Inferencia de retrasos en reporte de síntomas

Gonzalo Mena

May 5th, 2020

Abstract

Conocer la verdadera dimensión de una epidemia en tiempo real y a escalas geográficas es fundamental para la correcta estimación de parámetros epidemiológicos, y así mejorar la toma de decisiones. En la escasez de tests, el conteo de casos es inevitablemente una medida sesgada. Se propone aquí considerar retrasos en diagnosticar como variable auxiliar: en este reporte mostramos algunos métodos sencillos para inferir retrasos usando los datos disponibles en el repositorio del ministerio de ciencias. Además, ilustramos el curso de esa variable en las últimas semanas y comentamos cómo podría reportarse como complemento a la incidencia para ayudar a la investigación epidemiológica. En particular, nuestros análisis muestran retrospectivamente la gravedad de la situación en comuna de Quilicura, y sugieren preocupantes situaciones durante la semana epidemiológica 18 en comunas como Angol, Yungay, La Pintana, Maipú, San Bernardo, entre otras.

1 Introducción

En el contexto de la pandemia de COVID-19, es crucial tomar decisiones apropiadas con el fin de proteger poblaciones vulnerables. Estas decisiones están basadas en los resultados de modelos epidemiológicos, los cuales están alimentados de datos. Lamentablemente, sesgos en la recolección de datos pueden tener un gran impacto en la estimación de dichos parámetros, y en consecuencia, en la toma de decisiones [Zhao et al., 2020, Donnat and Holmes, 2020]

El principal insumo de los modelos (y de la toma de decisiones) es el número de casos nuevos en función del tiempo y la ubicación geográfica. Lamentablemente, debido a escasez diferencial de tests, y retrasos en diagnóstico de casos, es posible que esta cantidad entregue una señal débil y sesgada. Una manera de corregir este tipo de artefactos sería, por ejemplo, considerar la cantidad de tests disponibles, o el porcentaje que da positivo. Desafortunadamente, evidencia internacional señala que datos de testeo suelen ser muy ruidosos como para ser útiles [Kaashoek and Santillana, 2020]. Esto ciertamente es el caso en Chile: a pesar de que los datos de tests se publican cada día, la forma de tabulación corrompe la correspondencia entre el lugar de testeo y el origen del caso, dado que muchos tests se

mandan a analizar fuera de la región del caso. Además, el desglose se hace sólo por región, lo cual impide el más fino análisis por comuna.

La alternativa que describmos en este reporte es el uso de retrasos en diagnóstico como medida posiblemente útil para corregir los sesgos de conteo de casos. La motivación es doble: en primer lugar, investigación reciente basada en métodos bayesianos jerárquicos [Stoner et al., 2019] ha mostrado la utilidad de esa variable para la estimación de la incidencia real de una epidemia a escalas geográficas. Segundo, en un nivel más descriptivo, creemos que el hecho de disponer de esa cantidad puede ser útil a la hora de analizar en tiempo real, y tomar medidas.

2 Métodos

Los análisis presentados acá se basan en los reportes de casos nuevos (producto 1) (entendidos como diagnóstico por PCR) por comuna disponibles en los informes epidemiológicos desde el 30 de Marzo (cada 3 o 4 días) y los reportes de casos por fecha de inicio de síntomas (producto 15), disponible sincrónicamente con la aparición cada nuevo informe epidemiológico desde el 15 de Abril. En este informe se muestra, por comuna, el número de casos que presentaron síntomas en cada una de las semanas epidemiológicas. Importantemente, aunque sólo se encuentra disponible los datos correspondientes al último informe, el historial de ediciones de Github permite el accesso a la totalidad de aquellos, desde su primera publicación.

2.1 Retraso promedio

Los datos disponibles permiten estimar, en cada comuna y semana epidemiológica (SE), la cantidad de días que en promedio tarda una persona desde que presenta síntomas hasta que es diagnosticado.

Para tales fines, se propone el siguiente método: para cada una de las SE y comuna, calcular el número de casos (confirmados PCR) que fueron reportados en cada uno de los informes epidemiológicos. Lo anterior da una distribución de cuándo se reportan los casos con aparición de síntoma en una determinada SE. El valor esperado de esa distribución (relativo a la SE en cuestión) corresponde al retraso promedio.

2.1.1 Limitaciones y detalles

Datos previos al 30 de Marzo Existen reportes de aparición de síntomas por comuna en la SE7, tan temprano como mediados de febrero. El primer informe por fecha de inicio de síntomas (15 de Abril) da cuenta de todos aquellos casos, pero el desglose comunal en los informes epidemiológicos sólo está disponible desde el 30 de Marzo. Para extender el análisis en la carencia de los informes se propone un simple método de programación (lineal) dinámica, que se basa en un principio de parsimonia: los datos de los informes por aparición de síntoma deben "colmar" los datos por reporte PCR de la manera más sencilla posible, es decir, los síntomas reportados con mayor antigüedad han de ser atribuidos a informes más antiguos también.

Variabilidad El reporte de síntomas existe con la resolución de SE (similarmente, los informes epidemiológicos muestran la evolución agregada durante 3 o 4 días), lo que implica la imposibilidad de medir retrasos con la precisión absoluta de días. Un modelamiento más consciente de aquello se basaría en la simulación (*bootstrap*) de varios "estados del mundo" correspondientes a la asignación de aparición de síntomas en días particulares, lo que permitiría construir intervalos de confianza no-paramétricos alrededor de los promedios calculados. En este informe no incluimos dichos intervalos pero apelamos al sentido común: retrasos tienen una variabilida de al menos 3 días a cada extremo, y por lo tanto son más útiles en su aspecto comparativo que absoluto (por ejemplo, para detectar atrasos groseros).

Datos insuficientes La estimación de retraso promedio requiere como paso intermedio la estimación de una distribución, un problema no-paramétrico. No es claro en qué medida esto significa que se necesiten muchos datos para converger a la cantidad 'real'. Como medida básica se debe pensar que los casos donde hay pocas observaciones tienen menor confiabilidad. De hecho, anecdóticamente debido a errores de inconsistencia entre informes, alguna veces puede ocurrir que los retrasos sean una cantidad negativa. Eso no proviene de una falla del método pero sí de los errores de tabulación, e ilustran la necesidad de tener criterio a la hora de evaluar los resultados.

Sesgo por datos que aún no llegan La estimación de retraso se hace más precisa a medida que pasa el tiempo, debido a que muchos casos de inicio de síntoma en una determinada SE serán reportados en las semanas siguientes. Más aún, mientras mayor sea el retraso mayor será ese sesgo que tenderá a subestimar el retraso. Para análisis basados en información más reciente se recomienda la siguiente medida

2.2 Temporalidad

También consideramos la temporalidad (*timeliness*) de los reportes, que definimos como la proporción de los casos reportados durante una SE que presentaron síntomas dentro de esa misma semana. Para el cálculo de ésta se require clasificar cada informe epidemiológico dentro de una SE. Nosotros usamos la convención de que el reporte con fecha R corresponde a la SE S si la mayor parte de los días entre los reportes $R-1$ y R estuvieron en la semana S. Esto naturalmente lleva a un artefacto de *aliasing*, ya que reportes correspondientes a los primeros días de cada semana muy probablemente contienen casos con síntomas correspondientes a la semana previa, sin que ésto signifique necesariamente un retraso. Nuevamente, bajo este entendimiento, aún es posible usar la temporalidad para hacer comparaciones entre comunas y SE.

3 Resultados

En esta sección comentamos dos figuras que utilizan los métodos descritos en este informe. La figura 1 muestra el retraso (en días) en las semanas

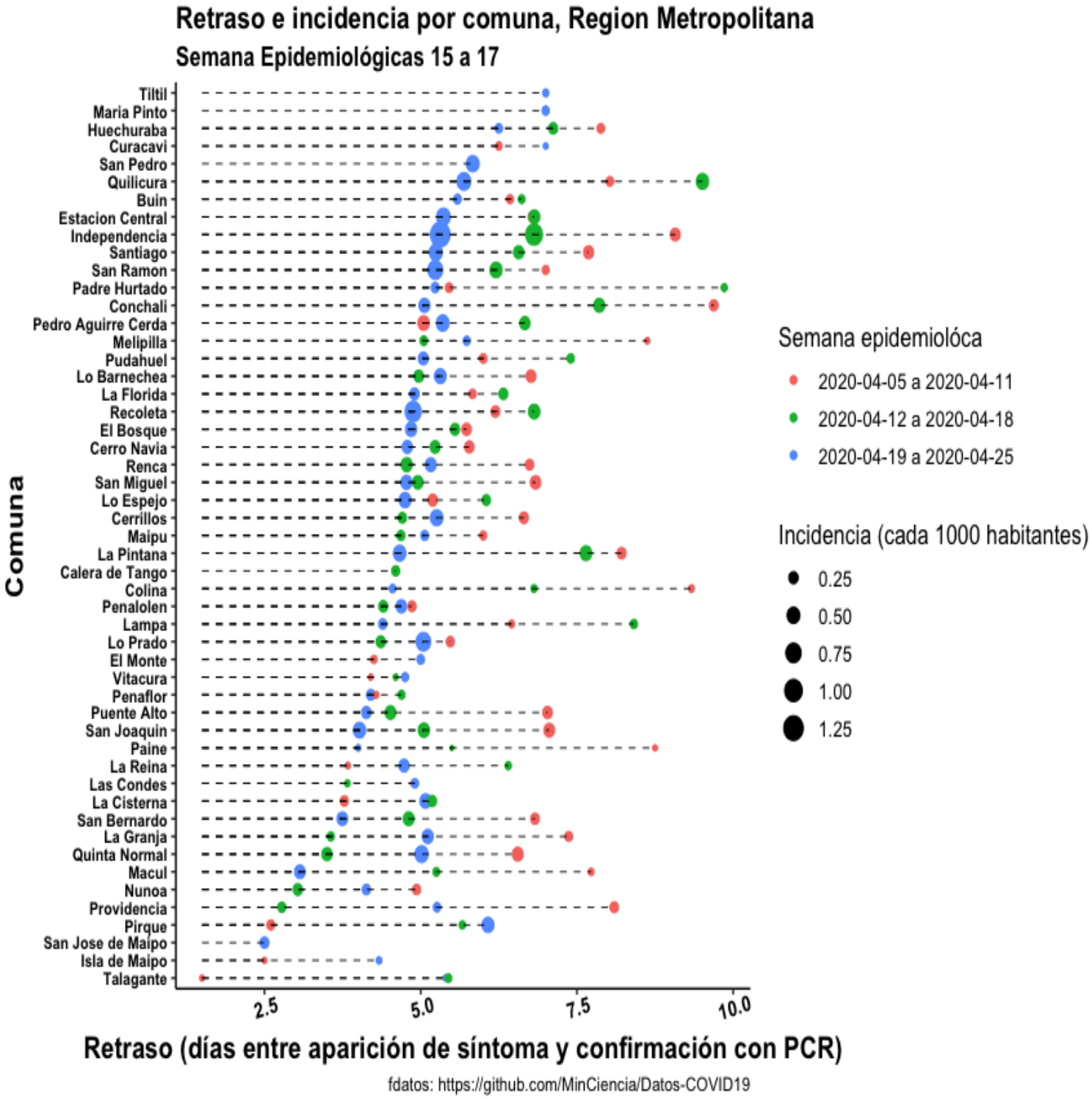

Figure 1: Graficar en función de la comuna y SE en conjunción con la incidencia puede ser útil para analizar retrospectivamente situaciones problemáticas

epidemiológicas 15, 16 y 17, así como la incidencia en dichas comunas durante esas semanas (por fecha de inicio de síntomas).

Este gráfico permite detectar **retrospectivamente** lugares donde ha habido retrasos importantes, y verificar que se han traducido en aumento de incidencia (reportada) en las semanas siguientes. Por ejemplo, la gravedad de la situación de Quilicura queda en evidencia: el retraso aumentó considerablemente entre la SE 15 y SE 16 (donde se observa el máximo regional), como también lo hizo la incidencia semanal. De manera similar, se observan situaciones preocupantes en las comunas de Pedro Aguirre Cerda, Independencia, Conchalí y lo Prado (transición de la SE16 a la SE17).

La figura 2 muestra la incidencia durante la SE18, donde también se muestran los casos diagnosticados durante esa misma semana. Al encontrarse en escala logaritmica, la distancia entre los dos círculos correspondientes a la misma comuna da una medida directa de la temporalidad, que permite comparaciones. Dentro de las comunas con alta incidencia registrada las situaciones de Quilicura y La Pintana destacan por su nivel de retraso. Se observan también importantes retrasos en Pirque, San Bernardo y Maipú, los cuales podrían estar evidenciando un brote para la cual existe una insuficiente capacidad de diagnóstico oportuno.

Finalmente, la figura 3 es análoga a la figura 2 pero a nivel nacional. Del punto de vista de reportes, se observan situaciones preocupantes en Comunas de Angol, Yungay y muchas otras. Estas situaciones podrían ser investigadas por expertos a nivel regional.

4 Agradecimientos

Agradecemos a Joaquín Fontbona, Alejandro Maass, Pamela Martinez, Alejandra Avalos, Mauricio Vargas, Orlando Rivera, Jorge Prado, Mauricio Santillana, Fred Lu, Oliver Stoner, Theo Economou, Jorge Pérez y Cristóbal Cuadrado por sus valiosos comentarios.

References

[Donnat and Holmes, 2020] Donnat, C. and Holmes, S. (2020). Modeling the heterogeneity in covid-19's reproductive number and its impact on predictive scenarios. *arXiv preprint arXiv:2004.05272*.

[Kaashoek and Santillana, 2020] Kaashoek, J. and Santillana, M. (2020). Covid-19 positive cases, evidence on the time evolution of the epidemic or an indicator of local testing capabilities? a case study in the united states. *A Case Study in the United States (April 10, 2020)*.

[Stoner et al., 2019] Stoner, O., Economou, T., and Drummond Marques da Silva, G. (2019). A hierarchical framework for correcting underreporting in count data. *Journal of the American Statistical Association*, pages 1–17.

[Zhao et al., 2020] Zhao, Q., Ju, N., and Bacallado, S. (2020). Bets: The dangers of selection bias in early analyses of the coronavirus disease (covid-19) pandemic. *arXiv preprint arXiv:2004.07743*.

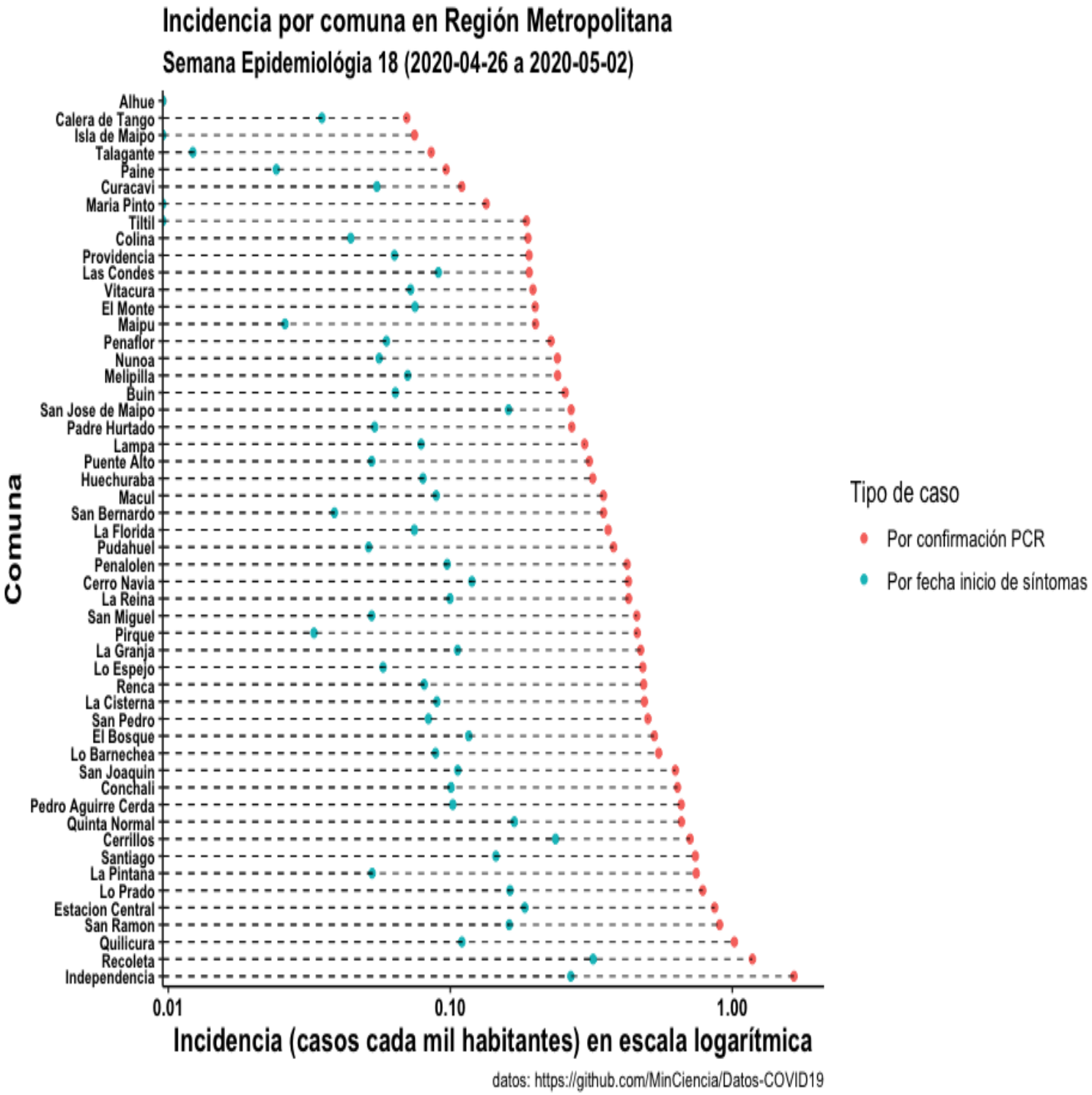

Figure 2: La distancia entre círculos representa la temporalidad, y ayuda a detectar retrasos en reportes

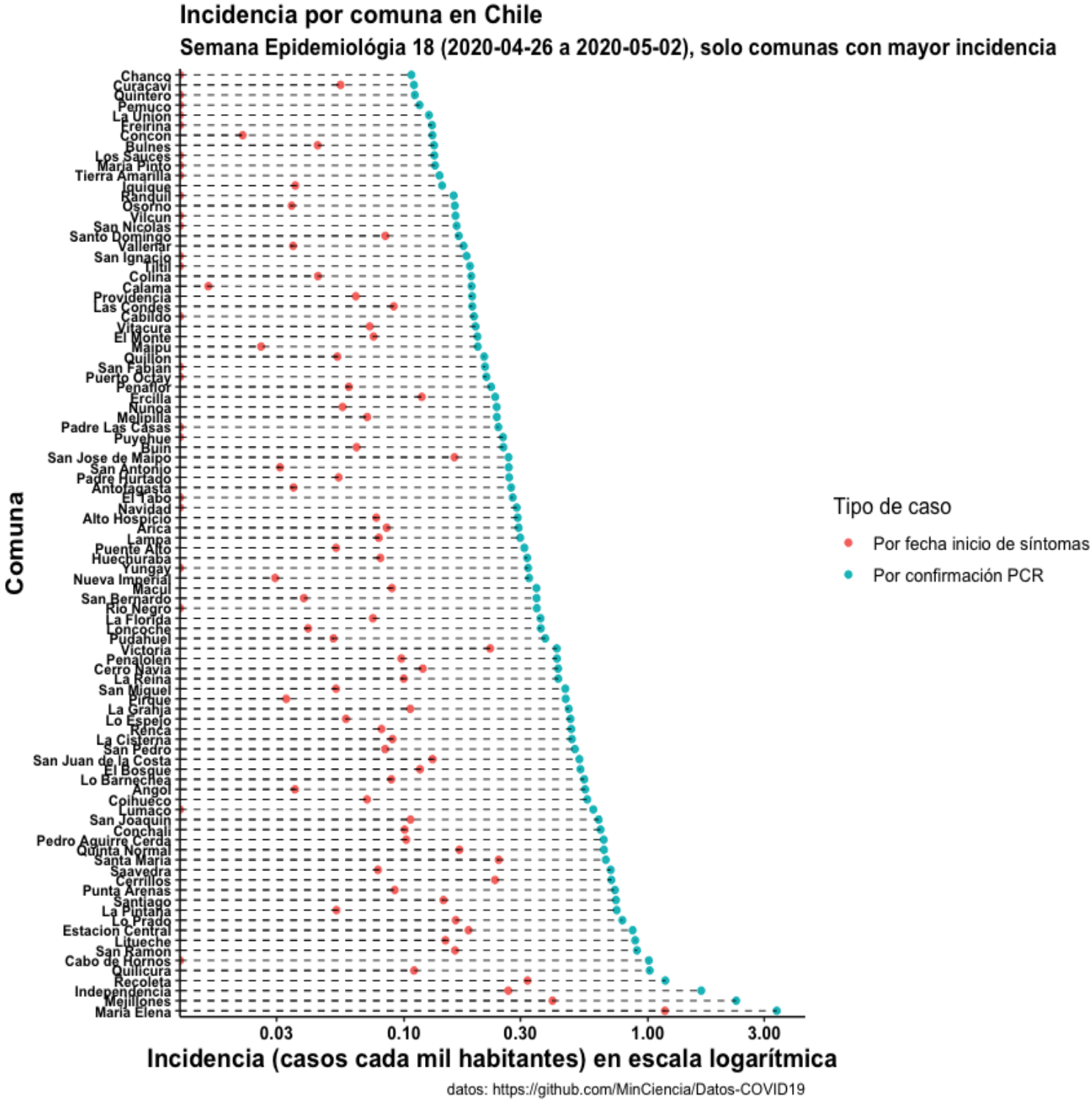

Figure 3: La distancia entre círculos representa la temporalidad, y ayuda a detectar retrasos en reportes