

Aprendizaje Estadístico e Inteligencia Artificial

Aplicaciones en Clima y Pronósticos, COVID-19, Finanzas y Legal-Tech

Manuel Pulido

Departamento de Física, Universidad Nacional del Nordeste

Instituto de Modelado e Innovación Tecnológica, CONICET

Instituto Franco-Argentino de Estudios Climáticos, IFAECl, CONICET/CNRS

<http://pulidom.github.io>

Líneas de investigación



Johann Kepler.

← Conectando →

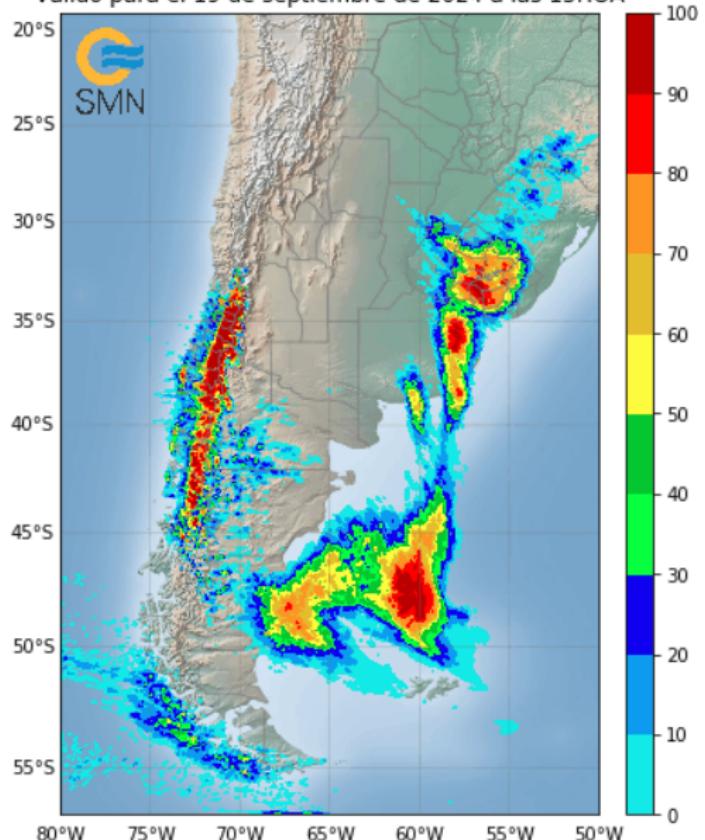


- ▶ Asimilación de datos. Desarrollo de algoritmos no-Gaussianos.
- ▶ Procesamiento de radares meteorológicos: precipitación, nowcasting.
- ▶ Procesamiento de imágenes SAR/SAOCOM: humedad de suelo.
- ▶ Asimilación de datos epidemiológicos: Dengue, COVID-19.
- ▶ Asimilación de datos y machine learning en finanzas.
- ▶ Entendimiento de procesos dinámicos climáticos/atmosféricos a través de aprendizaje estadístico.
- ▶ Inteligencia artificial aplicado a datos de empresas.

¿Qué es un pronóstico probabilístico?

Ensamble WRF - Probabilidad de precipitación (%) >1mm

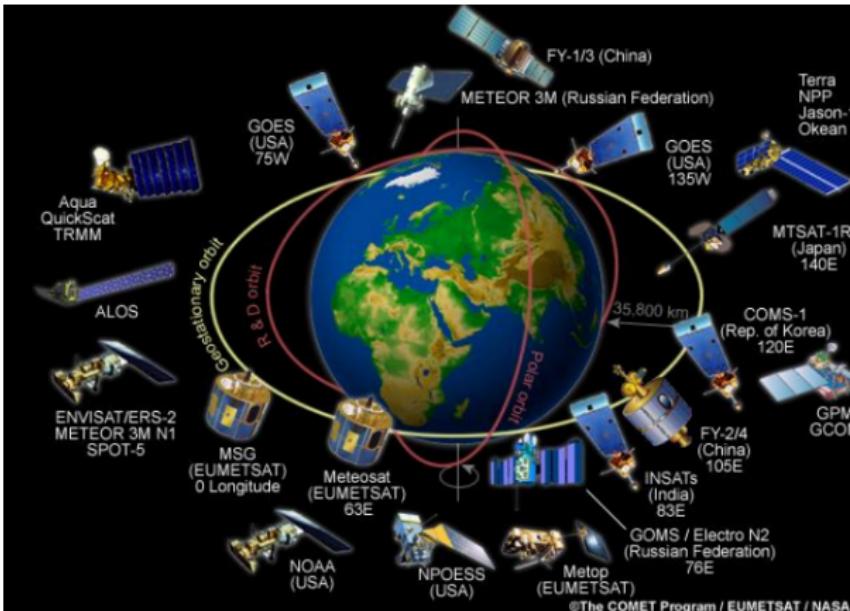
Válido para el 19 de septiembre de 2024 a las 15HOA



Inicializado el 18/9/2024 03 HOA

¿Como determinamos el clima o una tormenta?

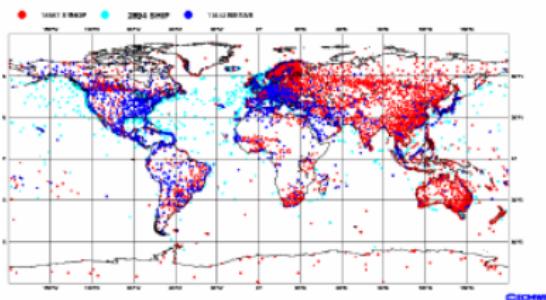
- ▶ Mediciones **localizadas** y **parciales**.
 - ▶ Los instrumentos realizan mediciones **indirectas** de variables.
 - ▶ Mediciones **ruidosas**.
 - ▶ Múltiples instrumentos.
-
- ▶ Sistemas muy complejos, i.e. la **atmósfera**, son imposibles de “medir” con un solo instrumento.



Observaciones en un periodo de 6 horas

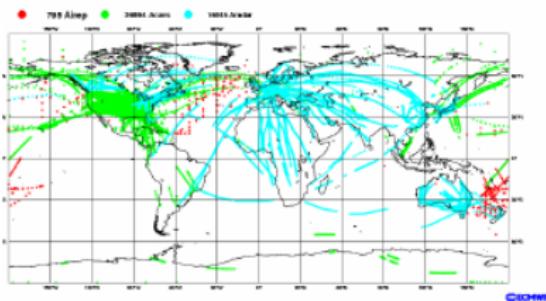
surface stations

ECMWF Data Coverage (All obs DA) - SYNOP/SHIP
12/NOV/2010; 00 UTC
Total number of obs = 31923



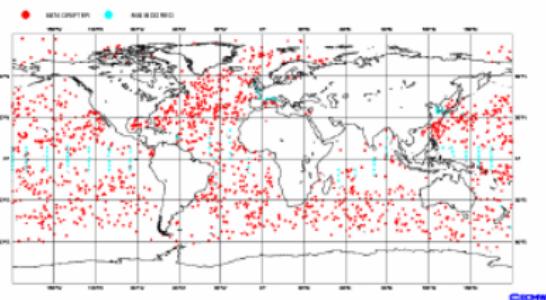
aircraft

ECMWF Data Coverage (All obs DA) - AIRCRAFT
12/NOV/2010; 00 UTC
Total number of obs = 53704



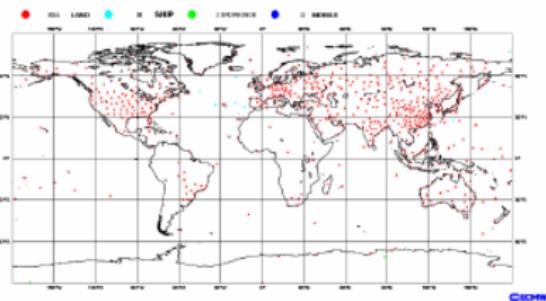
buoys

ECMWF Data Coverage (All obs DA) - BUOY
12/NOV/2010; 00 UTC
Total number of obs = 9423



radiosondes

**ECMWF Data Coverage (All obs DA) - TEMP
12/NOV/2010; 00 UTC**
Total number of obs = 644



Fusión de información

Como dice el proverbio “la unión hace la fuerza”: Combinamos todos los datos posibles, mediciones de todo tipo, y predicciones de modelos

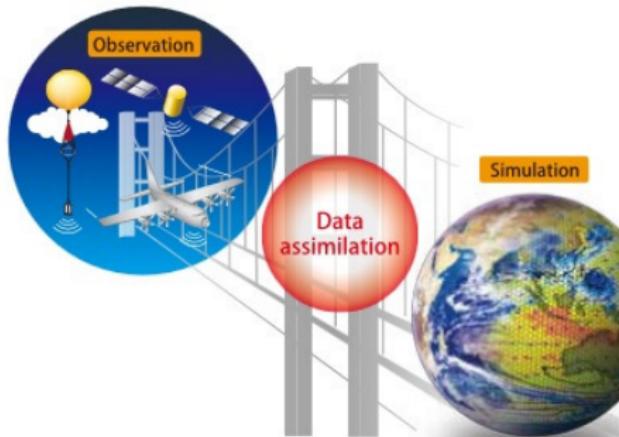
La fusión **ponderada** de información hace la fuerza...

Las observaciones **no** son la única fuente de información.

Conocemos las **leyes físicas o dinámicas del sistema**.

Podemos desarrollar modelos que nos **simulen** la evolución del sistema.

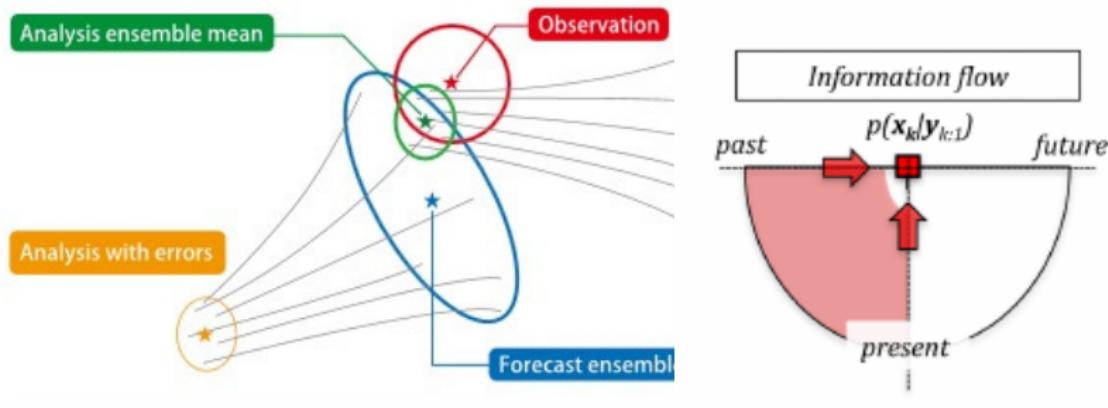
¿Como hacemos para fusionar dos datos que nos dan información disímil???



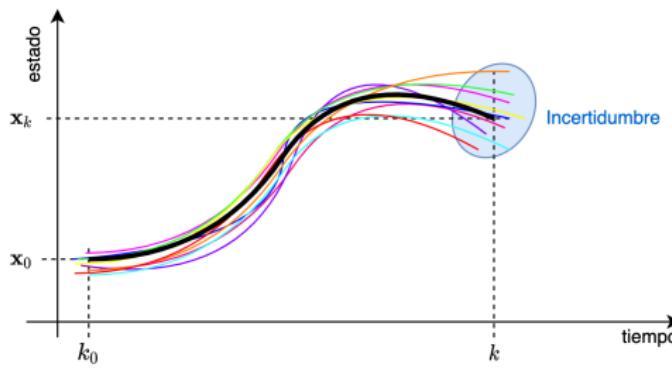
Incertezas y ensambles

Queremos **estimar el estado del sistema** (la situación en que se encuentra) dadas diversas fuentes de datos y con errores de mediciones.

Asimilación de datos es: **Cuantificar incertezas** de múltiples fuentes de datos.



Tratamiento de las incertidumbres



Fuentes de error

- ▶ Condiciones iniciales
No conocemos en forma exacta el estado del sistema.
- ▶ Error de modelo Los modelos son una simplificación de la realidad.

Técnica de Monte Carlo

Modelamos muchas realidades a través de distintos modelos que muestran una posible evolución del estado del sistema (Ensamble).

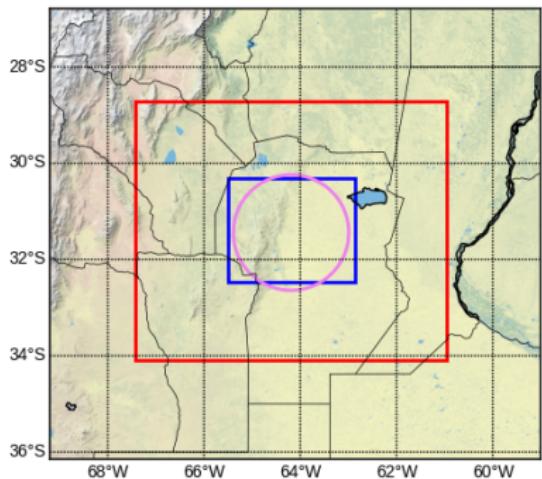


Pronóstico y Alerta de Eventos de Inundaciones Repentinasy

- ▶ Desarrollar sistemas de predicción hidrometeorológica de vanguardia para **prevenir inundaciones urbanas repentinasy (IUR)**.
- ▶ Transferencia inmediata de las advertencias a expertos y responsables del manejo de emergencias.
- ▶ Desarrollar una aplicación para teléfonos celulares que comunique eficientemente los pronósticos basados en impacto a la comunidad.
- ▶ Aumentar la conciencia y el conocimiento de la población sobre las IUR
- ▶ Acciones adecuadas que debe tomar la población en caso de emergencia.



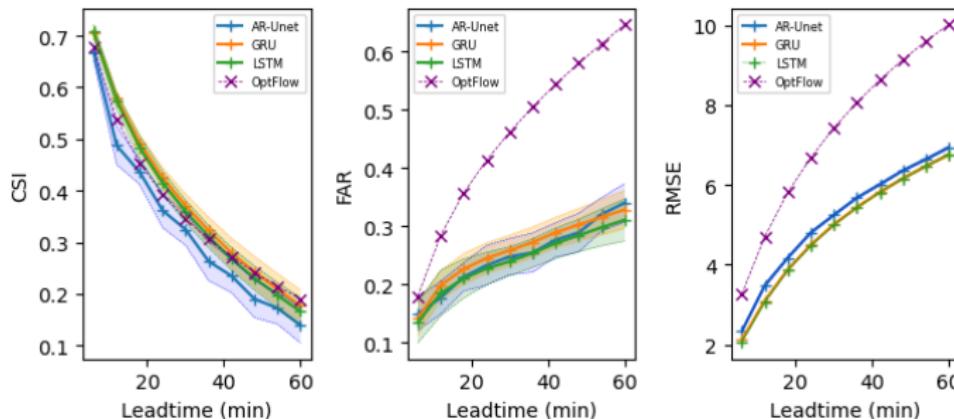
Base de datos: tormentas en Córdoba entre 2016 y 2023.



Datos sintéticos

- ▶ Se simulan las tormentas con el modelo de pronóstico WRF (Weather and Research Forecast).
- ▶ Dominio de 480km x 480km centrado en el radar RMA1
- ▶ 2km de resolución espacial. 6 minutos temporal.
- ▶ 188 tormentas en total.
- ▶ 5500 muestras para entrenamiento. 1000 para validación. 1000 para testing.

Performance UNET para pronóstico de eventos extremos.



CSI: Critical Success Index.

Cantidad de veces que fui capaz de predecir el evento.

FAR: False Alarm Ratio

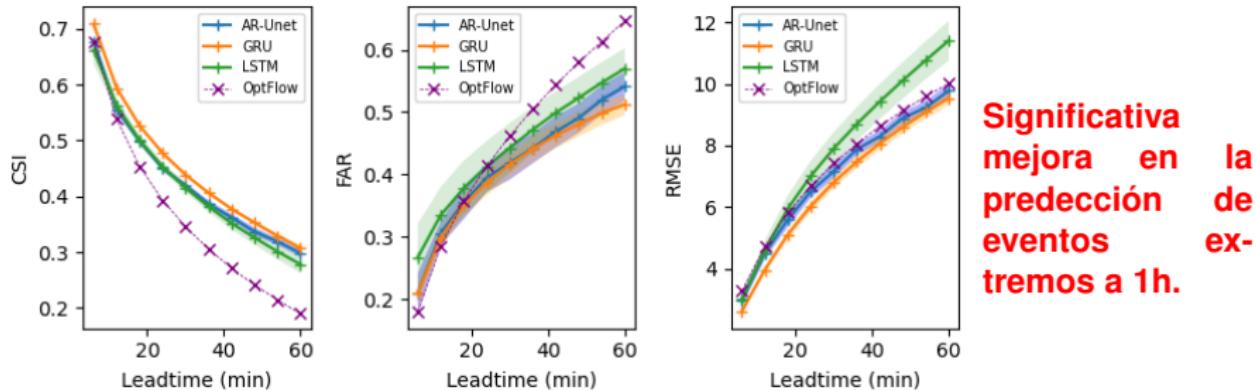
Cantidad de veces que predigo un evento y NO ocurre.

Se compara con la técnica clásica de flujo óptico.

Gran dificultad para predecir eventos extremos.

Las redes neuronales entranan para predecir eventos/casos típicos.

¿Como detectar eventos extremos con redes neuronales?



Se propone pesar en la función de pérdida por la probabilidad de ocurrencia de un evento:

$$L(\mathbf{w}) = \mathcal{E} \left[(\mathbf{y} - f(\mathbf{x}, \mathbf{w}))^\top (\mathbf{y} - f(\mathbf{x}, \mathbf{w})) \right]$$
$$L_w(\mathbf{w}) = \mathcal{E} \left[\frac{1}{p_y(\mathbf{x})} (\mathbf{y} - f(\mathbf{x}, \mathbf{w}))^\top (\mathbf{y} - f(\mathbf{x}, \mathbf{w})) \right]$$

Precio: Aumento de las falsas alarmas. Pero el FAR sigue siendo menor a optical flow.

Lo importante NO es la arquitectura. Gran impacto de la función de pérdida.

Cuantificación de la incertidumbre

- ▶ En pronósticos meteorológicos y sistemas de alertas **la asimilación de datos** juega un rol fundamental.

Otras aplicaciones que desarrollamos con la misma metodología:

- ▶ Estimación y predicción de contagios de **COVID-19**.
- ▶ Predicciones de arribos de clientes (o llamadas telefónicas)
- ▶ **Finanzas:** Predicción del valor de las acciones y su volatilidad.



**Podemos ganar plata usando
asimilación de datos???**



Cuantificación de la incertidumbre con redes neuronales

El **objetivo general** cuantificar la **incertidumbre** asociada a un estado de un sistema dinámico.

Nuestra **hipótesis** es que las redes neuronales artificiales pueden ser usadas para cuantificar la incertidumbre de un pronóstico y su dependencia con el estado del sistema a partir de un pronóstico determinístico .

Pronóstico Determinístico

$$\mathbf{x}_l^d = \mathcal{M}_l(\bar{\mathbf{x}}_a)$$

$$\mathbf{x}_0^d = \langle x_0^{d,0}, x_0^{d,1}, \dots, x_0^{d,s} \rangle$$

:

$$\mathbf{x}_l^d = \langle x_l^{d,0}, x_l^{d,1}, \dots, x_l^{d,s} \rangle$$

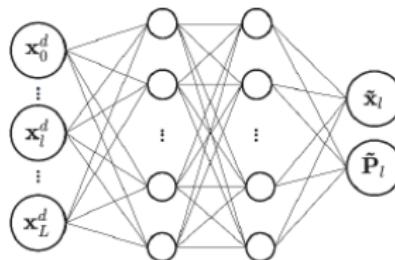
:

$$\mathbf{x}_L^d = \langle x_L^{d,0}, x_L^{d,1}, \dots, x_L^{d,s} \rangle$$

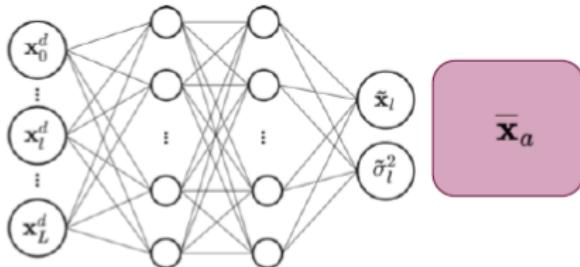
Pronóstico Probabilístico

$$\mathcal{N}(\tilde{\mathbf{x}}_l, \tilde{\mathbf{P}}_l)$$

$$\mathcal{N}(\tilde{\mathbf{x}}_l, \tilde{\sigma}_l^2), \quad \tilde{\sigma}_l^2 = \text{diag}(\tilde{\mathbf{P}}_l)$$



Metodología - Entrenamiento



Nosotros proponemos ([Sacco et al. 2022](#)), como alternativa, minimizar la siguiente función de costo que denominamos **Extended Root Mean Squared Error (eMSE)**

$$\mathcal{L}_{eMSE} = \|\tilde{\mathbf{x}}_l - \bar{\mathbf{x}}_l^a\|^2 + \gamma \|\tilde{\sigma}_l^2 - (\tilde{\mathbf{x}}_l - \bar{\mathbf{x}}_l^a)^2\|^2$$

Nos referiremos a esta estrategia como aprendizaje indirecto o entrenamiento indirecto

Monitoreo y predicción de COVID-19 con asimilación de datos

Proyecto CORR 01 COVID FEDERAL

Datos:

- Base de datos del SNVS
- Reporte diario de salud.
- Scraping de sitios web.

Modelo:

Modelo compartimental de metapoblación
del tipo SEIRHD por edades.

$$\frac{\partial S_j}{\partial t} = - \frac{S_j}{\tau^I N_j} \sum_{k=1}^n \lambda_{jk} I_k + \frac{R_j}{\tau^R}$$

$$\frac{\partial E_j}{\partial t} = \frac{S_j}{\tau^I N_j} \sum_{k=1}^n \lambda_{jk} I_k - \frac{E_j}{\tau^E}$$

$$\frac{\partial I_j}{\partial t} = \frac{E_j}{\tau^E} - \frac{I_j}{\tau^I}$$

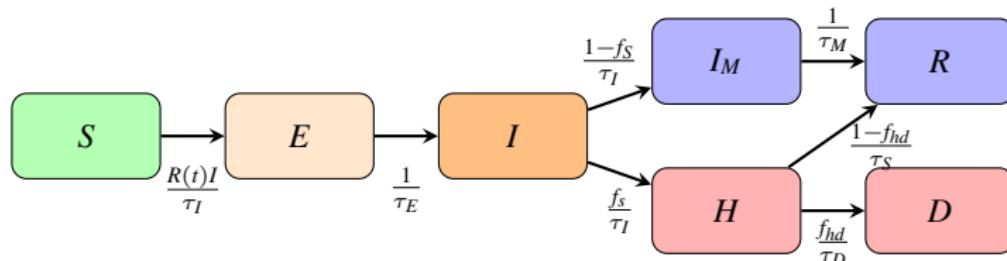
$$\frac{\partial T_j}{\partial t} = f_j^T \frac{I_j}{\tau^I} - \frac{T_j}{\tau^T}$$

$$\frac{\partial C_j}{\partial t} = f_j^C \frac{I_j}{\tau^I} - \frac{C_j}{\tau^C}$$

$$\frac{\partial M_j}{\partial t} = (1 - f_j^T - f_j^C) \frac{I_j}{\tau^I} - \frac{M_j}{\tau^M}$$

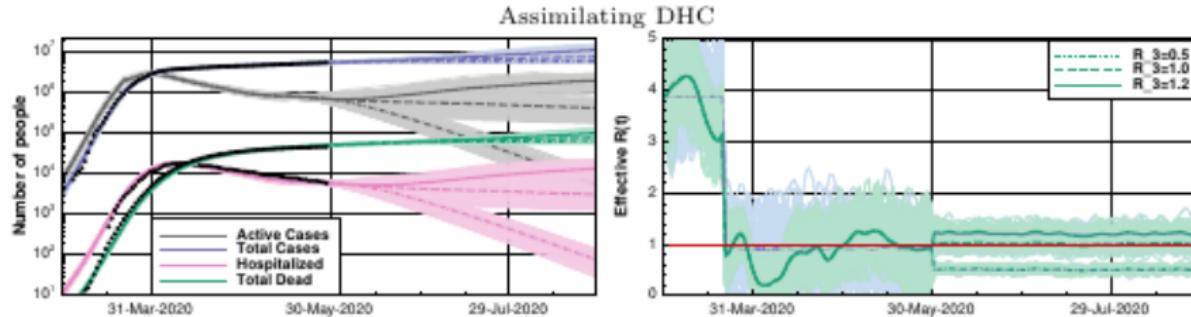
$$\frac{\partial D_j}{\partial t} = \frac{T_j}{\tau^C}$$

$$\frac{\partial R_j}{\partial t} = \frac{M_j}{\tau^M} + \frac{T_j}{\tau^T} - \frac{R_j}{\tau^R}$$

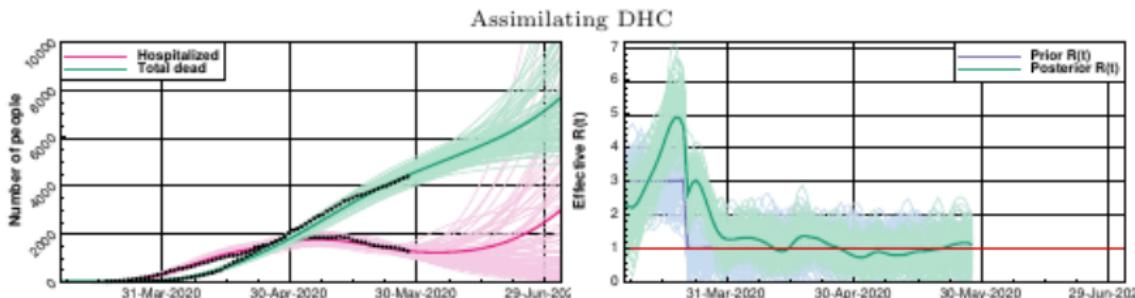


Incógnitas en COVID-19 con asimilación de datos

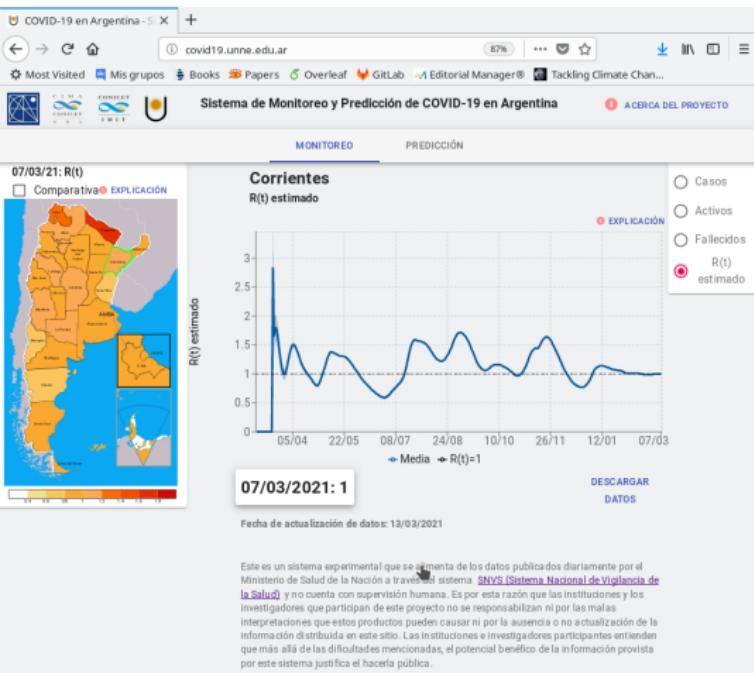
¿Cómo modelamos los contagios? ¿Cómo detectamos la cuarentena?



¿Cómo predecimos la evolución a partir de "hoy"?



Desarrollo de una aplicación web pública



El sistema en forma automática baja los datos de las distintas fuentes y los procesa **en tiempo real**.

Los resultados → son transferidos a una aplicación web:

- Desarrollada nativamente en React JS.
- Visualización de los datos de análisis
- Permite en forma interactiva elegir variables de distintas provincias y compararlas.

<http://covid19.unne.edu.ar/>

AN INTERNATIONAL INITIATIVE OF PREDICTING THE SARS-COV-2 PANDEMIC USING ENSEMBLE DATA ASSIMILATION

PLOS ONE

GEIR EVENSEN^{*,1}, JAVIER AMEZCUA², MARC BOQUET³,
ALBERTO CARRASSI^{2,4}, ALBAN FARCHI³, ALISON FOWLER²,
PIETER L. HOUTEKAMER⁵, CHRISTOPHER K. JONES⁶,
RAFAEL J. DE MORAES⁷ MANUEL PULIDO⁸,
CHRISTIAN SAMPSION⁶, AND FEMKE C. VOSSEPOEL⁷

¹NORCE and NERSC
Bergen, Norway

²Dept. of Meteorology
University of Reading and NCEO, UK

³CEREIA, joint laboratory École des Ponts ParisTech and EDF R&D
Université Paris-Est, Champs-sur-Marne, France

⁴Mathematical Institute
University of Utrecht, Netherlands

⁵Environment and Climate Change Canada
Dorval, Québec, Canada

⁶Renaissance Computing Institute
University of North Carolina, Chapel Hill, USA

⁷Department of Geoscience and Engineering
Delft University of Technology, Delft, Netherlands

⁸FaCENA, UNNE and IMIT, CONICET
Corrientes, Argentina



RESEARCH ARTICLE

Inference in epidemiological agent-based models using ensemble-based data assimilation

Tadeo Javier Cocucci^{1,2*}, Manuel Pulido^{2,3,4}, Juan Pablo Aparicio^{5,6}, Juan Ruiz^{7,8}, Mario Ignacio Simoy^{9,9}, Santiago Rosa^{1,2}

¹FaMAF, Universidad Nacional de Córdoba, Córdoba, Córdoba, Argentina, ²FaCENA, Universidad Nacional del Nordeste, Corrientes, Corrientes, Argentina, ³IAECA, CONICET, Corrientes, Corrientes, Argentina, ⁴IMIT, CONICET, Corrientes, Corrientes, Argentina, ⁵INENCO, CONICET, Universidad Nacional de Tucumán, Tucumán, Argentina, ⁶Saint Louis University, St. Louis, Missouri, United States of America, ⁷CIMA, CONICET, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Buenos Aires, Argentina, ⁸Southern High Altitude Atmosphere and the Oceans Department, FCEN, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Buenos Aires, Argentina, ⁹ Instituto Multidisciplinario sobre Ecosistemas y Desarrollo Sustentable, Universidad Nacional del Centro de la Provincia de Buenos Aires, Tandil, Argentina

* tadeojcocucci@gmail.com

OPEN ACCESS

Citation: Cocucci TJ, Pulido M, Aparicio JP, Ruiz J, Simoy MI, Rosa S (2022) Inference in epidemiological agent-based models using ensemble-based data assimilation. PLoS ONE 17(3): e0264992. <https://doi.org/10.1371/journal.pone.0264992>

Editor: Universitat Autònoma de Barcelona

Abstract

To represent the complex individual interactions in the dynamics of disease spread informed by data, the coupling of an epidemiological agent-based model with the ensemble Kalman filter is proposed. The statistical inference of the propagation of a disease by means of ensemble-based data assimilation systems has been studied in previous works. The models used are mostly compartmental models representing the mean field evolution through ordinary differential equations. These techniques allow to monitor the propagation of the infec-

Cocucci et al, Plos One (2022).

ABSTRACT. This work demonstrates the efficiency of using iterative ensemble smoothers to estimate the parameters of an SEIR model. We have extended a standard SEIR model with age-classes and compartments of sick, hospitalized, and dead. The data conditioned on are the daily numbers of accumulated deaths and the number of hospitalized. Also, it is possible to condition the model on the number of cases obtained from testing. We start from a wide prior distribution for the model parameters; then, the ensemble conditioning

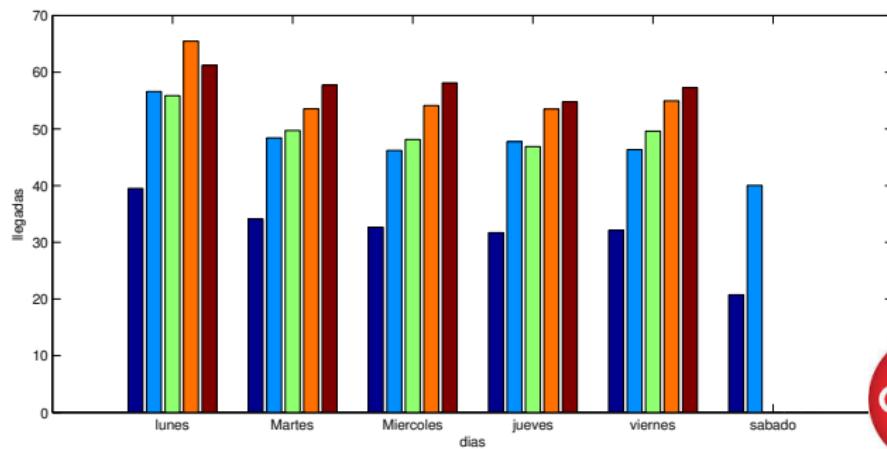
Sistema de espera para atención al cliente de empresa telefónica

¿Cuantos servidores se requieren para atender dado un flujo de ingreso de clientes?



Se aplica Erlang (Queueing theory)

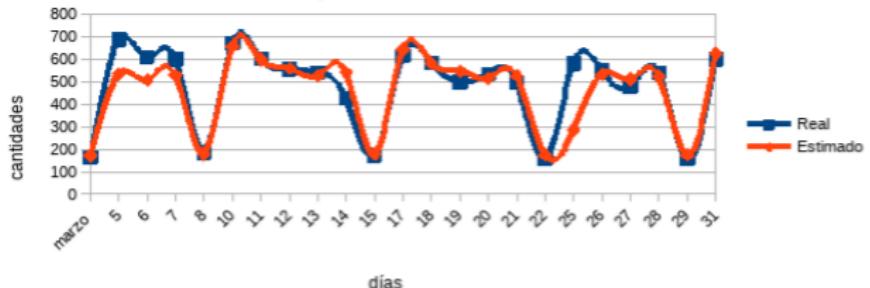
- ▶ Distintas líneas de espera
- ▶ Distintos servidores
- ▶ Calidad de atención: Tiempo de espera máximo?



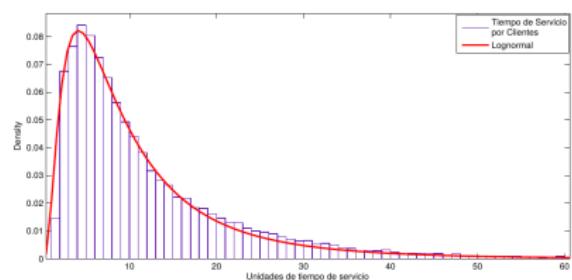
Predicción llegada de clientes



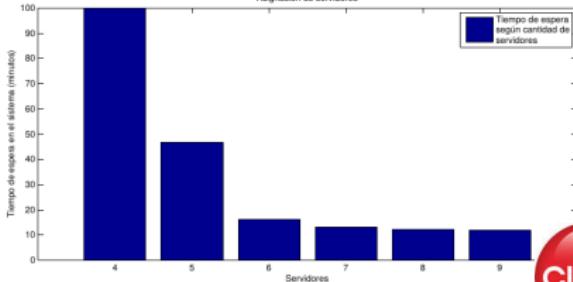
comparación marzo centro 55



Predicción de Clientes (basado en ARIMA)



Tiempo de atención



Tiempo de Espera

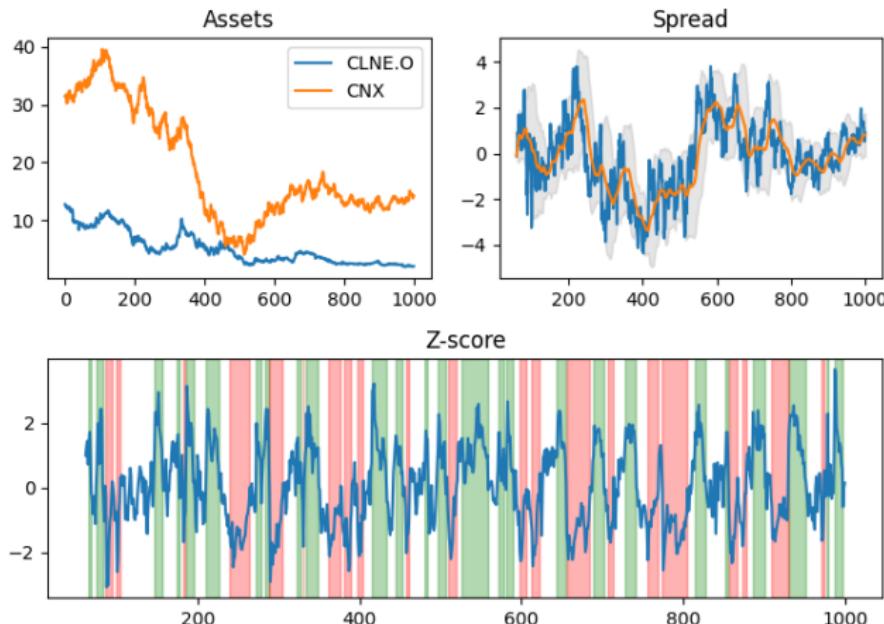


Incertidumbre en finanzas

Las bolsas en el mundo muestran periodos de fuerte volatilidad.

Podemos detectar los cambios de volatilidad de antemano con asimilación?

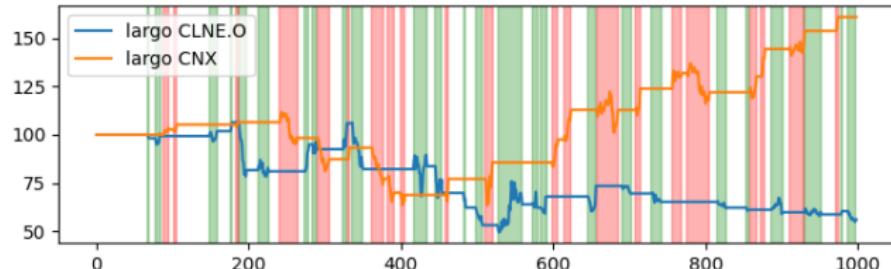
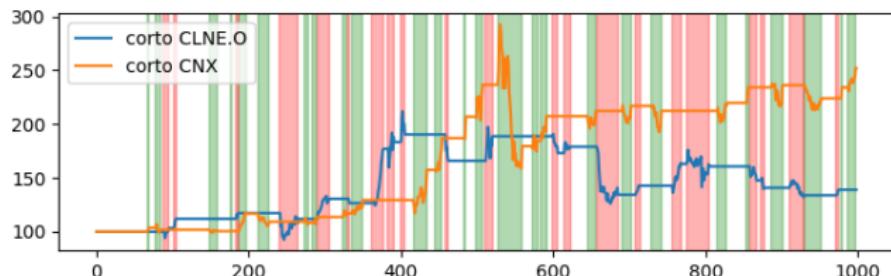
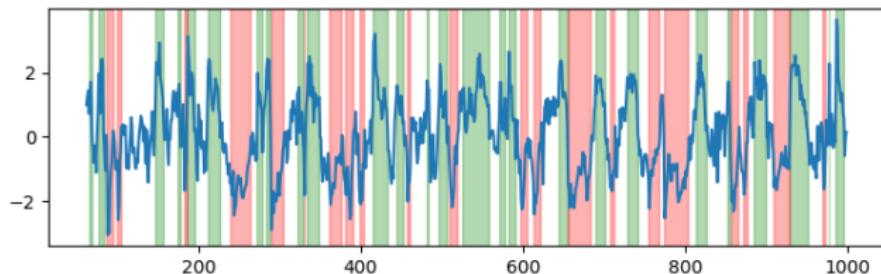
Proyecto para el fondo de inversión Dobby Quantitative Investments (DQI). Gracias a Trump!



- ▶ Muy similar a COVID.
- ▶ Se van asimilando datos a tiempo real y se van realizando predicciones.
- ▶ Se buscan desequilibrios del mercado para invertir

Estrategia de inversión por pair trading

Z-score



Uso de LLMs para Legal-Tech

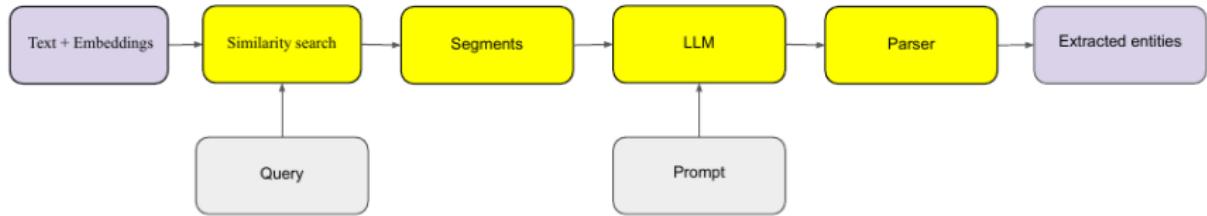
- ▶ La extraccion de informacion de accidentes de trabajo contenidas en sentencias judiciales es crucial para las ART para poder cuantificar los costos de los seguros.
- ▶ Los jueces asignan un valor punto y un porcentaje de discapacidad monto que debe ser previsto por las aseguradoras.

Objetivo:

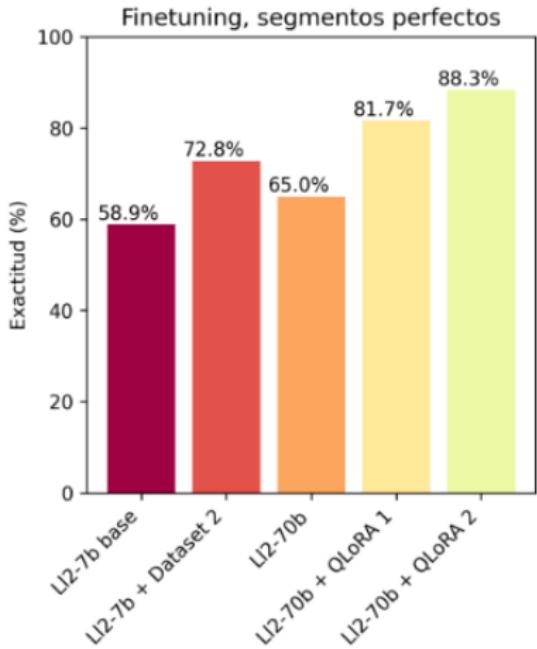
Se debe extraer entidades de sentencias que nos permitan determinar el grado de incapacidad fisica, psicologica, el tipo, y los respectivos montos.

Desafio: Este proceso es muy complejo aun para expertos legales, ya que en las sentencias se usan numerosos argumentos y se discuten los puntos de vista de las distintas partes.

Pipeline propuesto

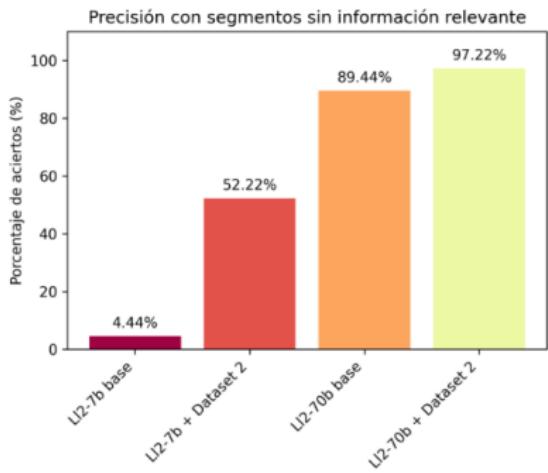


Entrenamiento de grandes modelo de lenguaje Llama-2 vs GPT4



- ▶ Significativa mejora cuando al modelo Llama le hacemos finetuning.
- ▶ El modelo Llama-7b finetuned tiene una mejor performance que el modelo Llama-70b base.
- ▶ Se obtiene para el mejor modelo una precision de 88.3%. El modelo GPT-4 turbo tienen una de 86.1%!.

¿Porque la gran mejora en la performance?



- ▶ La precision para segmentos que no tienen información relevante.
- ▶ El modelo Llama-7b alucina en esos segmentos.
- ▶ El modelo fine-tuned mejora significativamente y **disminuye las alucinaciones**.
- ▶ El impacto de las alucinaciones es relativamente menor en modelos grandes (70b)

Primer premio en el ASAID!

Extracción de entidades en sentencias judiciales usando LLaMA-2

Francisco Vargas

Alejandro Gonzalez Coene

Gaston Escalante

Exequiel Lobón

Manuel Pulido



Palabras clave: Reconocimiento de entidades nombradas, Grandes modelos de lenguaje, Textos legales



ASAID - Simposio Argentino de Inteligencia Artificial y Ciencia de Datos

Muchas gracias