Ficha Técnica Predicción Cooperativa

Contenido

1. Objetivo **1**

2. Variables y registros oficiales **1**

2.1. Modificaciones en los registros oficiales 2

3. Participantes **2**

4. Predictores cooperativos **3**

4.1. Motivación 3

4.2. Limitaciones 3

4.3. Predictores cooperativos 4

Referencias **5**

1. Objetivo

Obtener y evaluar predicciones cooperativas de cinco variables de interés en el contexto de la expansión del virus Covid-19 utilizando predicciones diarias con un amplio de abanico de modelos desarrollados por investigadores en el ámbito de la comunidad Matemática / Estadística / Científica de Datos que participan en esta iniciativa. Las predicciones se obtienen para cada una de las Comunidades Autónomas y también agregadas para el global de España. Los horizontes de predicción a considerar son de 1 a 7 días, ambos incluidos.

2. Variables y registros oficiales

Las variables de interés en el estudio son:

Número de ingresos en UCI	*uci*
Número de enfermos hospitalizados	*hospitalizados*
Número de fallecimientos	*fallecidos*
Número de nuevos casos	*nuevos*
Número de casos confirmados	*confirmados*

La precisión de las predicciones se evalúa conforme a la base de datos oficiales que diariamente publica el Instituto de Salud Carlos III (ISCIII) para cada una de las CCAA en la url `https://covid19.isciii.es/` y que se puede descargar directamente bajando el archivo `https://covid19.isciii.es/resources/serie_historica_acumulados.csv`. Los totales para España se obtienen agregando los registros de la totalidad de CCAA.

Las variables *uci*, *hospitalizados* y *fallecidos* reciben el mismo nombre en el archivo del ISCIII. En cambio la variable *confirmados* se denomina *CASOS* en este archivo y la variable *nuevos* no se incluye. Para obtener los valors de *nuevos* se calcula el incremento en el número de casos acumulados (*confirmados*). Por consiguiente, todas las variables excepto *nuevos* registran datos acumulados hasta la última fecha. En el repositorio `en https://rubenfcasal.github.io/COVID-19/` se puede bajar el archivo `acumula2.RData` con los datos del ISCIII en formato adecuado para esta iniciativa de elaborar predictores cooperativos.

2.1. Modificaciones en los registros oficiales

Es muy importante enfatizar que el ISCIII ha venido proporcionando información sobre cambios sustanciales en los registros publicados. Por ejemplo, desde 2020-04-02 se ha venido informando que los valores de *hospitalización* y *UCI* reportados por Castilla-La Mancha (CM), Castilla y León (CL), Comunidad Valenciana (VC), Madrid (MD) y Galicia (GA), son datos de prevalencia (personas ingresadas en la correspondiente fecha) y no reflejan el total de personas que han sido hospitalizadas o ingresadas en UCI a lo largo del periodo de notificación (al contrario de lo que en principio reporta el resto). Esto ha ido cambiando a lo largo del tiempo. Actualmente, únicamente Madrid (MD) reporta valores de prevalencia de hospitalizados, y en el caso de UCI, los valores de Madrid (MD), Castilla y León (CL) y Galicia (GA) son de prevalencia. Obviamente, estos diferentes criterios quitan además sentido al agregado de estas variables para España (ES).

Desde CEMat se ha solicitado reiteradamente a las autoridades la necesidad de proporcionar datos congruentes y homogéneos para poder obtener análisis y predicciones rigurosas. Más allá de esto, solo nos resta advertir de estas inconsistencias en los datos para poder entender algunos comportamientos anómalos de las series.

3. Participantes

En el momento de elaborar esta ficha (16 de abril), un total de 58 investigadores/grupos se han inscrito formalmente en Predicción Cooperativa, de los cuales 42 han enviado ya predicciones para alguna/s variable/s en una o más CCAA.

El abanico de técnicas empleadas es realmente muy amplio, incorporando modelos de regresión funcional, ecuaciones diferenciales ordinarias, modelos espacio-temporales, modelos SEIR y SEAIDR con ajuste a partir de los datos de Italia, modelos VAR de series temporales, modelos logístico discretos, Monte Carlo predictivo bayesiano, regresión loglineal con tendencia cuadrática, curvas de crecimiento tipo Gompertz, simulación de eventos discretos, modelo de Richards, modelo lineal autoregresivo no estacionario, modelos de

Física de partículas, boosting, random forest, modelos autoregresivos, SIR con simulación multiagente, suavizado exponencial, combinación de SIOPRED con un método naive con tendencia, sistemas dinámicos, modelos de regresión para datos composicionales, modelos ocultos de Markov, sistemas expertos con metodología Bayesiana, regresión no lineal con modelos compartimentales, modelos SIRV y SIRM, dinámica de poblaciones por ajuste de curvas, modelos lineales generalizados, modelos SIR con tasa de transición dinámica, ecuaciones de estimación generalizada con suavizado tipo spline, modelo de Hurdle, predicción de series temporales mediante procesos Gaussianos, métodos combinados de aprendizaje, modelos SEIR con compartimentos adicionales y modelos de regresión dinámica.

4. Predictores cooperativos

4.1. Motivación

La estrategia de combinar predicciones obtenidas desde diferentes métodos fue inicialmente propuesta por Bates y Granger [1]. El objetivo es encontrar combinaciones óptimas de predicciones individuales que conduzcan a predicciones más precisas y estables. Desde el trabajo seminal de Bates y Granger, un buen número de criterios de combinación de predictores han sido propuestos en la literatura (ver p.e. Timmerman 2006 [4] o Clements *et al.* 2012 [3]). Sin embargo, mientras que no parece cuestionable el interés de combinar predicciones, no hay un soporte teórico bien establecido que justifique qué procedimiento puede arrojar mejores resultados. En ocasiones, combinaciones sencillas como un simple promedio o medidas de tendencia central robustas pueden mostrar mejor comportamiento que criterios más sofisticados considerando pesos óptimos estimados en base a errores en el pasado (Claeskens *et al* 2016 [2]). En general, el criterio apropiado dependerá notablemente de los datos en estudio. Desde un punto de vista práctico, es relevante mencionar la existencia de alguna librería de R donde se implementan diferentes métodos de combinación de predicciones. Por ejemplo, la librería `ForecastComb` (Weiss *et al* 2018 [5]) incluye una batería de procedimientos que van desde predictores cooperativos sencillos como la media o mediana, a predictores cooperativos basados en estimar los pesos: (a) a partir de ajustes de regresión entre las respuestas observadas en el pasado respecto a las correspondientes predicciones individuales, y (b) minimizando el error cuadrático medio de predicción (MSPE) sujeto a una condición de normalización que conduce a procedimientos basados en los autovalores de la matriz de errores MSPE.

4.2. Limitaciones

En la iniciativa de Predicción cooperativa de CEMat, estimar los pesos en base a la precisión los predictores individuales en fechas pasadas es complejo por varios motivos, incluyendo (i) un período de entrenamiento muy corto (particularmente para horizontes elevados) y (ii) numerosos datos faltantes, toda vez que los predictores individuales se han ido incorporando a esta iniciativa escalonadamente en el tiempo. Esto por ejemplo hace inviable considerar predictores combinados como los mencionados basados en regresión al disponer de menos instantes de entrenamiento que predictores.

4.3. Predictores cooperativos

Los predictores cooperativos que se han considerado de inicio han sido cuatro predictores sencillos basados en la media y en medidas robustas de localización. Desde el 10 de abril, con último dato observado de 9 de abril, se incluyen resultados con tres nuevos criterios de combinación. Todos ellos se describen brevemente a continuación.

En adelante, $f_{i,t}$ denota la predicción del predictor individual i-ésimo en el día t para $i = 1, \ldots, npre$. Nótese que como los participantes no tienen porque enviar predicciones para la totalidad de variables, CCAA y horizontes, el número de predictores individuales $npre$ puede variar para cada combinación (variable, CCAA, horizonte).

CP01: Simple Average Todos los predictores reciben el mismo peso:

$$f_{CP01,t+h} = \frac{1}{npre} \sum_{i=1}^{npre} f_{i,t+h}.$$

CP02: Median Proporciona una combinación más robusta a predicciones extremas:

$$f_{CP02,t+h} = \begin{cases} f_{(\frac{npre+1}{2}),t+h} & \text{si } npre \text{ impar} \\ \frac{1}{2}\left(f_{(\frac{npre}{2}),t+h} + f_{(\frac{npre}{2}+1),t+h}\right) & \text{si } npre \text{ par} \end{cases}$$

donde $f_{(i),t+h}$ denota la predicción que ocupa el lugar i-ésimo cuando se ordenan en sentido creciente.

CP03: Trimmed Mean Otra vía robusta que consiste en calcular la media tras eliminar un porcentaje $100\lambda\,\%$, con $0 < \lambda < 1$, de las observaciones más extremas. Si $K = \lambda \cdot npre$:

$$f_{CP03,t+h} = \frac{1}{npre - 2K} \sum_{i=K+1}^{npre-K} f_{(i),t+h}.$$

Se ha considerado $\lambda = 0{,}2$.

CP04: Winsorized Mean El porcentaje $100\lambda\,\%$ de valores más extremos se reemplaza por los valores más extremos del resto de predicciones. Si como antes $K = \lambda \cdot npre$:

$$f_{CP04,t+h} = \frac{1}{npre}\left(K\left(f_{(K+1),t+h} + f_{(npre-K+1),t+h}\right) + \sum_{i=K+1}^{npre-K} f_{(i),t+h}\right).$$

De nuevo se ha considerado $\lambda = 0{,}2$.

CP05: Bates/Granger(mod) Predicciones combinadas con pesos normalizados e inversamente proporcionales al error en las predicciones de días anteriores, siguiendo así el procedimiento propuesto inicialmente por Bates y Granger (1969) pero con diferentes pesos. Específicamente

$$f_{CP05,t+h} = \sum_{i=1}^{nphis} \omega_i f_{i,t+h}, \text{ siendo } \omega_i = \frac{1/\rho_i}{\sum_{i=1}^{nphis} 1/\rho_i},$$

siendo $nphis$ el número de predictores que reportaron predicciones en días anteriores, $f_{i,t-1}, \ldots, f_{i,t-n_i}$, y ρ_i un valor del error promedio cometido por el i-esimo predictor en esos días previos.

Para un día t y una combinación (variable, CCAA, horizonte), los valores ρ_i se calculan como sigue.

1. Sea $ndhis = \max\limits_{1 \le i \le nphis} n_i$, el mayor número de días previos a t en el que se dispone de alguna predicción. Sea $\mathcal{M}$ la matriz de dimensión $nhdis \times nphis$ con las predicciones previas:

$$\mathcal{M} = \begin{pmatrix} f_{1,t-1} & f_{2,t-1} & \cdots & f_{nphis,t-1} \\ f_{1,t-2} & f_{2,t-2} & \cdots & f_{nphis,t-2} \\ \vdots & \vdots & \ddots & \vdots \\ f_{1,t-nhdis} & f_{2,t-nhdis} & \cdots & f_{nphis,t-nhdis} \end{pmatrix}$$

   La primera fila y al menos una columna de $\mathcal{M}$ tendrán todas las predicciones, pero en la mayoría de los casos $\mathcal{M}$ no estará completa.

2. Si y_{t-k} denota el valor real de la serie en estudio en el día $t-k$, entonces la matriz $\mathcal{E} = (e_{ki})$, con $e_{ki} = |y_{t-k} - f_{i,t-k}|$ contiene los errores absolutos de los predictores en el pasado. Los valores faltantes en $\mathcal{E}$ se imputan con $\max\limits_{k,i}\{e_{ki}\}$.

3. A partir de cada columna de $\mathcal{E}$ obtiene el error promedio ρ_i del predictor i-ésimo mediante alguno de los siguientes criterios:

MAE (Mean Absolute Error)	$\rho_i = \dfrac{1}{nhdis} \displaystyle\sum_{k=1}^{nhdis} e_{ki}$
RSME (Root Mean Squared Error)	$\rho_i = \left(\dfrac{1}{nhdis} \displaystyle\sum_{k=1}^{nhdis} e_{ki}^2 \right)^{1/2}$
MAPE (Mean Absolute Percentage Error)	$\rho_i = \dfrac{1}{nhdis} \displaystyle\sum_{k=1}^{nhdis} \dfrac{e_{ki}}{y_{t-k}}$

Los resultados publicados en los informes diarios corresponden al uso de ρ_i =MAE.

CP06: Lowess

CP07: Loess+Bates/Granger(mod)

Referencias

[1] Bates J.M. y Granger C.W.J (1969) The combination of forecasts. *Operations Research Quarterly*, 20:451–468.

[2] Claeskens G., Magnus J.R., Vasnev A.L. y Wang W. (2016) The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762.

[3] Clements M.P., Hendry D.F., Aiolfi M., Capistrán C. y Timmermann A. (2012) *Forecast Combinations*. Oxford University Press.

[4] Timmermann A. (2006) *Forecast combinations*. Handbook of economic forecasting, 1:135–196.

[5] Weiss C.E., Raviv E. y Roetzer G. (2018) Forecast Combinations in R using the `ForecastComb` Package. *The R Journal*, 10(2): 262–281.