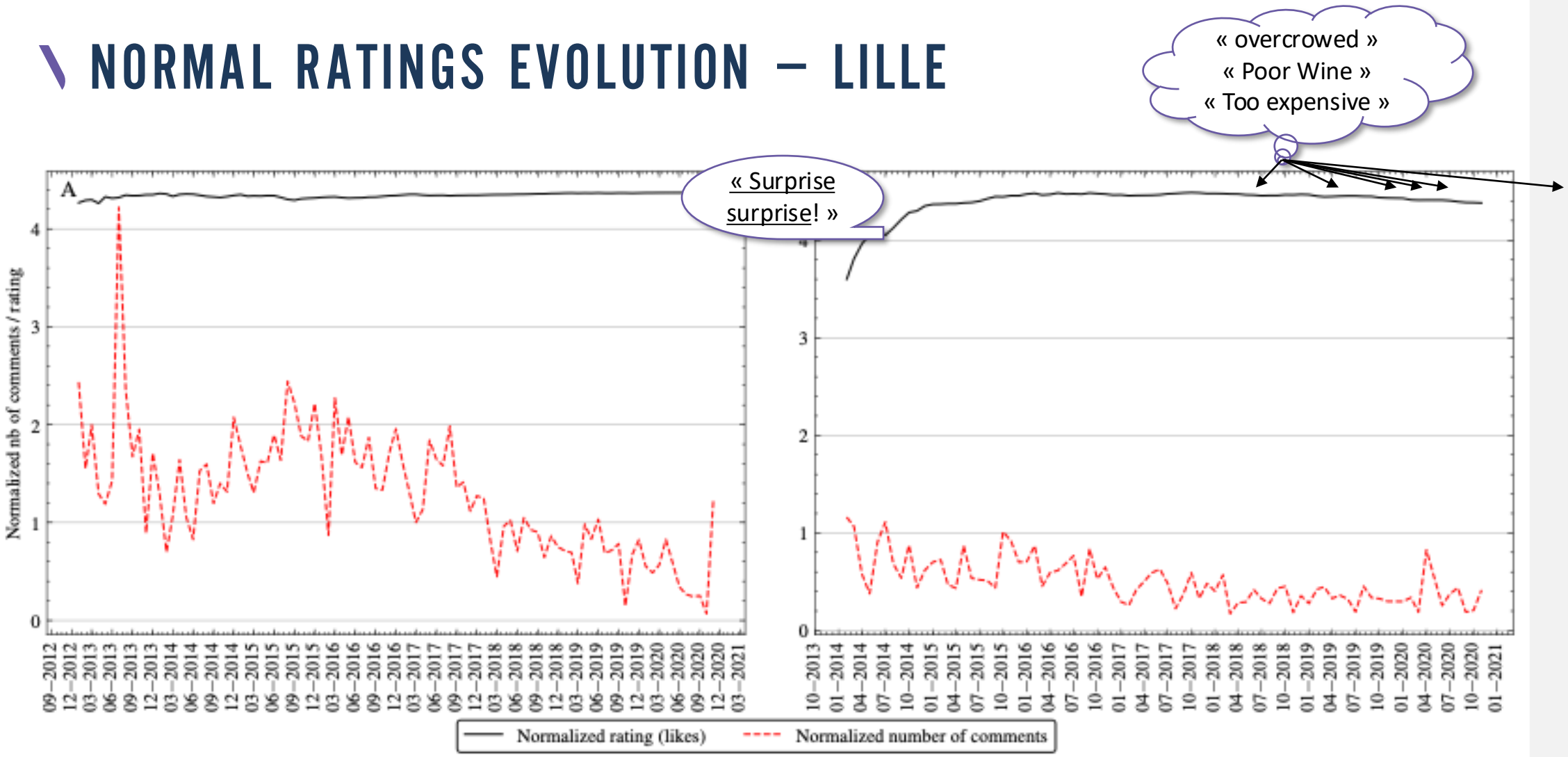
NORMAL RATINGS EVOLUTION – TAMPERE



Park and Pyynikki observation tower

Pyynikin Nakotorninkahvila

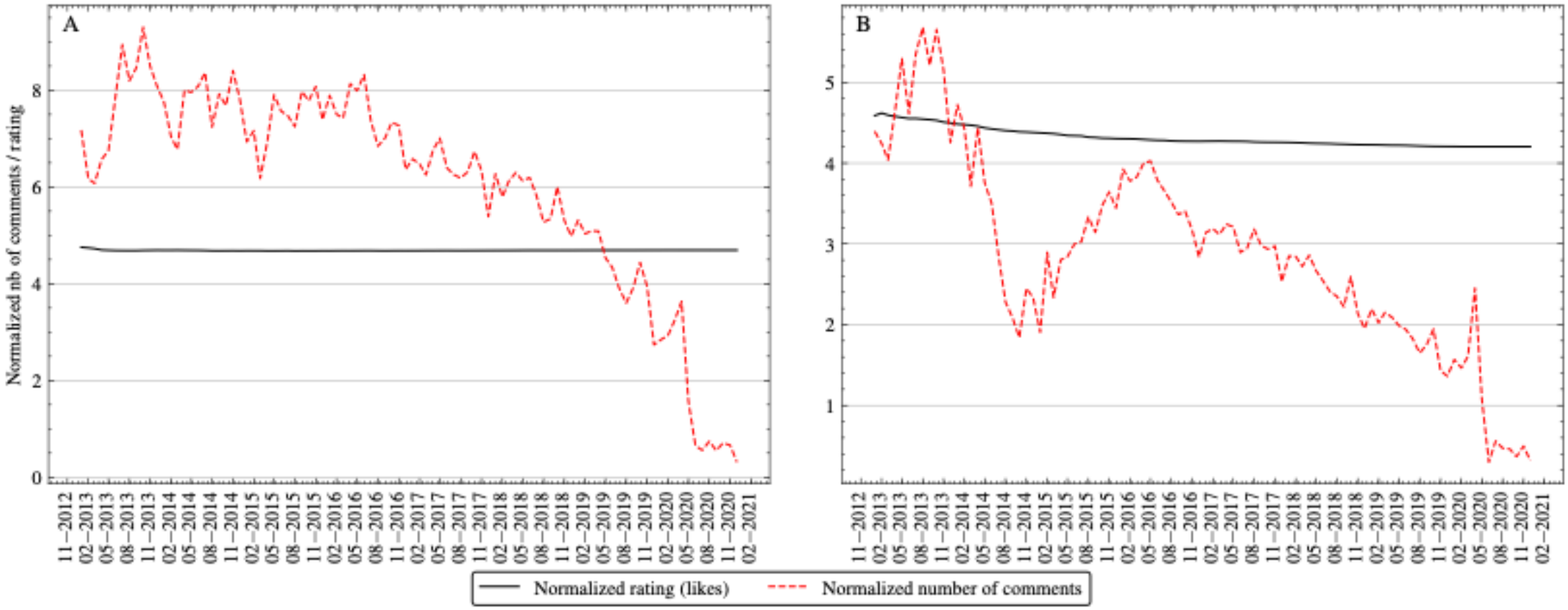
NORMAL RATINGS EVOLUTION – LILLE



Palais des Beaux-Arts

Bloempot restaurant

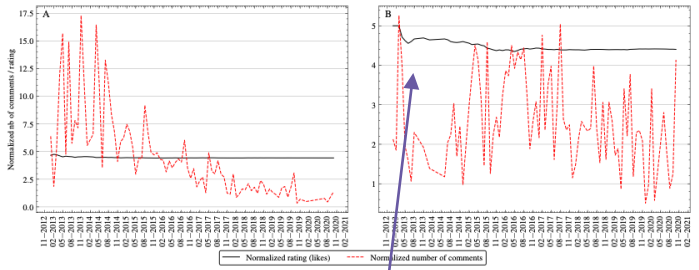
NORMAL RATINGS EVOLUTION – BARCELONA

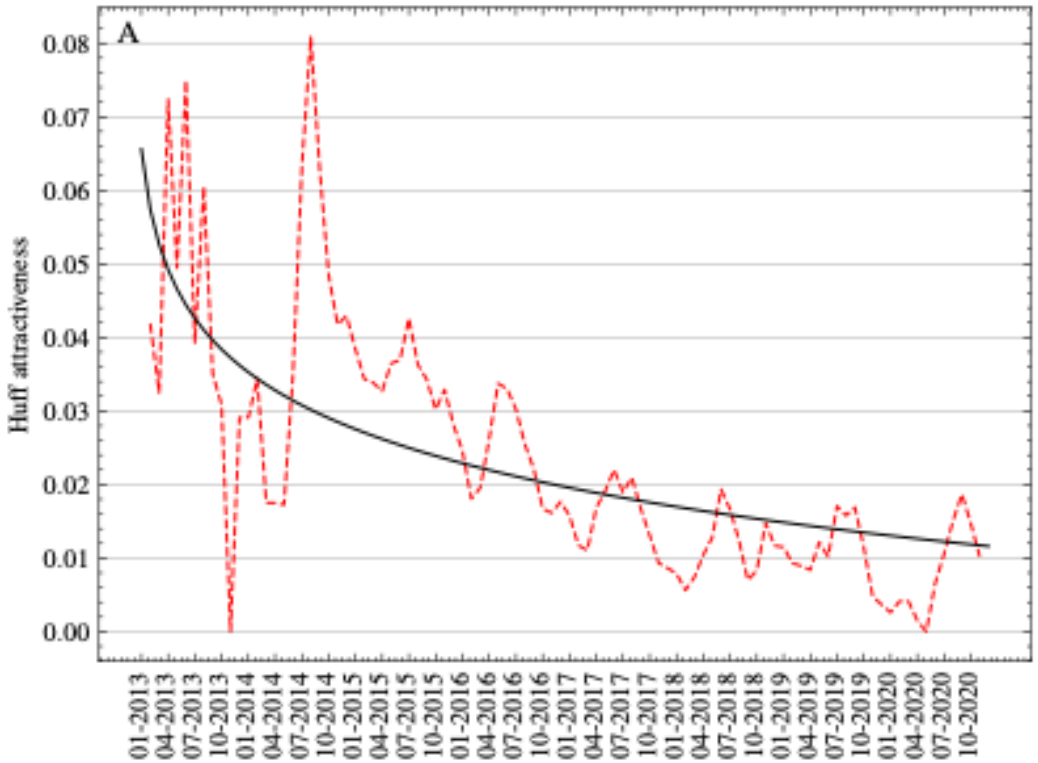


Sagrada Famìlia

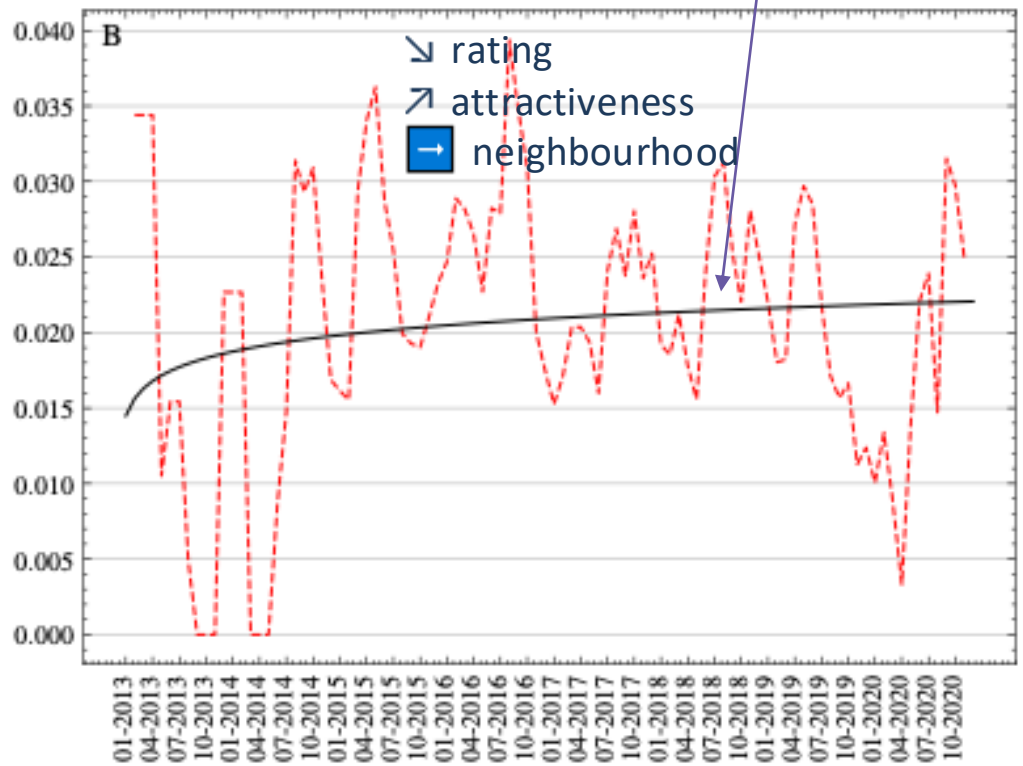
Guëll Parc

HUFF MODEL EVOLUTION – TAMPERE



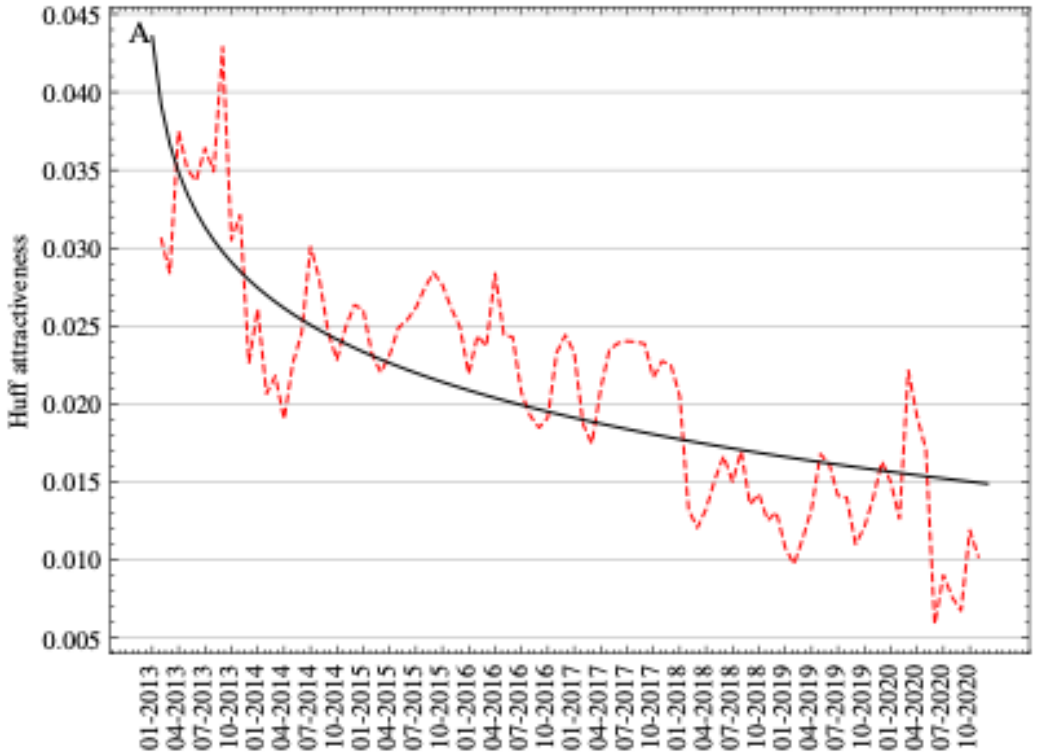


Park and Pyynikki observation tower

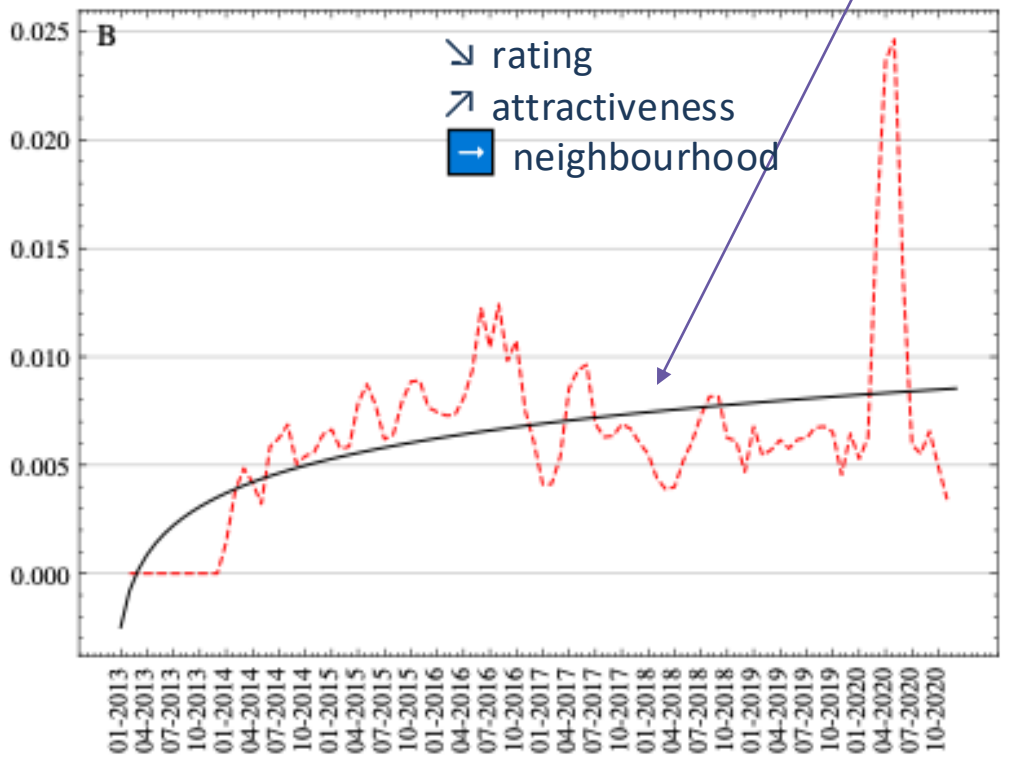


Pyynikin Nakotornin Kahvila

HUFF MODEL EVOLUTION – LILLE

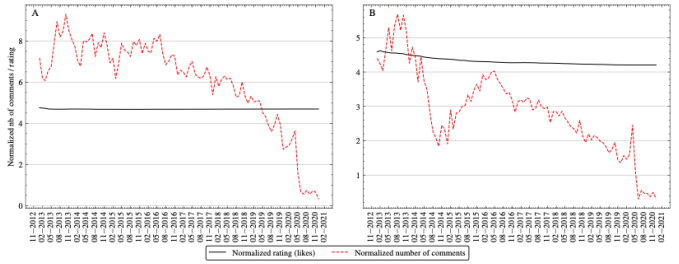


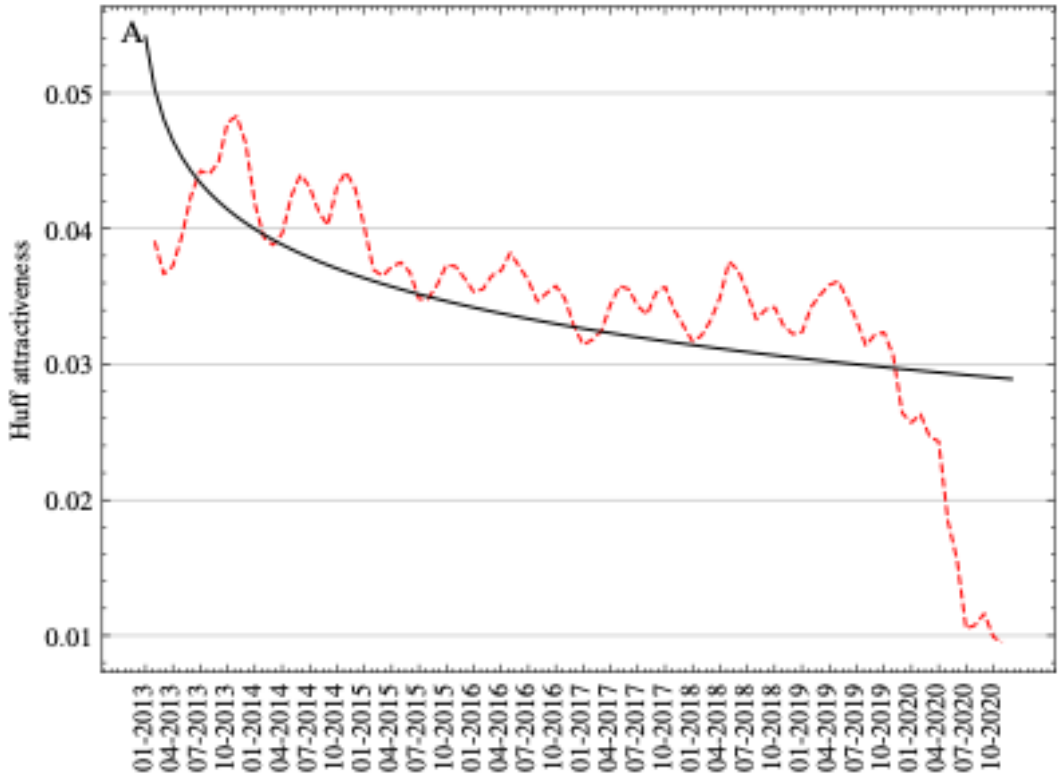
Palais des Beaux-Arts



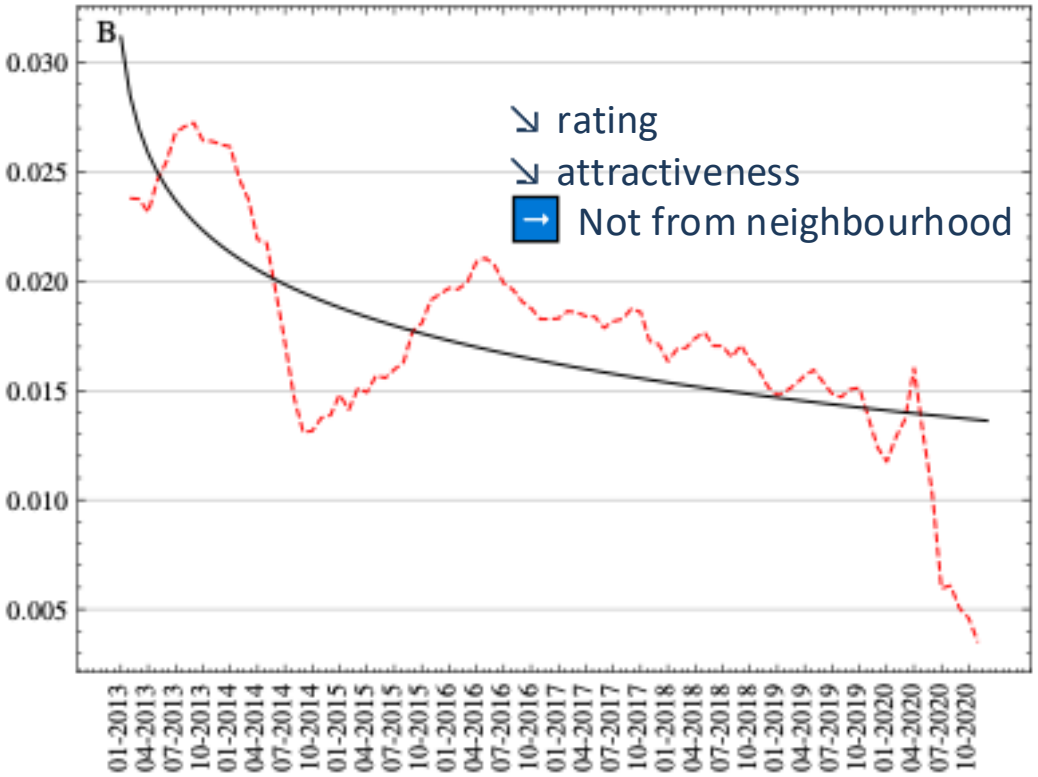
Bloempot restaurant

HUFF MODEL EVOLUTION – BARCELONA





Sagrada Famìlia



Guëll Parc

CONCLUSION & PERSPECTIVES

A data-driven approach to the early detection of latent overtourism:

- Temporal graph modeling,
- Graph simplification + frequent subgraph mining,
- Spatial attractiveness metrics

Perspectives:

- Contextual variables (short-term fluctuation vs long-term trends)
- Integrate text sentiment analysis in Huff model
- Overtourism patterns characterization
- Overtourism graph embeddings
- Overtourism area clustering

➡ ANR GRACEO (resubmission)

