

XXI Jornadas de Estadística de las Comunidades Autónomas

Una evaluación de modelos de pronóstico para COVID-19 en Cantabria

Alberto Lezcano Lastra
Gonzalo Llamosas García
Alejandro López Cagigas
Francisco Javier Parra Rodríguez

(Instituto Cántabro de Estadística)

Índice

Introducción

- Motivación
- Evidencia empírica
- Principales resultados de la bibliografía

Metodología

- Datos
- Modelos de series temporales

Resultados

- Comparativa de modelos
- Aplicativo

Conclusiones



Introducción

❑ Motivación:

I. Contexto global

- COVID-19: pandemia global
- Elevados ratios de contagio y transmisión
- Incertidumbre sobre la evolución de la crisis
- Ausencia de información estadística sobre la pandemia
- Necesidad de anticipar los brotes epidémicos



Introducción

❏ Motivación:

II. Contexto regional

- La Consejería de Sanidad del Gobierno de Cantabria propone al ICANE trabajar sobre las siguientes tareas:
 - ✓ Pronósticos diarios de casos PCR positivos
 - ✓ Estimación del número de reproducción instantáneo (R_0)
 - ✓ Desarrollo de un sistema de geolocalización de casos COVID-19



Introducción

Evidencia empírica



ARTICLE

<https://doi.org/10.1038/s41467-021-22457-w> OPEN

Predictive performance of international COVID-19 mortality forecasting models

Joseph Friedman^{1,4}, Patrick Liu^{2,4}, Christopher E. Troeger³, Austin Carter³, Robert C. Rein Ryan M. Barber³, James Collins³, Stephen S. Lim³, David M. Pigott³, Theo Vos³, Simor Christopher J. L. Murray^{3,5} & Emmanuela Gakidou^{3,5}

Forecasts and alternative scenarios of COVID-19 mortality have been critical inputs for pandemic response efforts, and decision-makers need information about predictive performance. We screen $n = 386$ public COVID-19 forecasting models, identifying $n = 7$ that are global in scope and provide public, date-versioned forecasts. We examine their predictive performance for mortality by weeks of extrapolation, world region, and estimation month. We additionally assess prediction of the timing of peak daily mortality. Globally, models released in October show a median absolute percent error (MAPE) of 7 to 13% at six weeks, reflecting surprisingly good performance despite the complexities of modelling human behavioural responses and government interventions. Median absolute error for peak timing increased from 8 days at one week of forecasting to 29 days at eight weeks and is similar for first and subsequent peaks. The framework and public codebase (<https://github.com/pylu47/covidcompare>) can be used to compare predictions and evaluate predictive performance going forward.



Analysis and forecast of COVID-19 spreading in China, Italy and France

Duccio Fanelli^{a,*}, Francesco Piazza^{b,c}

^aDipartimento di Fisica e Astronomia, Università di Firenze, INFN and CSDC, Via Sansone 1, Sesto Fiorentino 50019, Firenze, Italy
^bCentre de Biophysique Moléculaire (CBM), CNRS-UPR 4301, Rue C. Sadron, Orléans 45071, France
^cUniversité d'Orléans, Chateau de la Source, Orléans Cedex 45071, France

ARTICLE INFO

Article history:
Received 11 March 2020
Accepted 12 March 2020
Available online 21 March 2020

Keywords:
Covid-19
epidemic spreading
population model
non linear fitting

ABSTRACT

In this note we analyze the temporal dynamics of the coronavirus disease 2019 outbreak in China, Italy and France in the time window 22/01 – 15/03/2020. A first analysis of simple day-lag maps points to some universality in the epidemic spreading, suggesting that simple mean-field models can be meaningfully used to gather a quantitative picture of the epidemic spreading, and notably the height and time of the peak of confirmed infected individuals. The analysis of the same data within a simple susceptible-infected-recovered-deaths model indicates that the kinetic parameter that describes the rate of recovery seems to be the same, irrespective of the country, while the infection and death rates appear to be more variable. The model places the peak in Italy around March 21st 2020, with a peak number of infected individuals of about 26000 (not including recovered and dead) and a number of deaths at the end of the epidemics of about 18,000. Since the confirmed cases are believed to be between 10 and 20% of the real number of individuals who eventually get infected, the apparent mortality rate of COVID-19 falls between 4% and 8% in Italy, while it appears substantially lower, between 1% and 3% in China. Based on our calculations, we estimate that 2500 ventilation units should represent a fair figure for the peak requirement to be considered by health authorities in Italy for their strategic planning. Finally, a simulation of the effects of drastic containment measures on the outbreak in Italy indicates that a reduction of the infection rate indeed causes a quench of the epidemic peak. However, it is also seen that the infection rate needs to be cut down drastically and quickly to observe an appreciable decrease of the epidemic peak and mortality rate. This appears only possible through a concerted and disciplined, albeit painful, effort of the population as a whole.

© 2020 Elsevier Ltd. All rights reserved.

Journal of Econometrics 220 (2021) 2–22



Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Panel forecasts of country-level Covid-19 infections[☆]

Laura Liu^{a,1}, Hyungsik Roger Moon^{b,c,d,2}, Frank Schorfheide^{e,f,g,h,*}

^aIndiana University, United States of America
^bUniversity of Southern California, United States of America
^cSchaeffer Center for Health Policy & Economics, USC, United States of America
^dYonsei University, Republic of Korea
^eUniversity of Pennsylvania, United States of America
^fCEPR, United Kingdom
^gNBER, United States of America
^hFIER, United States of America

ARTICLE INFO

Article history:
Received 13 May 2020
Received in revised form 21 August 2020
Accepted 24 August 2020
Available online 16 October 2020

JEL classification:
C11
C23
C53

Keywords:
Bayesian inference
Covid-19
Density forecasts
Interval forecasts
Panel data models
Random effects
SIR model

ABSTRACT

We use a dynamic panel data model to generate density forecasts for Covid-19 infections for a panel of countries/regions. Our specification the growth rate of active infections can be represented by autoregressive around a downward sloping deterministic trend function with a break. Our approach allows us to flexibly estimate the cross-sectional distribution then implicitly use this distribution as prior to construct Bayes fore individual time series. We find some evidence that information from 1 an early outbreak can sharpen forecast accuracy for late locations. There a lot of uncertainty about the evolution of active infection, due to param

Chaos, Solitons and Fractals 142 (2021) 110512



Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

Frontiers

Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods

Serkan Balçi¹

¹Department of Information Systems Engineering, Faculty of Technology, Muğla Sıtkı Koçman University, 48000, Muğla, Turkey

ARTICLE INFO

Article history:
Received 29 September 2020
Accepted 23 November 2020
Available online: 28 November 2020

Keywords:
Covid-19
Machine learning
Support vector machines
Multi-layer perceptron
Statistical distribution

ABSTRACT

The Covid-19 pandemic is the most important health disaster that has surrounded the world for the past eight months. There is no clear date yet on when it will end. As of 18 September 2020, more than 31 million people have been infected worldwide. Predicting the Covid-19 trend has become a challenging issue. In this study, data of COVID-19 between 20/01/2020 and 18/09/2020 for USA, Germany and the global was obtained from World Health Organization. Dataset consist of weekly confirmed cases and weekly cumulative confirmed cases for 35 weeks. Then the distribution of the data was examined using the most up-to-date Covid-19 weekly case data and its parameters were obtained according to the statistical distributions. Furthermore, time series prediction model using machine learning was proposed to obtain the curve of disease and forecast the epidemic tendency. Linear regression, multi-layer perceptron, random forest and support vector machines (SVM) machine learning methods were used. The performances of the methods were compared according to the RMSE, APE, MAPE metrics and it was seen that SVM achieved the best trend. According to estimates, the global pandemic will peak at the end of January 2021 and estimated approximately 80 million people will be cumulatively infected.

© 2020 Elsevier Ltd. All rights reserved.



Travel Medicine and Infectious Disease 37 (2020) 101742



Contents lists available at ScienceDirect

Travel Medicine and Infectious Disease

journal homepage: www.elsevier.com/locate/tmaid

Original article

Time series modelling to forecast the confirmed and recovered cases of COVID-19

Mohsen Maleki^a, Mohammad Reza Mahmoudi^{b,c,*}, Darren Wraith^d, Kim-Hung Pho^e

^aDepartment of Statistics, University of Isfahan, Isfahan, Iran
^bInstitute of Research and Development, Duy Tan University, Da Nang, 550000, Vietnam
^cDepartment of Statistics, Faculty of Science, Fasa University, Fasa, Fars, Iran
^dInstitute of Health and Biomedical Innovation (IHBI), Queensland University of Technology (QUT), Queensland, Australia
^eFractional Calculus, Optimization and Algebra Research Group, Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

ARTICLE INFO

Keywords:

ABSTRACT

Coronaviruses are enveloped RNA viruses from the Coronaviridae family affecting neurological, gastrointestinal, in late 2019 a new member of this family belonging to the Betacoronavirus originated and spread quickly across the world calling for strict containment is in the world, the outbreak of the disease has been serious and the number of 'confirmed' and 'recovered' COVID-19 cases helps planning to control the health care resources. Time series models based on statistical methodology are 1 and for forecasting. Autoregressive time series models based on two-piece 1, called TP-SMN-AR models, is a flexible family of models involving many 1d light/heavy tailed autoregressive models. In this paper, we use this family 1d time series data of confirmed and recovered COVID-19 cases.



ven verified by the Centers for Disease Control and Prevention (CDC), 1d also reported that COVID-19 is spreading by touching surfaces, one contact, air, or objects that contain viral particles. The incubation period of COVID-19 is at least 14 days [7], and it can spread to others in re incubation period. Finally note that the incubation period and median age of confirmed cases are respectively 3 days and 47.0 years

- Además, experiencia previa en la modelización y pronóstico de otros brotes epidémicos anteriores (fiebre hemorrágica, brucelosis, gripe, etcétera).

Una evaluación de modelos de pronóstico para COVID-19 en Cantabria

Instituto Cántabro de Estadística

Introducción

❑ Principales resultados de la bibliografía:

- Incertidumbre en la evolución de los casos
- Coherencia en pronósticos para diferentes regiones
- En general, mejor rendimiento de métodos de machine learning
- Adecuación en el uso de procesos ciclo-estacionarios
- Ausencia de datos fiables y comparables en primera instancia
- Empleo de casos diarios y fallecimientos para pronóstico



Metodología

❏ Datos

- Plataforma “Información Coronavirus” del Servicio Cántabro de Salud (SCS)
- Datos diarios de casos PCR y hospitalizados
- Recogida de datos: desde primer caso positivo hasta presente
- Ámbito geográfico: CCAA de Cantabria



Metodología

❑ Modelos de series temporales

- ARIMA
- REGRESIÓN LOCAL (LOESS)
- RED NEURONAL
- BOSQUE ALEATORIO
- K VECINOS CERCANOS



Comparativa de
modelos:

(“Metric performance”)

Plan de validación
(entrenamiento + test)
+

Factores de rendimiento

Metodología

❑ Modelos: ARIMA (Autorregresive Integrated Moving Average)

- Es la metodología más empleada en el tratamiento de series que presentan patrones estacionales.
- Fue desarrollado a finales de los sesenta del siglo XX. Box y Jenkins (1976) lo sistematizaron
- Especificación:
$$Y_t = -(\Delta^d Y_t - Y_t) + \phi_0 + \sum_{i=1}^p \phi_i \Delta^d Y_{t-i} - \sum_{i=1}^q \theta_i \epsilon_{t-i} + \epsilon_t$$
- ARIMA:
 - ✓ Autorregresivo (AR): Se especifican los valores previos utilizados de la serie para predecir los valores actuales.
 - ✓ Integrado (I): Este componente es necesario activarlo si hay tendencia. Elimina ese efecto usando diferencias.
 - ✓ Media móvil (MA): Evalúa la variable de interés en relación a otras variables cambiantes.

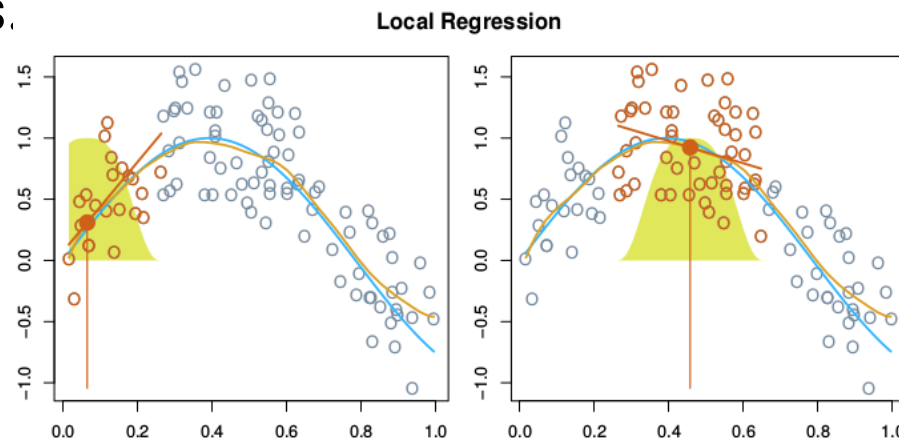
Metodología

❑ Modelos: Regresión local

- Aproximación no paramétrica para obtener modelos no lineales mediante reconfiguraciones de modelos clásicos.
- Se ajustan modelos simples a distintas subregiones del predictor, usando observaciones cercanas
- El proceso LOESS finaliza cuando el proceso se ha repetido para cada una de las "n" observaciones del conjunto de datos.

Parámetros necesarios:

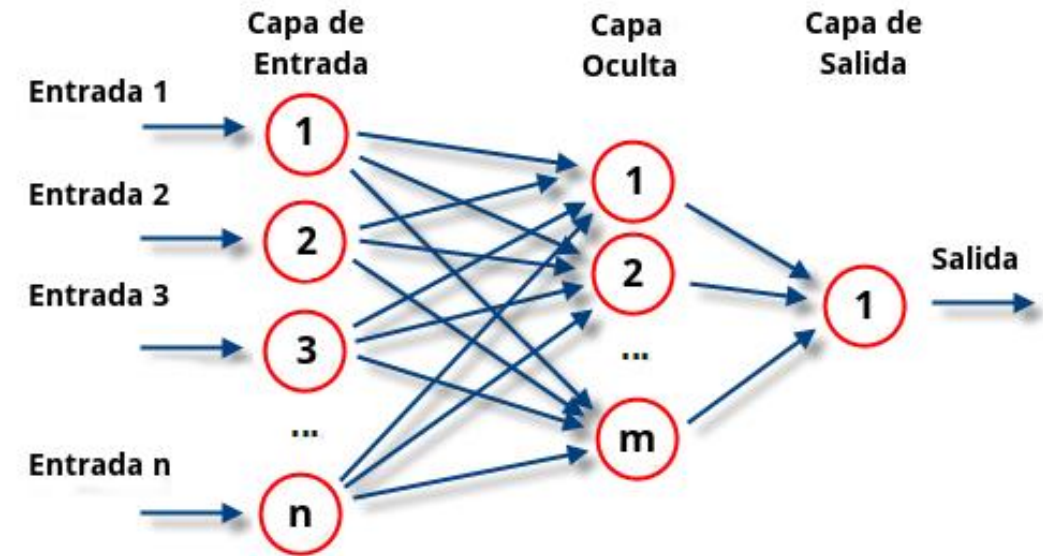
- ✓ **Span**
- ✓ **Grado del polinomio**
- ✓ **Función para designar pesos**



Metodología

Modelos: Red Neuronal

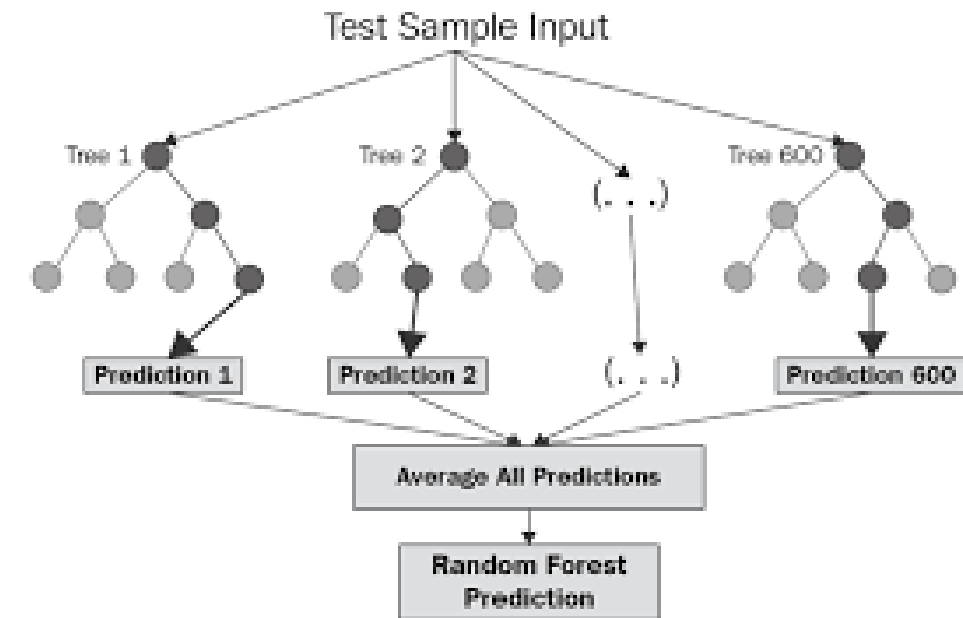
- Conjunto de unidades: “neuronas artificiales”
- Definición de pesos
- Proceso por etapas: aprendizaje y test
- Estructura:
 - ✓ Capas de entrada
 - ✓ Capas ocultas
 - ✓ Capas de salida
- Cálculo: suma ponderada inputs + función activación = output



Metodología

❑ Modelos: Bosque aleatorio

- Modelos predictivos de reglas binarias (no paramétrico)
- Conjunto de árboles de decisión individual
- Emplea técnicas de bootstrapping
- Útil para clasificación y predicción
- Sensibles a datos desbalanceados
- Evitar el overfitting (poda temprana)



Predicción
de Bosque
aleatorio

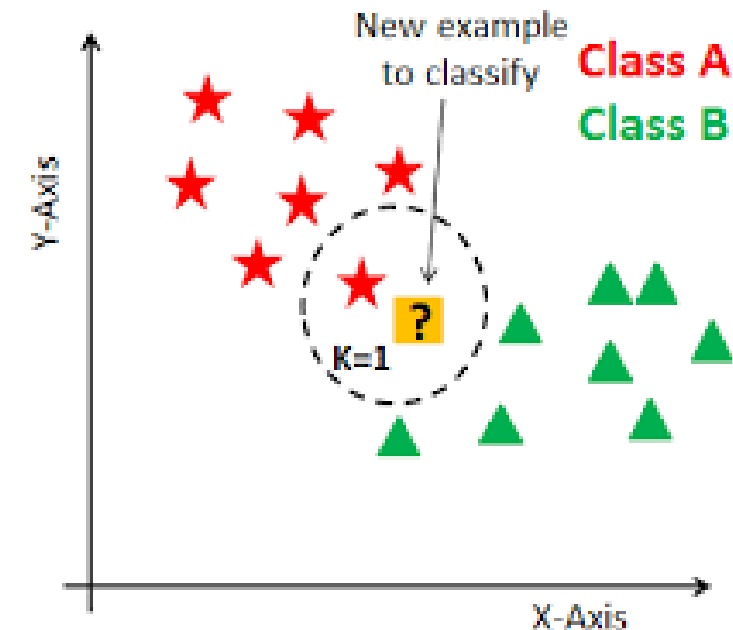
$$F(x) = \frac{1}{J} \sum_{j=1}^J c_{j_{full}} + \sum_{k=1}^K \left(\frac{1}{J} \sum_{j=1}^J contribution_j(x, k) \right)$$

Metodología

❑ Modelos: K-vecinos cercanos

➤ Algoritmo de clasificación

- ✓ Calcula distancia entre nuevas observaciones y otras en el conjunto de entrenamiento.
- ✓ Selecciona las "K" observaciones + cercanas.
- ✓ Determina la clase mayoritaria dentro del rango más cercano.
- ✓ Clasifica cada nueva observación según distancia en ese rango.
- ✓ En caso de empate, resuelve mediante esquemas predeterminados.



Resultados

I. Comparativa modelos: plan de validación a 30 días

1. Primera fase: **test**
2. Segunda fase: **training**

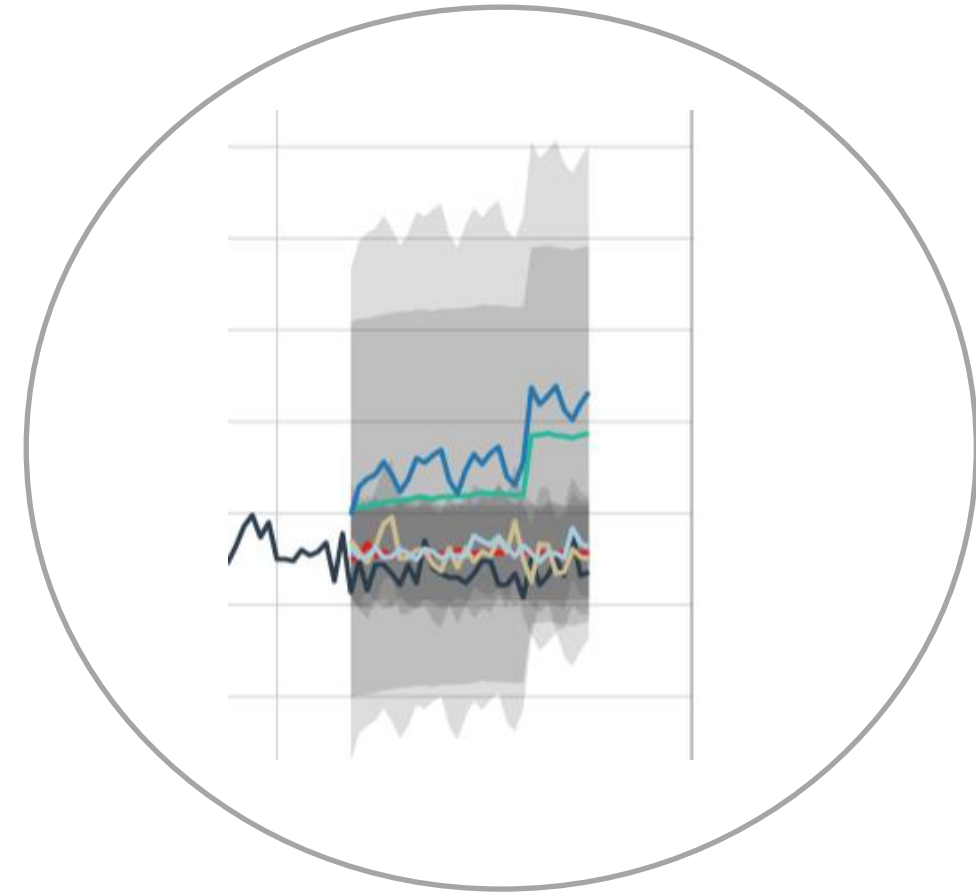
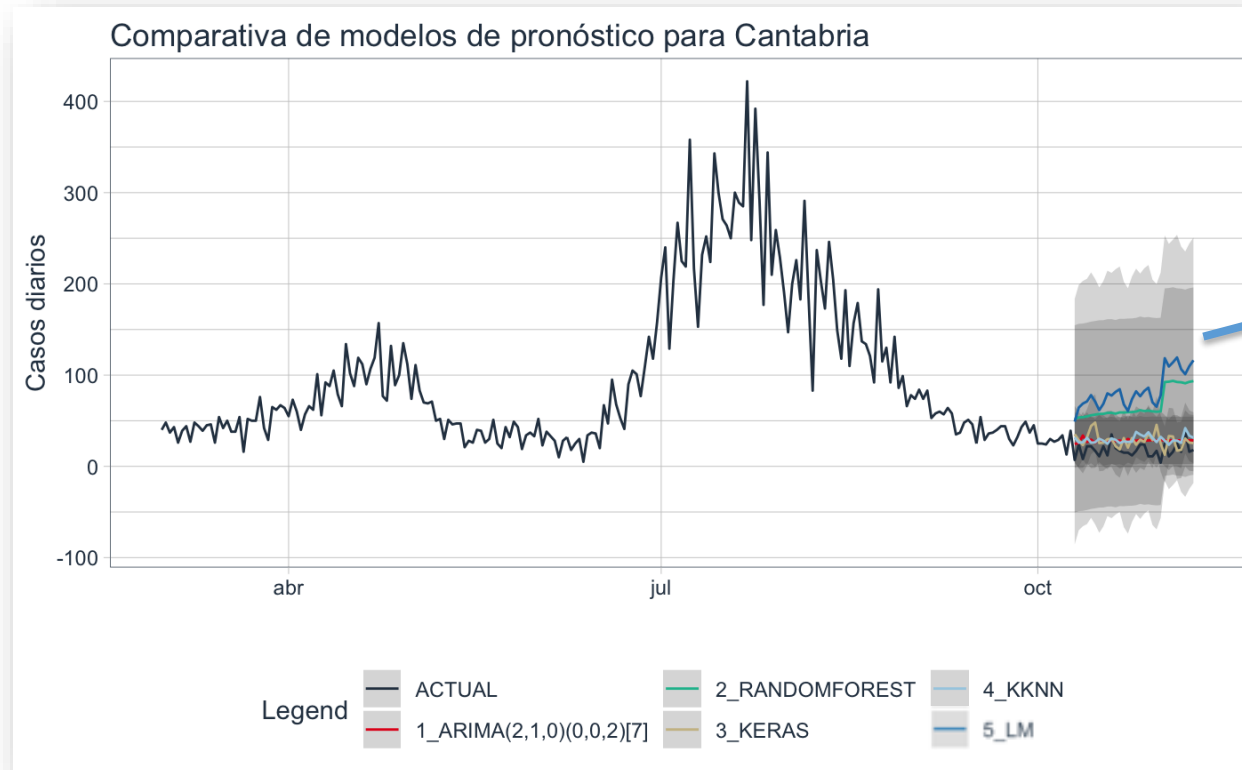
➤ Comparativa de modelos:

- ✓ ARIMA
- ✓ REGRESIÓN LOCAL (LOESS)
- ✓ RED NEURONAL
- ✓ BOSQUE ALEATORIO
- ✓ K VECINOS CERCANOS



Resultados

I. Comparativa de modelos



Resultados

I. Comparativa de modelos

	Factores de rendimiento				
MODELOS	MAE	MAPE	MASE	SMAPE	RMSE
ARIMA	11,61	102,09	1,23	55,03	13,11
BOSQUE ALEATORIO	49,54	358,13	5,26	117,41	51,92
RED NEURONAL	12,98	108,47	1,38	57,78	15,61
K-NN	12,26	109,16	1,30	56,29	14,10
REGRESIÓN LOCAL	65,43	462,85	6,95	130,21	68,02

■ Resultados comparativa:

a. ARIMA, red neuronal y K-NN 😊

b. Regresión local y bosque aleatorio 😞

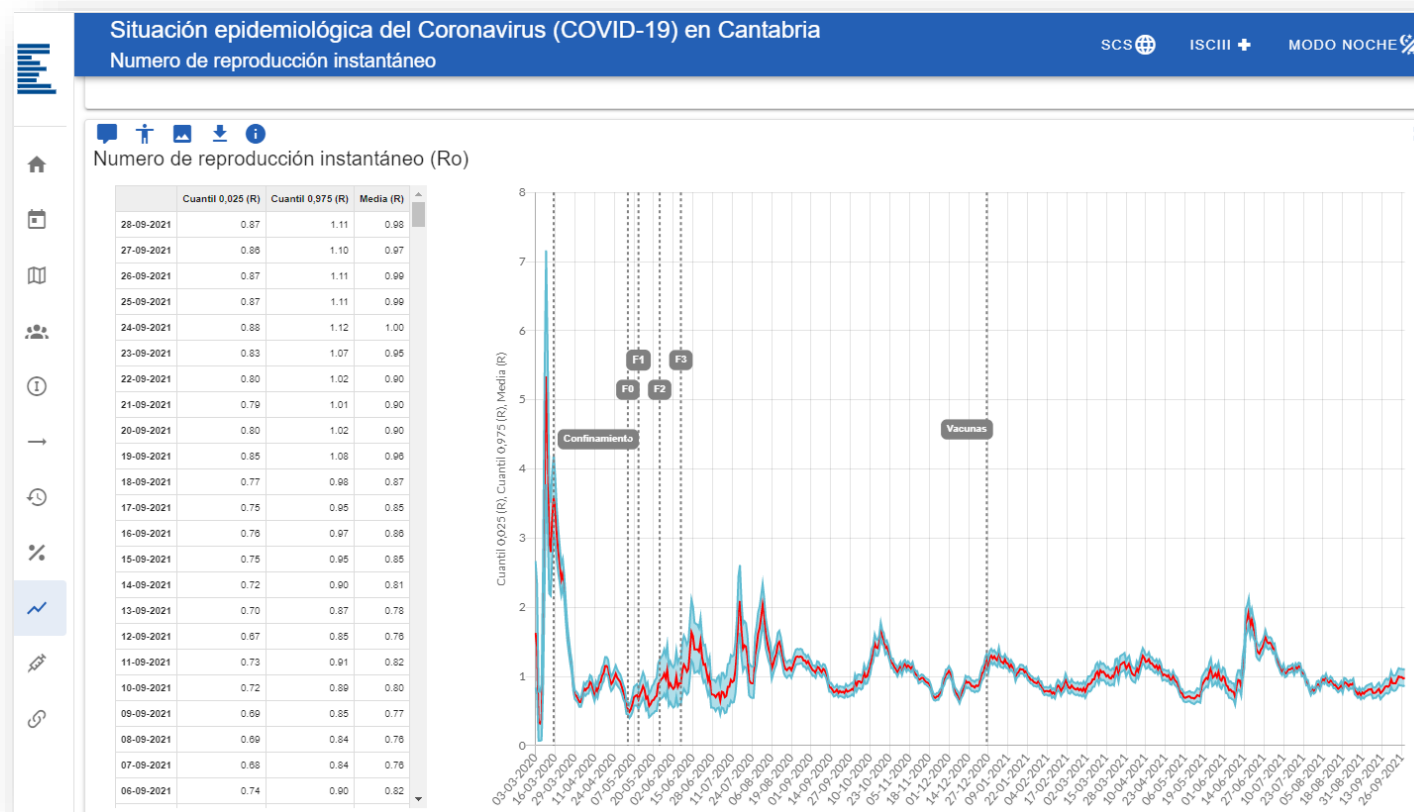
Los factores de rendimiento informan de la desviación de estimación de los pronósticos

Resultados

II. Aplicativo ICANE

➤ Número R_0

- ✓ N° promedio de casos nuevos que genera un caso dado a lo largo de un periodo infeccioso
- ✓ Cuando $R_0 < 1$ la infección muere tras un largo periodo. Sin embargo, si $R_0 > 1$ la infección tiende a propagarse.
- ✓ Para su cálculo se utilizó la función "estimate_R" de la librería "EpiEstim".

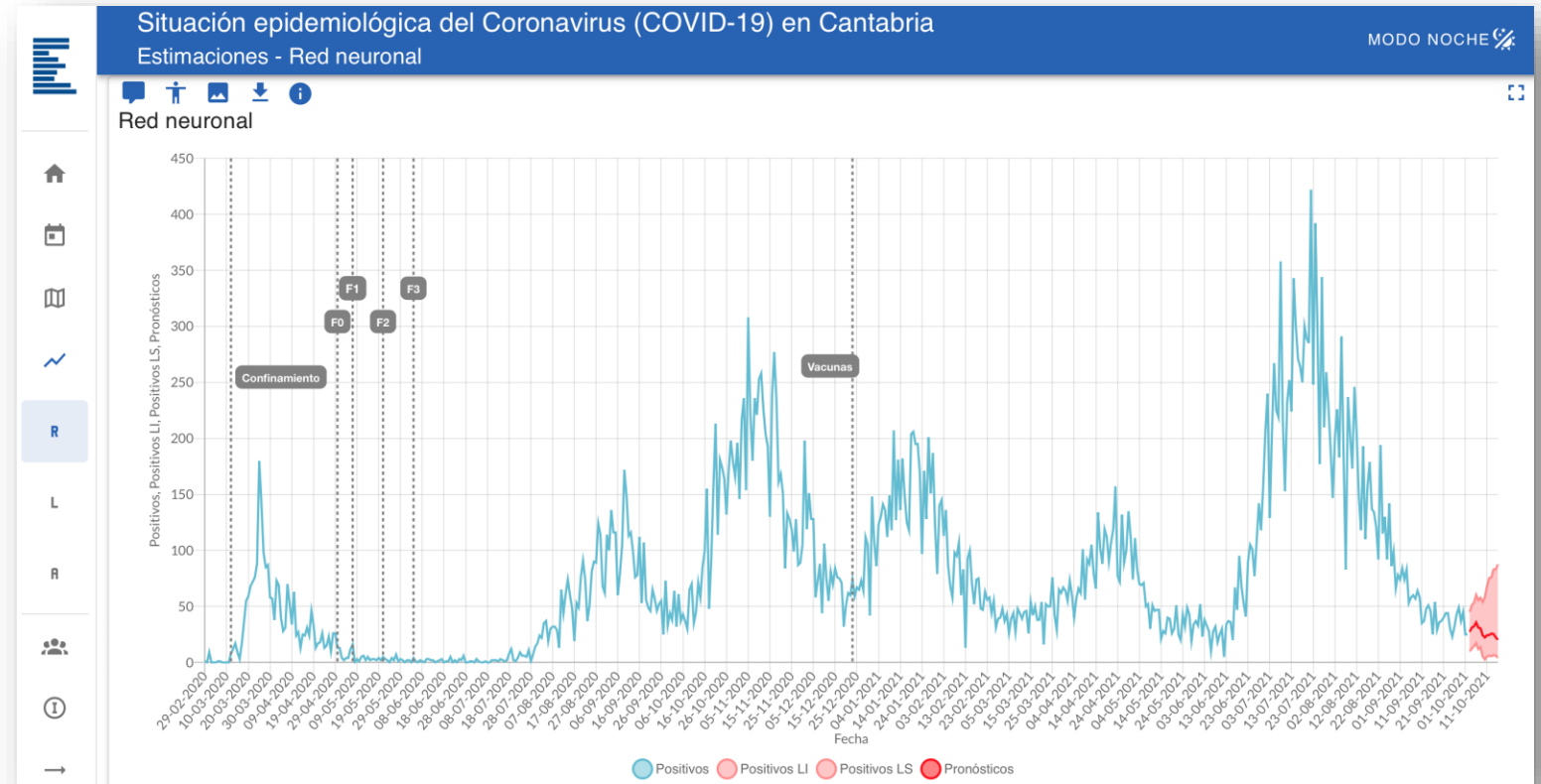


Resultados

II. Aplicativo ICANE

➤ Pronósticos

- ✓ Se incluyen modelos ARIMA, LOESS y red neuronal.
- ✓ Pronóstico de positivos SARS-CoV-2 por PCR en Cantabria.
- ✓ Los pronósticos fueron realizados utilizando R Studio Cloud.
- ✓ En nuestro análisis diario de los diferentes modelos, la red neuronal fue la que tuvo un mejor comportamiento predictor.



Resultados

II. Aplicativo ICANE:

- Principales tecnologías empleadas



Conclusiones

- I. Las redes neuronales, K-NN y la metodología ARIMA muestran en general un buen comportamiento en el pronóstico de casos PCR positivos.
- II. Sorprende el mal comportamiento predictor de otros modelos de “machine learning” como por ejemplo, el bosque aleatorio.
- III. La evidencia diaria del ICANE confirma los resultados mostrados en el comparativo de modelos.
- IV. A nivel informático, las herramientas de Cloud-Computing han sido fundamentales en el proceso de recogida y tratamiento de datos.
- V. El portal de pronóstico y número R_0 es solo una parte del aplicativo COVID-19, en el que el ICANE ha venido trabajando durante el último año y medio.

Más información en:

<https://www.icanes.es/covid19-home>



Islas Canarias
Del 15 al 19 de noviembre de 2021

¡Gracias!



Instituto Cántabro de **ESTADÍSTICA**



www.linkedin.com/in/gonzalo-llamosas-garcia



@gonzalollamosas