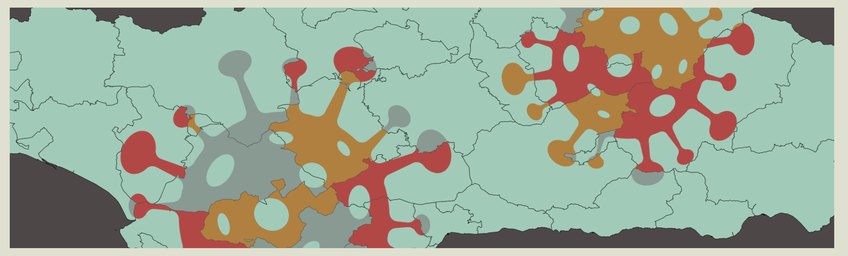
UNIVERSIDAD DE CÓRDOBA

DEPARTMENT OF COMPUTER SCIENCE AND NUMERICAL ANALYSIS

UNIVERSITY OF CÓRDOBA, CÓRDOBA, SPAIN

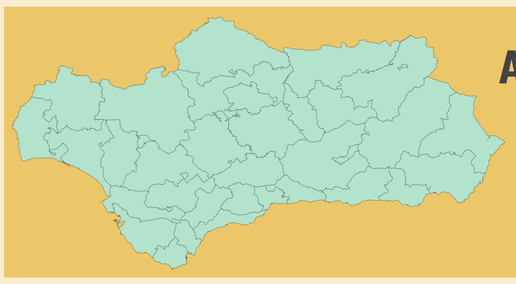
CLUSTERING OF COVID-19 TIME SERIES INCIDENCE INTENSITY IN ANDALUSIA, SPAIN

Miguel Díaz-Lozano, David Guijo-Rubio, Pedro Antonio Gutiérrez, César Hervás-Martínez



01 INTRODUCTION

In this paper, an approach based on a time series clustering technique is presented by extracting relevant features from the original time series. A curve characterization is applied to the daily contagion rates of the **34 sanitary districts** of Andalusia, Spain. Sanitary districts are administrative divisions that have local health management competencies in specific zones of Andalusia. By determining the maximum incidence instant and two inflection points for each wave, an outbreak curve can be described by six intensity features, defining its initial and final phases. These features are used to derive different groups using state-of-the-art clustering techniques with the objective of identifying Andalusian sanitary districts that behave similarly in terms of intensity in different wave periods.



ANDALUSIAN SANITARY DISTRICTS

02 MATERIAL AND METHODS

DATA ACQUISITION

The information has been obtained from the official **Andalusian government**, where COVID-19 statistics are daily reported. The information about positive diagnoses, cured and deceased people is available since February 26, 2020.

OUTBREAKS LIMITATIONS

ERRATIC

- 1ST March 12, 2020 to April 15, 2020
- 2ND July 31, 2020 to December 15, 2020
- 3RD December 15, 2020 to March 13, 2021

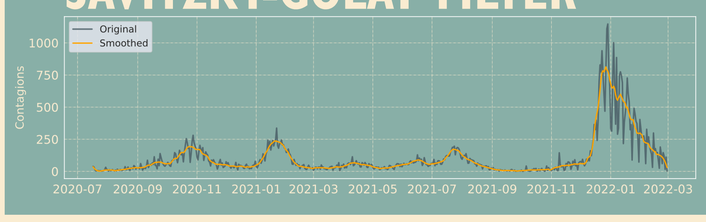
VACCINATION

- 4TH March 13, 2021 to June 15, 2021
- 5TH June 15, 2021 to September 15, 2021

PREPROCESSING

The time series present a weekly pattern in which fewer diagnoses are recorded during weekends. ▶ **HIGHLY NOISY**

SAVITZKY-GOLAY FILTER

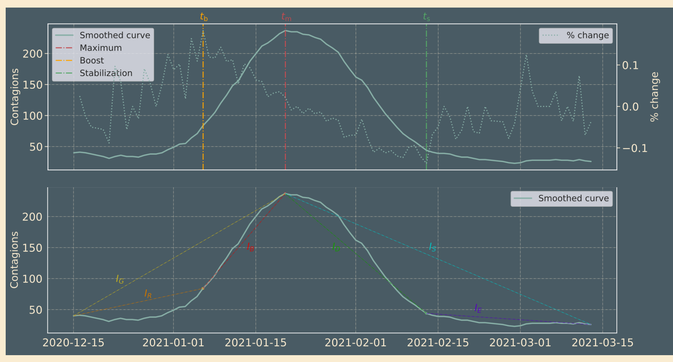


District of Córdoba

Original and smoothed data from July, 15 200 to September 15, 2021.

CURVE CHARACTERIZATION

From **TIME SERIES DATA** ▶ To **CHARACTERIZATION FEATURES**



3 INSTANTS
FOR EACH OUTBREAK AND DISTRICT

- **Boost (t_b):** rate accelerates.
- **Maximum (t_m):** highest incidence.
- **Stabilization (t_s):** rate stabilizes.

6 INTENSITIES FEATURES
SLOPE BETWEEN INSTANTS

- **Grow (I_G):** from begin to t_m
- **Rise (I_R):** from begin to t_b.
- **Boost (I_B):** from t_b to t_m.
- **Decrease (I_D):** from t_m to t_s.
- **End (I_E):** from t_s to end.
- **Stabilize (I_S):** from t_m to end.

CLUSTERING

ALGORITHMS

- **HIERARCHICAL**: Agglomerative
- **PARTITIONAL**: *k*-Means, *k*-Medoids

INTERNAL VALIDATION METRICS

No external information is available for validating groups.

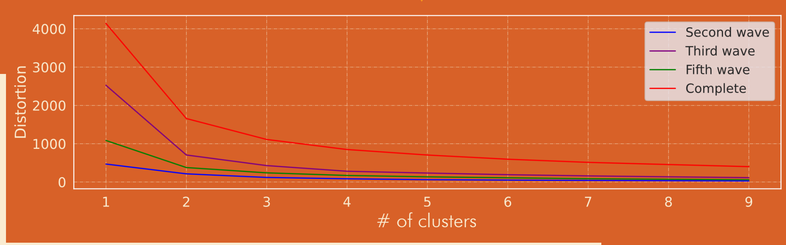
- Silouhette Index (SI) — MAXIMIZE
- Calinski-Harabasz (CH) — MAXIMIZE
- Dunn Index (DI) — MAXIMIZE
- Davies-Bouldin (DB) — MINIMIZE

03 RESULTS

 # OF CLUSTERS

ELBOW METHOD
FOR DETERMINING THE BEST NUMBER OF CLUSTERS

DISTORSION
Sum of squared distances from the patterns to the closest centroid.



CONSIDERING MORE THAN 3 CLUSTERS DOES NOT REDUCE SIGNIFICANTLY THE DISTORSION.

METRICS

***k*-means**
performs better overall in most internal validation metrics.

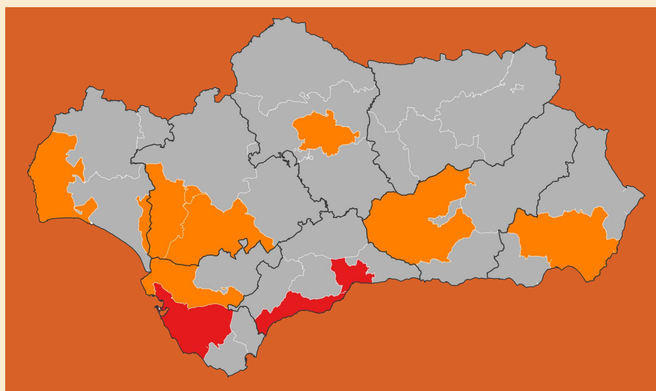
Wave	Agglomerative				*k*-means				*k*-medoids			
	SI	CH	DB	DI	SI	CH	DB	DI	SI	CH	DB	DI
2	0.68	39.11	0.59	0.34	0.62	45.46	0.75	0.16	0.46	28.62	0.92	0.09
3	0.64	73.59	0.63	0.13	0.66	75.81	0.64	0.13	0.67	75.12	0.64	0.12
5	0.58	52.95	0.80	0.15	0.58	54.77	0.81	0.12	0.40	47.34	0.82	0.12
Complete	0.66	34.05	0.83	0.33	0.68	42.26	0.80	0.44	0.41	31.59	0.98	0.14

Quality metrics for the three clustering techniques using $k = 3$.

GROUPS DEFINED BY INTENSITY FEATURE MAGNITUDES

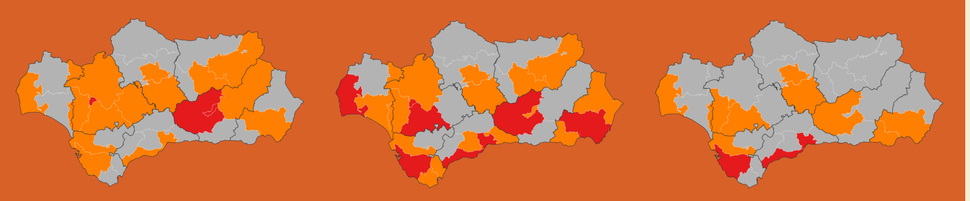
MAJOR INCIDENCE **HIGH INCIDENCE** **MODERATE INCIDENCE**

GEOGRAPHICAL RESULTS



OVERVIEW CLUSTER USING ALL OUTBREAKS

EVOLUTION OF CLUSTERS IN DISTRICTS OVER TIME



2ND OUTBREAK **3RD OUTBREAK** **5TH OUTBREAK**

04 DISCUSSION

- **Intensity features are clearly segregated into 3 classes.**
- **In the overview cluster, the virus spread was major in the districts of Málaga, Costa del Sol and Granada.** ▶ TOURISM AND MOBILITY
- **In the overview cluster, the virus spread was high in province capitals and districts close to the capitals.** ▶ POPULATION DENSITY
- **During the third outbreak, the number of districts with major incidence increased with respect the second outbreak.** ▶ CHRISTMAS HOLIDAYS
- **Málaga, Cádiz Bay and Costa del Sol show a extreme major incidence during the fifth wave, which took place in summer.** ▶ CROWD OF PEOPLE

05 CONCLUSION

The COVID-19 daily contagions curve **characterization** proposed in this article results in a descriptive dataset that is used to analytically describe the pandemic situation by means of the contagion rate intensities.

The resulting clusters may be used as **auxiliary information** to adopt similar prevention measures on different locations exhibiting similar behaviors. Moreover, with the aim of modeling the contagion rate to be used for forecasting purposes, this cluster analysis allows the possibility of **reducing the number of models required to forecast the transmission rate** in all the districts by using joint information from areas with similar behaviors.

ACKNOWLEDGEMENTS

This work was supported by the "Agencia Española de Investigación (España)" (grant reference: PID2020-115454GB-C22 / AEI / 10.13039 / 501100011033); the "Consejería de Salud y Familia (Junta de Andalucía)" (grant reference: PS-2020-780); and the "Consejería de Transformación Económica, Industria, Conocimiento y Universidades (Junta de Andalucía) y Programa Operativo FEDER 2014-2020" (grant references: UCO-1261651 and PY20_00074).

David Guijo-Rubio's research has been subsidised by the University of Córdoba through grants to Public Universities for the requalification of the Spanish university system of the Ministry of Universities, financed by the European Union - NextGenerationEU (grant reference: UCOR01MS).

BIO-SCIENCES INTERPLAY COMPUTATION

IWINAC2022
9th International Work-Conference on the Interplay between Natural and Artificial Computation

REFERENCES

Fraley, C., Raftery, A.E.: **How many clusters? which clustering method? Answers via model-based cluster analysis.** The Computer Journal 41(8), 578–588 (1998).

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M., Perona, I.: **An extensive comparative study of cluster validity indices.** Pattern recognition 46(1), 243–256 (2013).

Khan, M., Mehran, M.T., Haq, Z.U., Ullah, Z., Naqvi, S.R., Ihsan, M., Abbass, H.: **Applications of artificial intelligence in covid-19 pandemic: A comprehensive review.** Expert Systems with Applications 185, 115695 (2021).

Grupo de investigación AYRNA

Junta de Andalucía
Consejería de Salud y Familias
FUNDACIÓN PROGRESO Y SALUD

ESCUELA POLITÉCNICA SUPERIOR DE CÓRDOBA
Universidad de Córdoba