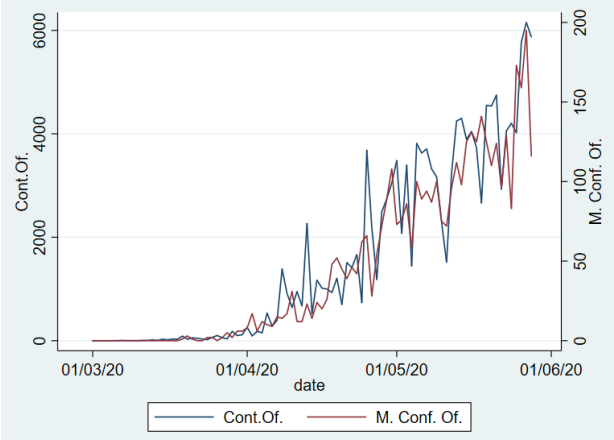
Estrujemos los datos para vencer el COVID-19
El pulpo Paul y los rankings de desempeño en la batalla contra la pandemia

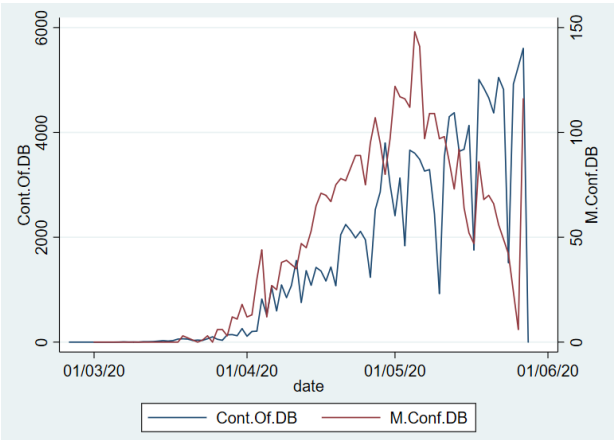
¿Por qué medimos y monitoreamos? La respuesta corta es para saber cómo vamos, y tener claro si es momento de celebrar o corregir. Pero también lo hacemos para aprender, en especial cuando nos enfrentamos a algo menos conocido. Los dos ingredientes clave en un proceso de monitoreo son la medición y la referencia, y ambas deben ser robustas para que de verdad el monitoreo en la toma de mejores decisiones. Cualquiera de ellas que falla, y estamos frente a un clásico caso de "garbage in, garbage out". Si los datos que usamos para evaluar son incorrectos, corremos el serio riesgo de llegar a conclusiones equivocadas y alegrarnos cuando no corresponde, o pensar que lo estamos haciendo pésimo cuando eso no es cierto.

Pensemos en todos los datos que recibimos sobre el Covid-19. Los medios constantemente comparan unos países con otros, hablando de lo bien que lo están haciendo algunos y lo mal que lo están haciendo otros. Sin embargo, estas comparaciones asumen que la información disponible es confiable y refleja la realidad, y eso está lejos de ser cierto, pues hoy sabemos que tanto los contagios como las muertes informadas pueden estar subdimensionadas. Por ejemplo, en el caso de la ciudad de Nueva York[1], se determinó que los contagios reales eran 10 veces los contagios confirmados, y que las muertes reales casi duplicaban a las muertes confirmadas por Covid. Estos cambios tienen un efecto brutal en la "infection fatality rate". Usando sólo los datos de casos confirmados se concluye que un 7.8% de los infectados muere, mientras que con los datos corregidos es un 1.4%, es decir menos de la quinta parte.

Si esto sucede en Estados Unidos, que son los campeones mundiales de las estadísticas y el análisis de datos, ¿qué estará sucediendo en países menos desarrollados? Hoy en día la mayoría de los países han hecho pública esta información, por lo que es fácil hacernos una idea si aplicamos los análisis adecuados.

Tomemos el caso de "Amazonia"[2], un país latinoamericano que ha puesto el detalle de la información sobre casos del Covid19 disponible en Internet. Abajo hay dos gráficos de la evolución de contagios y muertes en el tiempo, ambos obtenidos de fuentes oficiales. El de la izquierda son los datos agregados informados por la autoridad, obtenibles en Wikipedia, y el de la derecha proviene de una base de datos con detalle a nivel de persona, anonimizada. Podemos apreciar que hay varias cosas que no cuadran con las expectativas. Primero, la evolución diaria cambia de un set de datos al otro. Por ejemplo, las muertes diarias (en rojo) del gráfico de la izquierda crecen en forma sostenida, mientras que en el de la derecha primero crecen, caen a pique, y luego vuelven crecer en forma abrupta. Esto, a pesar de que la suma de las muertes en ambos gráficos es la misma.





[1] Ver detalles aquí: https://www.worldometers.info/coronavirus/coronavirus-death-rate/

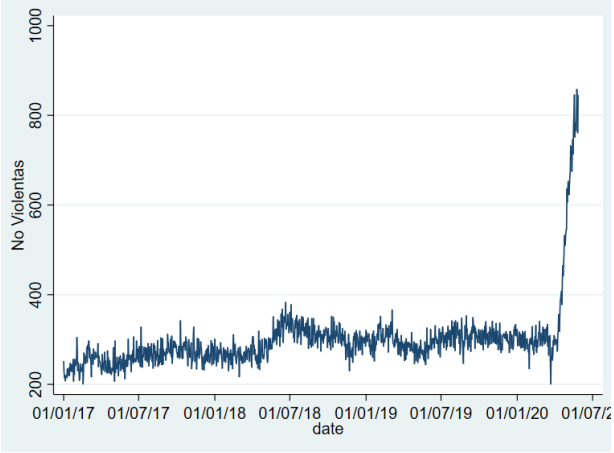
[2] Uso un nombre ficticio pues me interesa focalizar la atención en el método analítico, y no tanto en el país en particular, sin embargo, el lector interesado podrá identificarlo si dedica algunos minutos a googlear. Personalmente, hay muchas cosas que me gustan y admiro de ese país, y tengo amigos allí, por lo que, en esta batalla contra el Covid19, le deseo lo mejor.

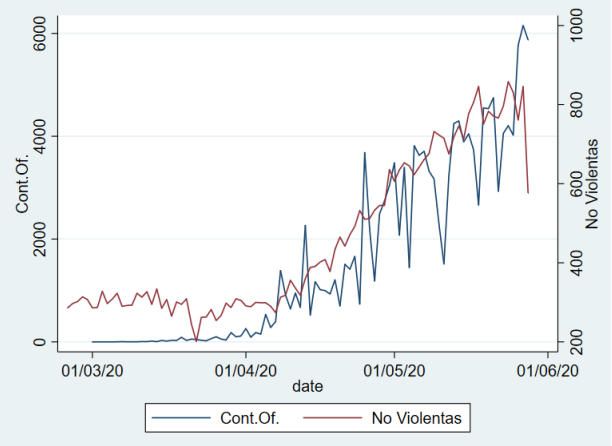
Por Patricio Rojas E., Ph.D., ESE Business School projas.ese@uandes.cl | 31 de mayo de 2020

En cuanto a los contagios, primero difiere la suma de todos los contagios a la fecha, aunque la diferencia es pequeña (2%), y segundo, el gráfico de la izquierda tiene quiebres mucho menos marcados que el de la derecha. Por otro lado, en ambos gráficos no se aprecia el rezago que sabemos que existe entre contagios y muertes, que está en torno a 14 días[3]. De ambos datasets, el de la derecha es el que posee patrones más alejados de lo que indica la lógica.

¿Dados estos patrones, será información confiable? ¿qué podemos hacer para salir de dudas? En el caso de la estimación de muertes, tenemos una muy buena alternativa, el dato de las muertes totales en el país. El registro de fallecidos de un país es un dato bastante confiable, poco susceptible a errores o manipulación, por lo que es buen candidato para estimar la cantidad de fallecidos. Además, en condiciones normales, la cantidad diaria de personas fallecidas en un país es un dato bastante estable, que tiende a equivaler a un porcentaje de la población.

El gráfico de abajo a la izquierda muestra las muertes no violentas diarias en Amazonia desde 2017. Queda claro que se han pegado un tremendo salto en las últimas semanas, con el covid-19 como sospechoso número uno. El gráfico de la derecha compara los contagios confirmados con las muertes no violentas, y vemos que desaparece la variabilidad abrupta que observamos en los gráficos previos para las muertes covid19 confirmadas.





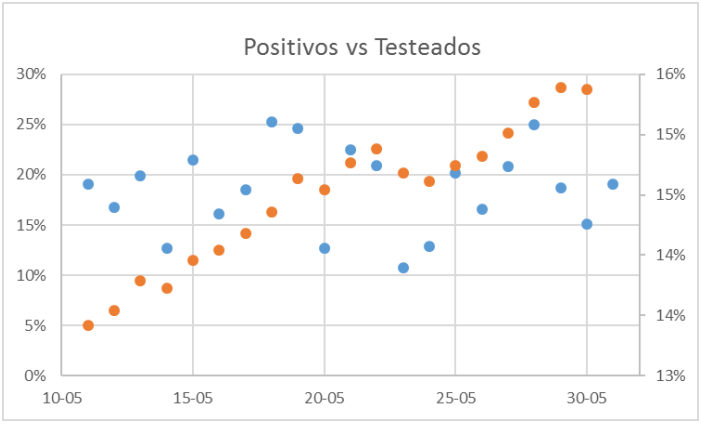
Con estos datos podemos estimar las "muertes en exceso", que es la diferencia entre las muertes reales y las esperables teniendo en cuenta factores estacionales y tendencias macro, que es justamente lo que hice corriendo un modelo multinivel de efectos mixtos. El resultado son 15.835 decesos en vez de los 4.099 oficiales por covid19 para el mismo período. Es decir, las muertes en exceso son 3.9 veces las muertes confirmadas, por lo que hay casi 12 mil muertos no considerados.

¿Por qué se dará esta tremenda diferencia? La explicación más sencilla es que la gente muere y simplemente no se sabe si tenían Covid19 o no. Y desde el punto de vista puramente práctico e inmediato, para la autoridad sanitaria hace más sentido preocuparse de los vivos y dedicar los recursos a testearlos y tratarlos, que precisar de qué fallecieron los muertos. Por muy terrible que suene, cada minuto y gramo de energía dedicado a un muerto es uno no disponible para salvar a los enfermos.

Pasemos ahora a los contagios. Si en Amazonia sucede lo mismo que en otras partes del mundo, entonces con alta probabilidad están sub-dimensionados. ¿Pero en cuánto?

[3] Ver https://onlinelibrary.wiley.com/doi/full/10.1002/jmv.25689?af=R - Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China.

Un método técnicamente adecuado para determinarlo sería tomar una muestra aleatoria de personas, y determinar si tienen anticuerpos Covid19, lo que indicaría que estuvieron infectados y se recuperaron. En el caso de la ciudad de Nueva York, les resultó que un 19.9% de las personas tenían anticuerpos[4]. Para el caso de Amazonia no disponemos de esa información, pero sí sabemos cuántas de las personas testeadas dan positivo. En las últimas tres semanas el valor acumulado ha ido subiendo, pasando de 13.4% a 15.5%, sin embargo, el valor instantáneo se mantiene estable con oscilación, con un promedio de 18.6%.



La gran pregunta es cuál es la base de población a la que habría que aplicar esta tasa, pues los contagios siguen aumentando y nada señala cuál es el estado de avance del proceso de infección. ¿Habría que aplicarlo a la población total del país?, ¿a la urbana?, ¿a la mitad de la urbana? Ni idea.

Pero sí podemos hacer una estimación inversa, usando las Infection Fatalitity rates conocidas, por ejemplo, la estimada para el mundo por WHO, o las que resultaron en China a nivel de ciudad, provincial, o nacional.

	WHO	Wuhan	Hubei	China
IFR	3,4%	4,9%	3,1%	2,1%
Cont. Estimados	465.735	323.163	510.806	754.047
Veces Cont. Oficial	5,3	3,7	5,8	8,5

Por ejemplo, si la tasa en Amazonia es como el promedio chino, para que tengamos hoy 15.835 decesos, hace 14 días habríamos requerido 754.047 contagiados (15.835/2.1%), valor que es 8.5 veces la cantidad de contagiados oficiales en Amazonia a esa fecha, equivalente a 88.498 personas. Es decir, Amazonia podría tener menos grado de error que la ciudad de Nueva York. Nada mal para un país en vías de desarrollo.

Nótese que, durante las últimas dos semanas, en Amazonia los nuevos contagios oficiales han aumentado en 67.130, que equivalen a 570.605 contagios efectivos del ejemplo previo (x8.5). Por lo tanto, en los próximos días deberían tener 11.923 (570.605 x 2.1%) nuevos decesos por Covid-19. Por lo que discutimos previamente, es probable que no se reflejen en la estadística como "muertos por Covid-19", pero sí como "muertes en exceso" en la estadística de muertes no violentas. Un panorama sombrío, y que me alegraría mucho fuese incorrecto.

Luego de tanto número y análisis, el lector se preguntará qué puede sacar en limpio de todos esto. Intentaré compartir algunas reflexiones.

Lo primero, es que las estadísticas del Covid-19 están lejos de ser datos robustos, y por lo tanto tiene muy poco sentido hacer comparaciones entre países con fines de saber quién lo está haciendo mejor o peor[5]. Comparar el desempeño relativo de los países con datos malos, es un poco cómo haber decidido quién ganaba la copa del mundial de fútbol del 2010 usando las predicciones del pulpo Paul, en vez de usar los resultados reales de los partidos.

[4] https://www.governor.ny.gov/news/amid-ongoing-covid-19-pandemic-governor-cuomo-announces-results-completed-antibody-testing, y https://www.worldometers.info/coronavirus/coronavirus-death-rate/

[5] Debo reconocer que yo mismo no tenía plena consciencia de lo variopinto de la calidad de los datos, hasta que quise ir más allá de los titulares y miré el detalle.

Por Patricio Rojas E., Ph.D., ESE Business School projas.ese@uandes.cl | 31 de mayo de 2020

Estrujemos los datos para vencer el COVID-19
El pulpo Paul y los rankings de desempeño en la batalla contra la pandemia

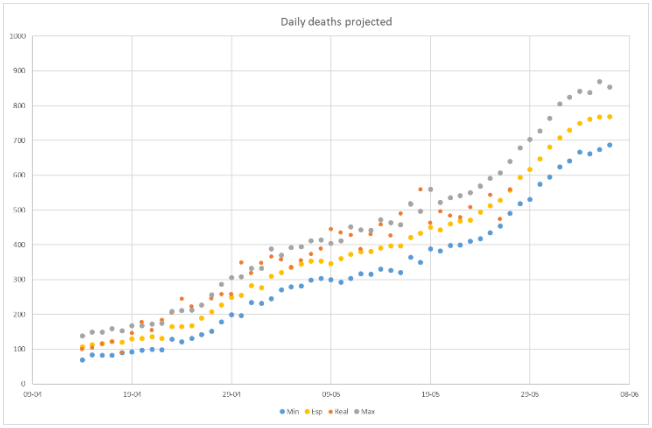
CUADRO: REPORTE DE CORONAVIRUS AL 31 DE MAYO DE 2020

Nº	PAIS	INFECTADOS	FALLECIDOS	RECUPERADOS	ACTIVOS
1	EEUU	1,778,515	104,081	416,461	1,257,973
2	Brasil	498,440	27,878	189,476	281,086
3	Reino Unido	276,156	38,571	1,190	236,395
4	Rusia	405,843	4,693	171,883	229,267
5	India	190,536	5,406	91,621	93,509

Lamentablemente, la opinión pública y los medios de comunicación desconocen lo difícil que es contar con datos robustos para estimar los contagios y muertes, y menos aún conocen las técnicas necesarias para hacer comparaciones menos malas. No es raro entonces que en las redes sociales uno vea tanto a los partidarios como detractores de las autoridades usando los malos datos disponibles para hacer comparaciones similares a las típicas de los campeonatos de fútbol. Pero a diferencia de los goles, que son un dato robusto, los contagiados y muertos por Covid19 están lejos de serlo. Por lo tanto, las comparaciones en los medios y redes sociales son tan buenas como los datos que los alimentan, y por lo tanto son incorrectas por definición.

Sin embargo, lo que sí podemos y deberíamos hacer, es usar los datos para entender mejor el Covid-19, y así combatirlo en forma más efectiva. Cómo se transmite, cómo frenarlo, cómo se viene la mano en términos de necesidad de respiradores, y cuál es la estimación más certera de personas que fallecerán en las próximas semanas. Si lo entendemos mejor, podemos también educar mejor a la ciudadanía.

El gráfico del lado muestra una posible aplicación de estas ideas. Se trata de un modelo de simulación estocástico[6] que predice, para el caso de Amazonia, el "exceso de muertes" a partir de los contagios oficiales. Dada toda la discusión previa, sabemos que el dato de contagios es de calidad dudosa. A pesar de eso, el modelo hace un trabajo razonable proyectando las muertes en exceso[7]. Según esta modelo, dentro de poco Amazonia va a superar las 700 muertes diarias vinculadas a la pandemia.



El Covid-19 es un flagelo terrible, e incluso a los países más ricos y desarrollados los golpeó con mucha fuerza. Si a ellos, con recursos y poblaciones educadas les fue mal, ¿qué podemos hacer el resto? Todo indica que deberíamos poner nuestro mejor esfuerzo en aprender rápido, y para eso es esencial un mejor uso de los datos. Tenemos que estrujarnos para sacarles hasta la última gota de información, y eso no es fácil como intenta mostrar este artículo.

Hoy en día el trabajo más estresante y amargo del mundo es, probablemente, liderar la batalla contra el Covid-19 a nivel de un país. Salvo unas pocas excepciones, hagas lo que hagas te van a llegar palos. De seguro nuestros líderes cometen errores, y es probable que en alguna parte haya gente mejor que ellos para hacer el trabajo. ¿Pero quiénes son? ¿Dónde están? ¿Qué tan rápido los podríamos poner a cargo, sin menoscabar la batalla diaria contra la pandemia? Son preguntas casi sin respuesta, porque requieren acuerdos políticos cuya probabilidad es muy baja. Lo que sí es cierto, es que a todos nos conviene que a la autoridad a cargo le vaya bien. Y quizás hay cosas a nuestro alcance que podemos hacer para colaborar con el bien común. En lo inmediato, desconfiemos de los rankings tipo liga de fútbol, hagamos caso de las medidas preventivas, y ayudemos al que pasa necesidad.

[6] Para detalles del modelo consultar **"covid19sim - Stochastic simulation model for covid19",** por P.Rojas.
[7] Ver anexo 1 por comentario adicional.

Por Patricio Rojas E., Ph.D., ESE Business School projas.ese@uandes.cl | 31 de mayo de 2020

Anexo 1

La información disponible de Amazonia es un tanto voluble. Por ejemplo, bajé los datos de muertes totales desde el 1/1/2017 hasta la fecha en dos ocasiones, con diferencia de horas.

Al comparar los datos noté variaciones, como muestra la tabla de abajo. Por ejemplo, en el año 2018 hubo 22 días con modificaciones, con un efecto neto de 2 muertes más en el año, las no violentas aumentando en 34 y las violentas bajando en 32. El total original de muertes para ese año era 112.8 mil.

Para el 2019 el cambio fue más fuerte, aumentaron las muertes en 1748, con cambios en 69 días. El total original de muertes para ese año era 113.2 mil.

Si bien son cambios muy menores en términos del total, el punto es que uno no esperaría ninguna modificación. Por lo tanto, a pesar de que el dato de decesos oficiales es bastante confiable, estos cambios muestran que los datos numéricos están lejos de una información contundente escrita en piedra.

A pesar de eso, la "ley de los grandes números" nos permite sacarles provecho vía técnicas estadísticas.

	Muertes			Cambios		
	Tot	Vio	N.Vio	Tot	Vio	N.Vio
2017	-1	-94	93	40	40	40
2018	2	-32	34	22	22	22
2019	1784	149	1635	69	69	69
2020	-104	-13	-91	67	67	67

Por Patricio Rojas E., Ph.D., ESE Business School projas.ese@uandes.cl | 31 de mayo de 2020