ANOVA Simultaneous Component Analysis for the Efficient Exploration of Massive Network Traffic Data

J. Camacho

Computational Data Science Laboratory (CoDaS Lab)
Research Centre for Information and Communication Technologies (CITIC-UGR)
University of Granada, Spain

josecamacho@ugr.es

Topic: Providing network observability is not only a matter of devising the best data measurement techniques (e.g., the network telemetry framework in RFC9232), but also of properly engineering good practices for data visualization, exploration, and understanding

Contribution: In this poster, we extend ANOVA Simultaneous Component Analysis (ASCA) [1] for the visualization of network Big Data

Results: We provide insights into the massive and complex Netmob 2023 Data Challenge

The **Netmob Data** [2,3] includes 77 days of traffic generated by 68 popular mobile services, in upload and download direction, geo-referenced over 20 metropolitan areas in France in 2019

The data includes more than 870,000 high-resolution regular tiles of 100×100 m2 each and a temporal resolution of 15 minutes

Data is massive and difficult to visualize and understand in terms of spatio-temporal patterns

ASCA is a combination of Analysis of Variance (ANOVA) and Principal Component Analysis (PCA)

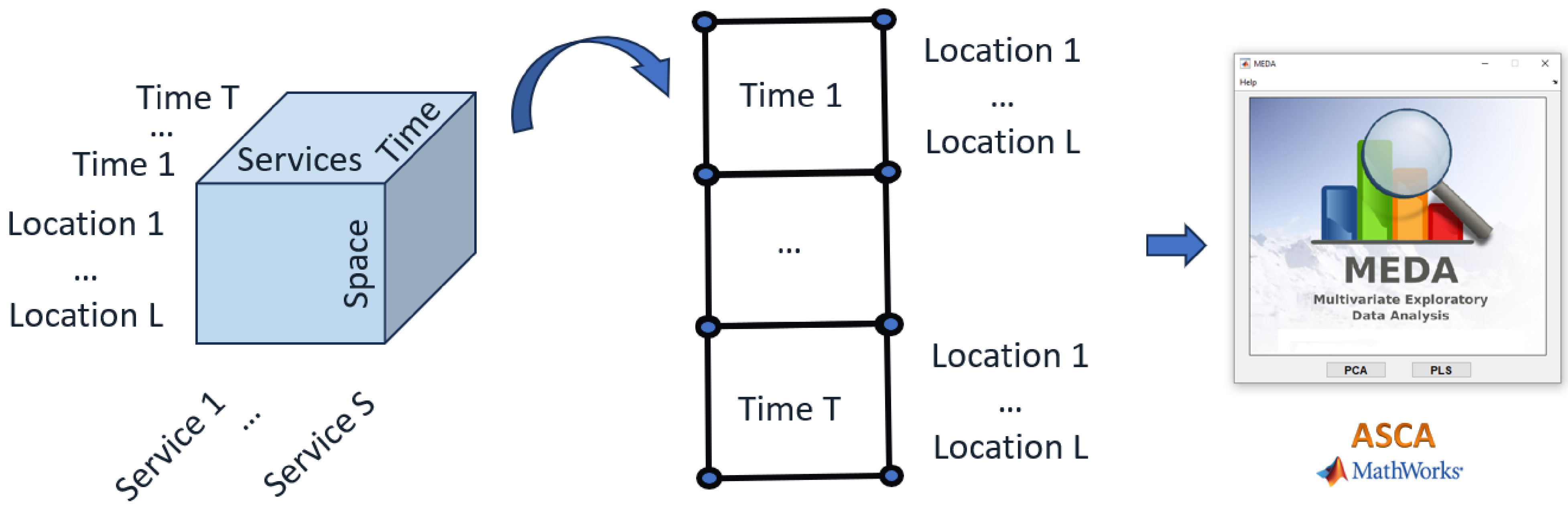
i) **Factorizes** the data according to a set of factors, like time or location, allowing us to untangle temporal and spatial patterns

ii) Performs **statistical inference**

iii) Allows visualizations of the patterns with **PCA** plots

Data Analysis Workflow

1. We organize the traffic statistics of the Netmob Data in 3-way tensors of **Space** x **Time** x **Service**
2. We unfold the spatio-temporal tensor in a matrix along the service mode
3. We use our own implementation of ASCA in the MEDA Toolbox [4], a software package in Matlab available to the community



(1) (2) (3)

High-level spatio-temporal ASCA model

We aggregate traffic into a high-level spatio-temporal tensor of

77 days x 20 cities x 136 services

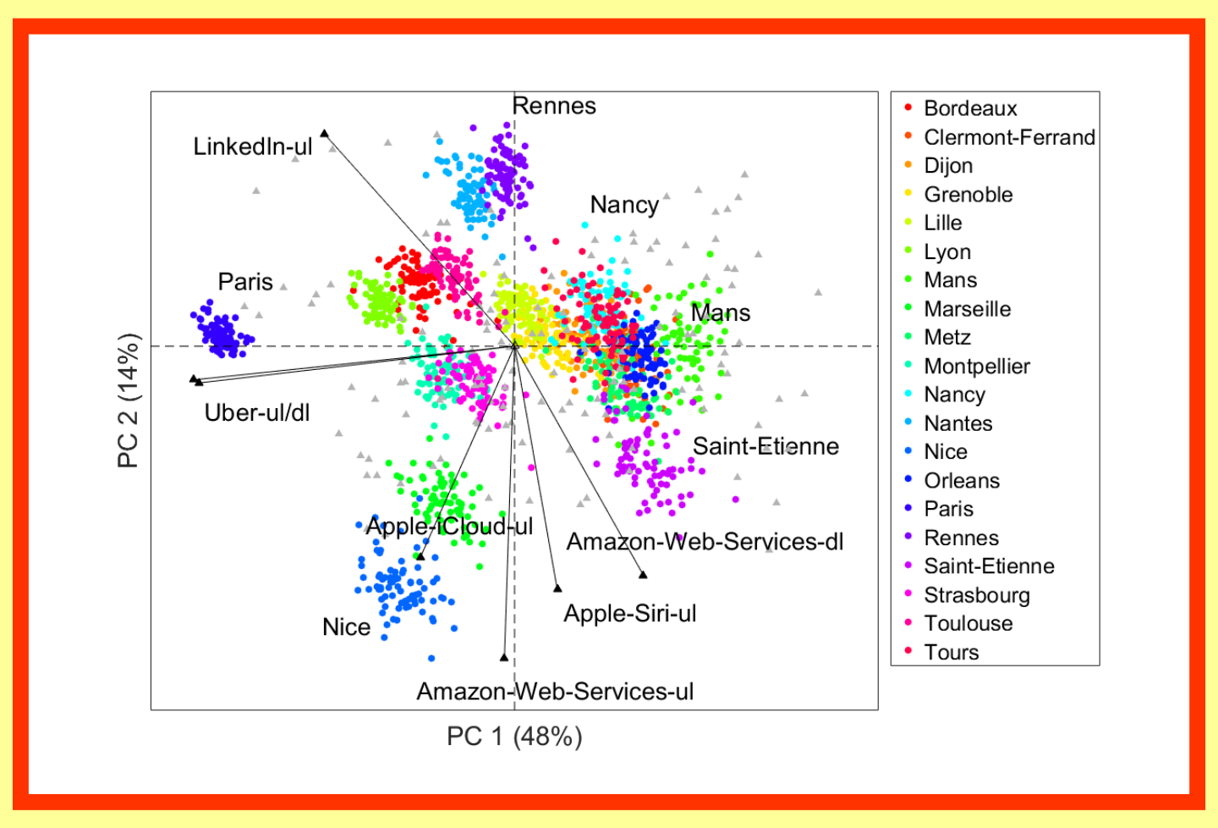
We apply ASCA following the workflow and with the model below

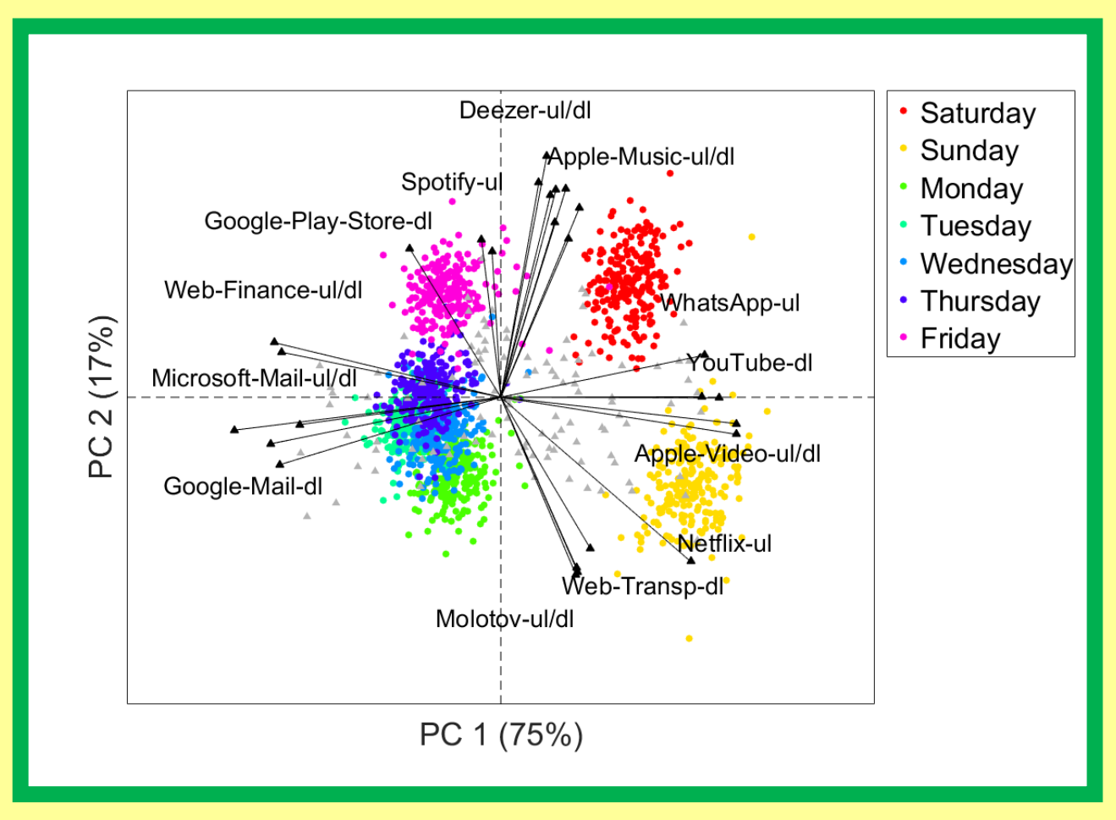
$$X = X_{City} + X_{Weekday} + X_{Date} + E$$

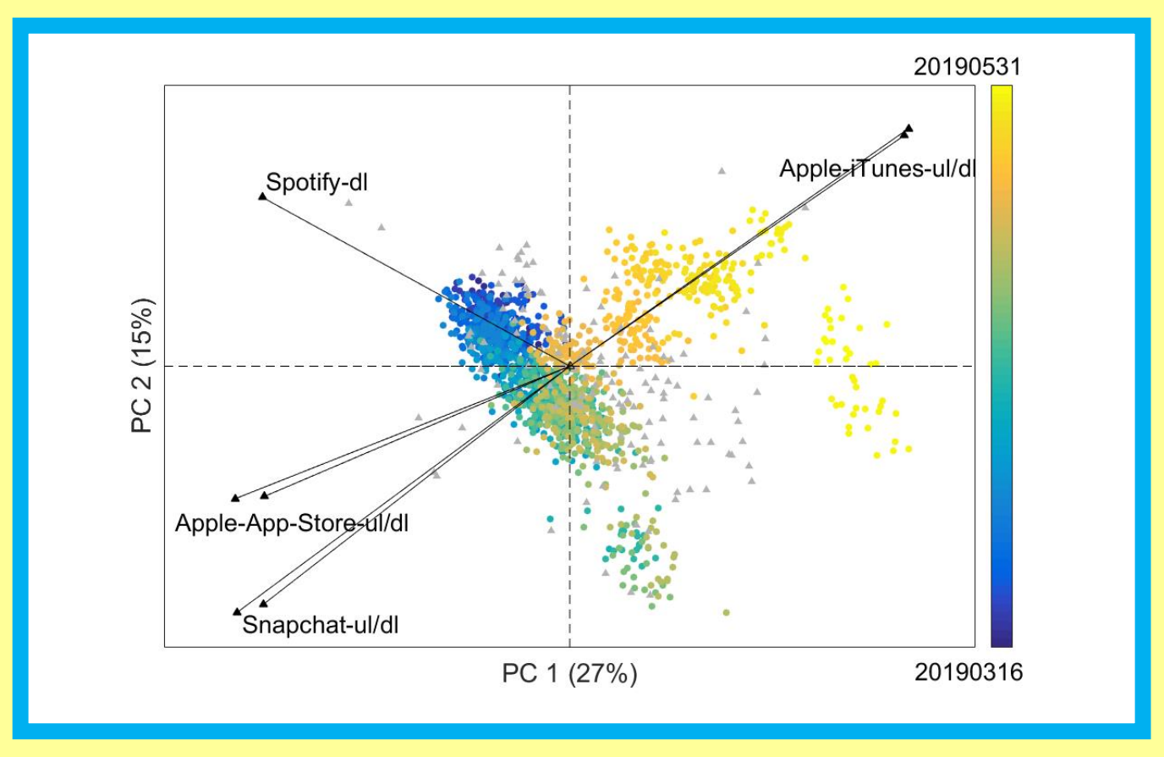
All factors are statistically significant

Weekday is the most important factor affecting the traffic, followed by City (3/5) and Date (1/8)

Spatio-temporal patters for specific services are discovered in the plots (check the paper for details)







City-level Big Data ASCA model

We build a different tensor per city, e.g., for Paris

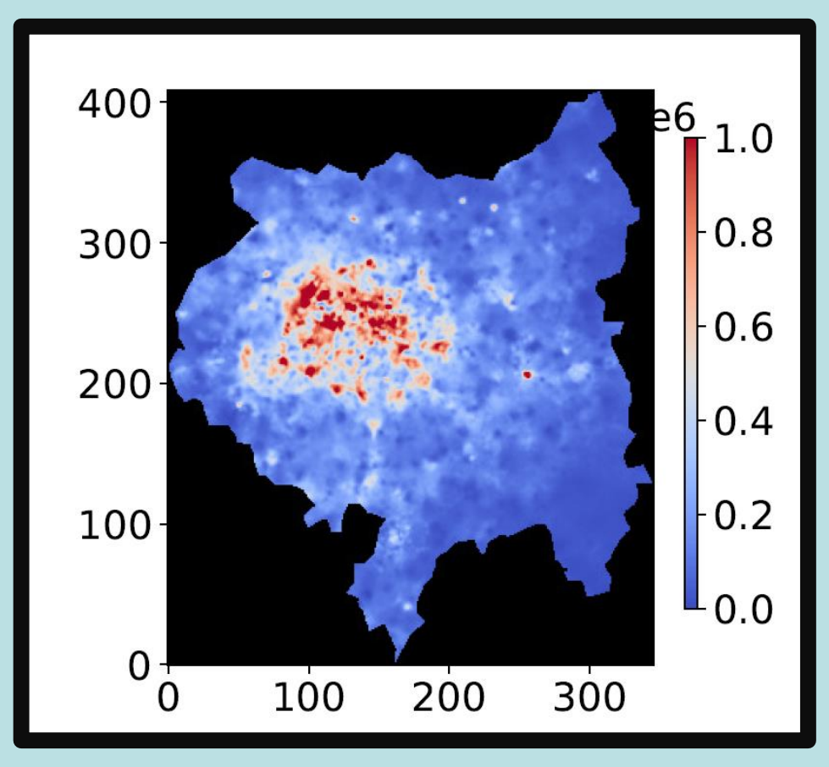
77 days x 6M tiles x 136 services

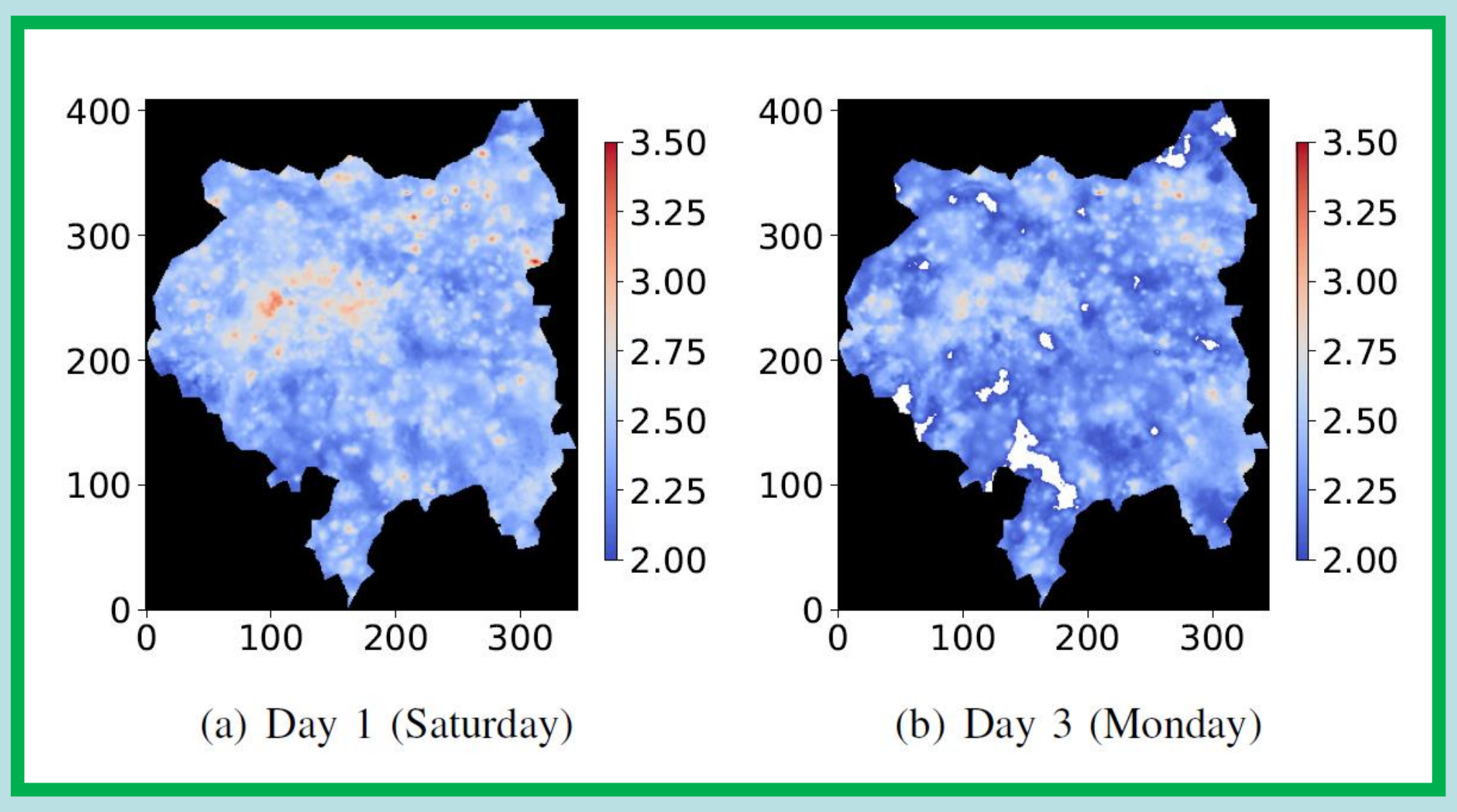
For the analysis, we contribute a **new extension of ASCA to Big Data** using Cross-product matrices and Clustering, following previous work [5]

$$X = X_{Row} + X_{Column} + X_{Weekday} + X_{Date} + E$$

Weekday is the most important factor in all cities. Spatio-temporal patters differ across cities.

In Paris, we find selected services with relevant spatial (e.g., **Uber upload**, left) and/or temporal (e.g., Instagram download, right) patterns





(a) Day 1 (Saturday) (b) Day 3 (Monday)

The capability of ASCA to untangle the effect of different factors in the data was leveraged to investigate spatio-temporal patterns of traffic for the first time. We demonstrate this capability with the Netmob 2023 data, for which we had to develop the first Big Data extension of ASCA

[1] Smilde et al., “Anova-simultaneous component analysis (asca): a new tool for analyzing designed metabolomics data,” Bioinformatics, vol. 21, no. 13, pp. 3043–3048, 2005.
[2] Netmob 2023 Data Challenge. https://netmob2023challenge.networks.imdea.org (Last Access 24th April, 2024)
[3] Martínez-Durive et al. 2023. The netmob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography. arXiv preprint arXiv:2305.06933..
[4] “GitHub repository for the MEDA Toolbox,” https://github.com/codaslab/MEDA-Toolbox (Last Access 24th April, 2024)
[5] Camacho. “Visualizing Big data with Compressed Score Plots: Approach and Research Challenges.” Chemometrics and Intelligent Laboratory Systems, vol. 135, pp. 110–125, 2014.

ACKNOWLEDGEMENTS: This work was supported by the Agencia Estatal de Investigación in Spain, MCIN/AEI/ 10.13039/501100011033, grant No PID2020-113462RB-I00.