Lenguaje de programación y testeo estadístico: El caso de Ventanas

Vera Sativa*

September 3, 2019

Abstract

Utilizando el lenguaje de programación Python, tras unificar los registros anuales de defunciones en Chile 1998-2016 (~1.7M), analizamos los diagnósticos primarios en defunciones de menores hasta 16 años, comparando la zona crítica bajo la contaminación del complejo industrial Quintero-Ventanas, contra el resto de Chile como control. Encontramos incidencias de malformaciones congénitas, deformidades y anomalías cromosómicas (CIE-10: Q00-Q99), 3.04 a 3.75 desviaciones estándar sobre el resto del país, con P-values de 0.0001 a 0.00002 en un millón de simulaciones, estimando un impacto de entre 29.73 a 37.8 muertes de menores en la zona crítica por sobre la norma nacional. La metodología podría ser escalada a todo el país para detectar focos de contaminación desconocidos.

Una comparativa entre la zona crítica y el resto de Chile

Introducción

Mediante programación en *Python* fue posible estandarizar los registros de defunciones oficiales, que tienen una serie de variaciones año a año. Una vez construido un *registro unificado de defunciones en Chile, entre el año 1998 y 2016*[1], surge la pregunta general: ¿Se podrán observar en éste, rasgos distintos en una zona crítica al resto de Chile? Utilizando el mismo lenguaje de programación testearemos esa hipótesis.

Un aporte clave es la metodología, y por esta razón se disponibiliza todo el *código fuente*[2], posibilitando la implementación y extensión de ésta en otros problemas, territorios o datasets.

Antecedentes

El impacto ambiental y sobre la salud humana del complejo industrial Quintero-Ventanas ha sido ampliamente documentado al punto de que recientemente el Colegio Medico chileno dedicó un número completo de su revista de salud pública *Cuadernos Médico Sociales* al problema[3]

El reciente informe *The Lancet Commission on Pollution and Health* señala que la polución del aire puede ser vinculada al aumento de nacimientos prematuros y con bajo peso. Y que algunos estudios han mostrado asociación entre polución ambiental del aire y aumento del riesgo de síndrome de muerte súbita del lactante[4].

Comparación de diagnósticos primarios

Para buscar una respuesta a la asociación entre contaminación ambiental y enfermedades de la gestación, es posible usar con todas sus limitaciones los registros de defunciones oficiales del país. Una forma de avanzar por sobre esas restricciones es hacer uso de *la integración jerárquica de los códigos de diagnóstico CIE-10*[5] en el dataset. Una comparación de éstos se presenta como la opción más evidente y atractiva, mediante un trabajo de programación en Phyton. Este trabajo desarrolla una metodología de programación orientada al objeto de estudio y pone a Phyton como una alternativa para su uso en salud pública.

*Corresponding author – hola@verasativa.com

Limitaciones de los datos

Para entender que podemos investigar desde estos datos, debemos reconocer sus limitaciones. Dado que este dataset solo incluye las defunciones y no contiene información sobre la población general, no es posible hacer un análisis respecto a tasas de ocurrencia sin tener que usar datos como censos. La ruralidad de la zona, conjugada con la migración campo-ciudad, produce un movimiento poblacional que en ese período aumenta la incertidumbre de las cifras. Tampoco podemos hacer un análisis sobre la distribución etaria de la mortalidad, ni la distribución de diagnósticos primarios en la población general, sin normalizar primero con datos adicionales.

Pregunta de investigación

Con esas limitaciones en mente, podemos plantear una pregunta sencilla, pero contestable:

¿Cómo se comparan los diagnósticos primarios de defunciones, en la zona de interés, con respecto al resto del país en menores de 16 años?

Proceso exploratorio: definiendo la "zona de interés"

Inicialmente, se exploró como zona de interés solamente las comunas de Quintero y Puchuncaví, puesto que son colindantes (Quintero) o el lugar mismo del foco industrial de contaminación .

```
Zona de interés: Quintero, Puchuncaví
Total defunciones en el grupo de interés: 119
Total defunciones en los 10 principales diagnósticos primarios del grupo de interés: 118
Fracción del total: 0.992
```

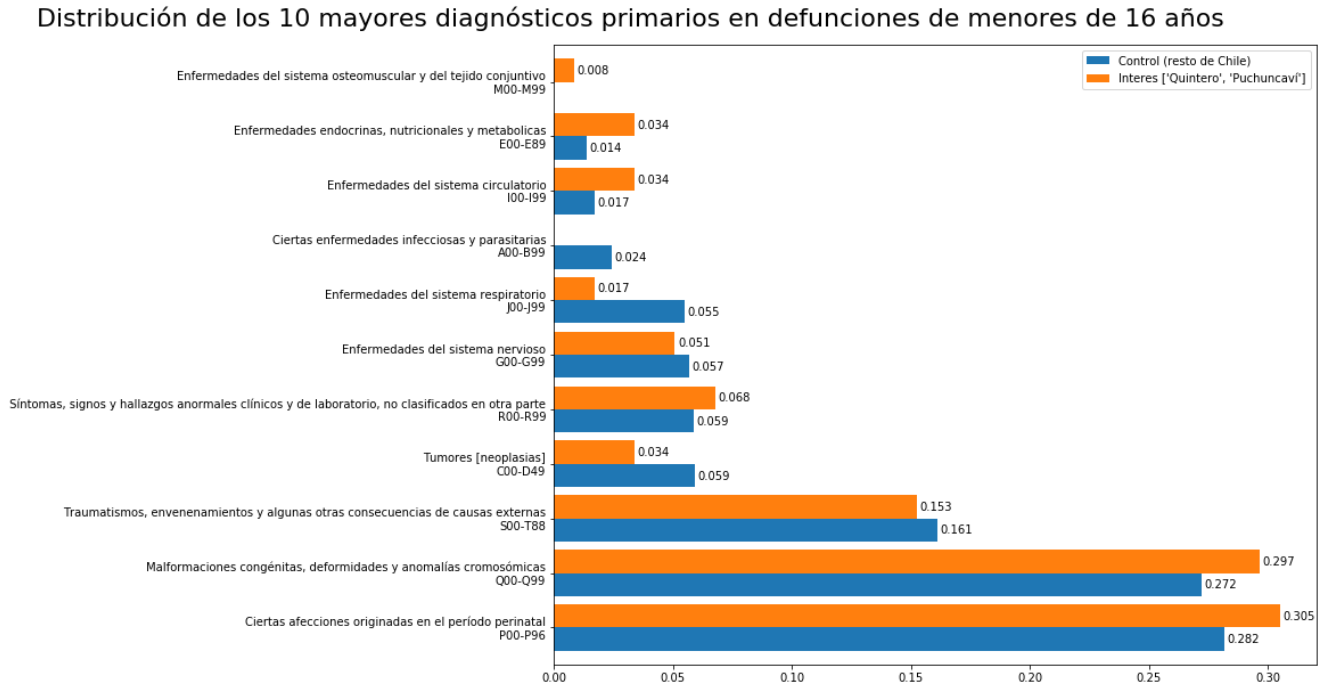

Figura 1: Distribución de los 10 mayores diagnósticos primarios en defunciones de menores hasta 16 años (Quintero-Puchuncaví)

Sin embargo, no se econtró algo claro ya que podemos observar en la figura 1 que hay dos diagnósticos primarios que presentan incidencias superiores a la nacional. Con la intención de buscar una tendencia más clara y validable (tamaño de la muestra), exploramos la incidencia de estos dos diagnósticos primarios en todas las comunas de la Quinta Región.

Ciertas afecciones originadas en el período perinatal (P00-P96)

#	Comuna	Incidencia comuna	Incidencia otros	Proporción comuna	Proporción otros	Total comuna	Total otros
0	La Cruz	26	15957	0.509804	0.274836	51	58060
1	Rinconada	15	15968	0.46875	0.274936	32	58079
2	San Felipe	109	15874	0.37457	0.274542	291	57820
3	Olmué	17	15966	0.369565	0.274968	46	58065
4	El Tabo	7	15976	0.368421	0.275012	19	58092
5	San Esteban	15	15968	0.357143	0.274983	42	58069
6	Cartagena	18	15965	0.339623	0.274984	53	58058
7	Llaillay	28	15955	0.337349	0.274953	83	58028
8	Calle Larga	13	15970	0.333333	0.275003	39	58072
9	Algarrobo	11	15972	0.323529	0.275014	34	58077
10	Quintero	23	15960	0.315068	0.274992	73	58038
11	Hijuelas	23	15960	0.310811	0.274997	74	58037

Tabla 1: Incidencias comunales y nacionales de P00-P96

Malformaciones congénitas, deformidades y anomalías cromosómicas (Q00-Q99)

#	Comuna	Incidencia comuna	Incidencia otros	Proporción comuna	Proporción otros	Total comuna	Total otros
0	Puchuncaví	20	15428	0.434783	0.265702	46	58065
1	Zapallar	6	15442	0.428571	0.265797	14	58097
2	Papudo	3	15445	0.375	0.265821	8	58103
3	La Ligua	37	15411	0.37	0.265657	100	58011
4	Concón	30	15418	0.352941	0.265708	85	58026
5	Nogales	20	15428	0.333333	0.265766	60	58051
6	Cabildo	20	15428	0.333333	0.265766	60	58051
7	Putaendo	19	15429	0.322034	0.265779	59	58052
8	El Tabo	6	15442	0.315789	0.26582	19	58092
9	Quillota	78	15370	0.3083	0.26565	253	57858
10	Calle Larga	12	15436	0.307692	0.265808	39	58072
11	Viña del Mar	287	15161	0.304671	0.265196	942	57169

Tabla 2: Incidencias comunales y nacionales de Q00-Q99

Al observar las tablas de incidencia, podemos notar que el diagnóstico primario *Ciertas afecciones originadas en el período perinatal*, está dominado por otras comunas, no primariamente las de la zona de interés. Y por otra parte, el diagnóstico primario de *Malformaciones congénitas, deformidades y anomalías cromosómicas* domina en comunas colindantes al centro industrial.

Ante esto, decidimos graficar la incidencia de interés en el mapa.

Se observa una distribución no-radial: ¿Cómo se explica?

En la figura 2 se puede apreciar que las comunas más afectadas parecen ser las directamente al norte (Puchuncaví, Zapallar, Papudo y La Ligua), al este (Nogales, La Calera y Quillota) e incluso directamente al sur (Concón) del

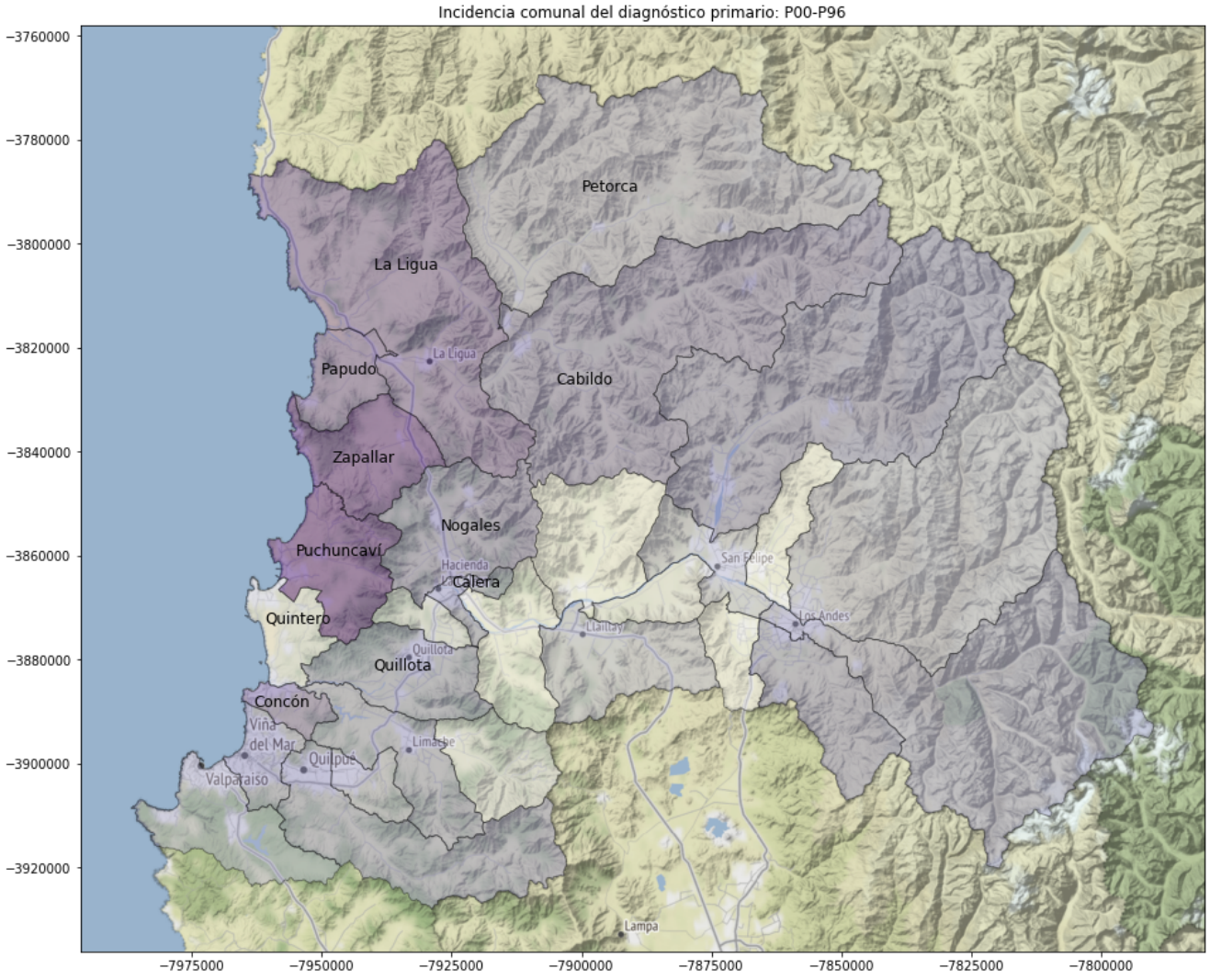

Figura 2: Mapa de incidencias comunales

complejo industrial.

Al observar esta distribución no radial, la investigación parecía no tener sentido y se estancó durante un tiempo. Se sospechaba de un patrón de vientos. Para seguir avanzando, fue necesario que se nos refiriera a la investigación de Patricio Cornejo, Juan López y Sergio Romano 1983[6], donde se creó un mapa que indica la dispersión de contaminación desde el complejo industrial Quintero-Ventanas.

Al sobreponer ese mapa sobre nuestras incidencias (Figura 3), observamos una coincidencia interesante, que nos llevó a continuar.

Diferentes zonas de interés

Debido a que el mapa y la incidencia de Q00-Q99 parecieran indicar en la dirección opuesta de la ciudad de Quintero, y la sospecha de que Concón tenga su propia fuente de contaminación (Refinería de petróleo ENAP), analizamos tres grupos en paralelo: El primero excluyendo Concón y Quintero, el segundo incluyendo ambas comunas, y el tercero incluyendo Concón y excluyendo Quintero.

A continuación, graficamos la incidencia del diagnóstico primario en estos grupos e imprimimos algunos indicadores de representatividad como tamaño del grupo, y cuántos de sus diagnósticos primarios están entre los 10 principales que graficamos.

```
Grupo 1
Zona de interés: Puchuncaví, Zapallar, Papudo, La Ligua, Petorca, Cabildo, Nogales
Total defunciones en el grupo de interés: 313
Total defunciones en los 10 principales diagnósticos primarios del grupo de interés: 308
Fracción del total: 0.984

Grupo 2
Zona de interés: Puchuncaví, Zapallar, Papudo, La Ligua, Petorca, Cabildo, Nogales, Concón, Quintero
Total defunciones en el grupo de interés: 471
Total defunciones en los 10 principales diagnósticos primarios del grupo de interés: 462
Fracción del total: 0.981

Grupo 3
Zona de interés: Puchuncaví, Zapallar, Papudo, La Ligua, Petorca, Cabildo, Nogales, Concón
Total defunciones en el grupo de interés: 398
Total defunciones en los 10 principales diagnósticos primarios del grupo de interés: 390
Fracción del total: 0.980
```

Observaciones

Al comparar estos gráficos, inmediatamente notamos que el diagnóstico primario *Malformaciones congénitas, deformidades y anomalías cromosómicas* (CIE-10: Q00-Q99), es considerablemente más alto en los grupos de interés que en el resto del país como grupo de control (36.7%, 34.2% y 36.7% por sobre 27.2%).

Validación

Para validar estas observaciones realizamos una prueba de permutación:

Por cada grupo, tomamos un millón de muestras al azar del mismo tamaño que el grupo de interés (308, 471 y 398) desde el grupo de control, y observaremos la distribución del diagnóstico primario de interés en estas muestras, en contraste con casos seleccionados por zona geografía de interés, a fin de responder:

¿Qué tan probable es observar las incidencias (36.7%, 34.2% y 36.7%) que se dan en nuestros grupos de interés en cualquier otro grupo del mismo tamaño muestreado al azar desde el grupo de control?

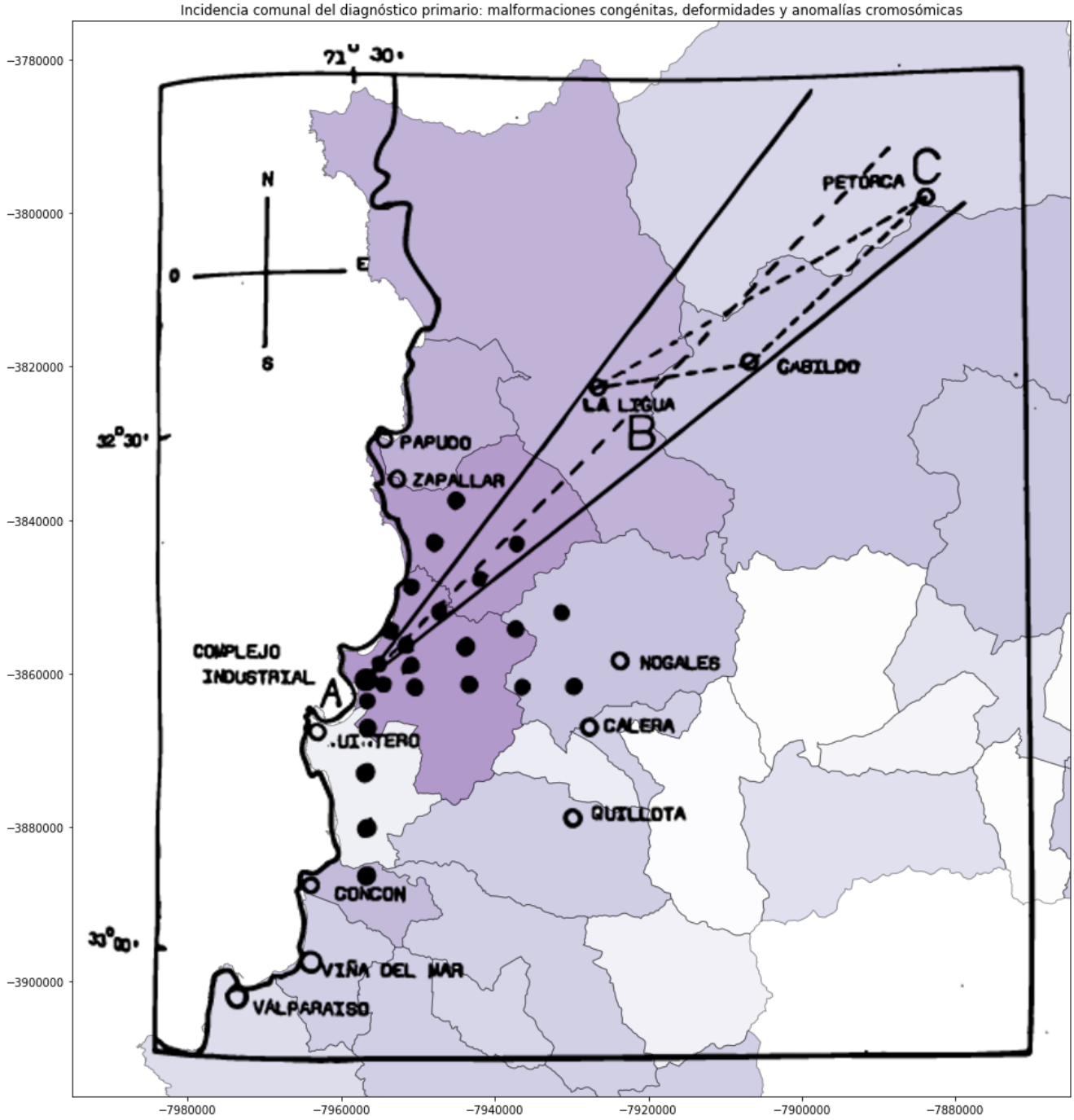

Figura 3: Mapa de incidencias comunales con modelo pluma (Cornejo, López y Romano 1983) sobrepuesto

Grupo 1

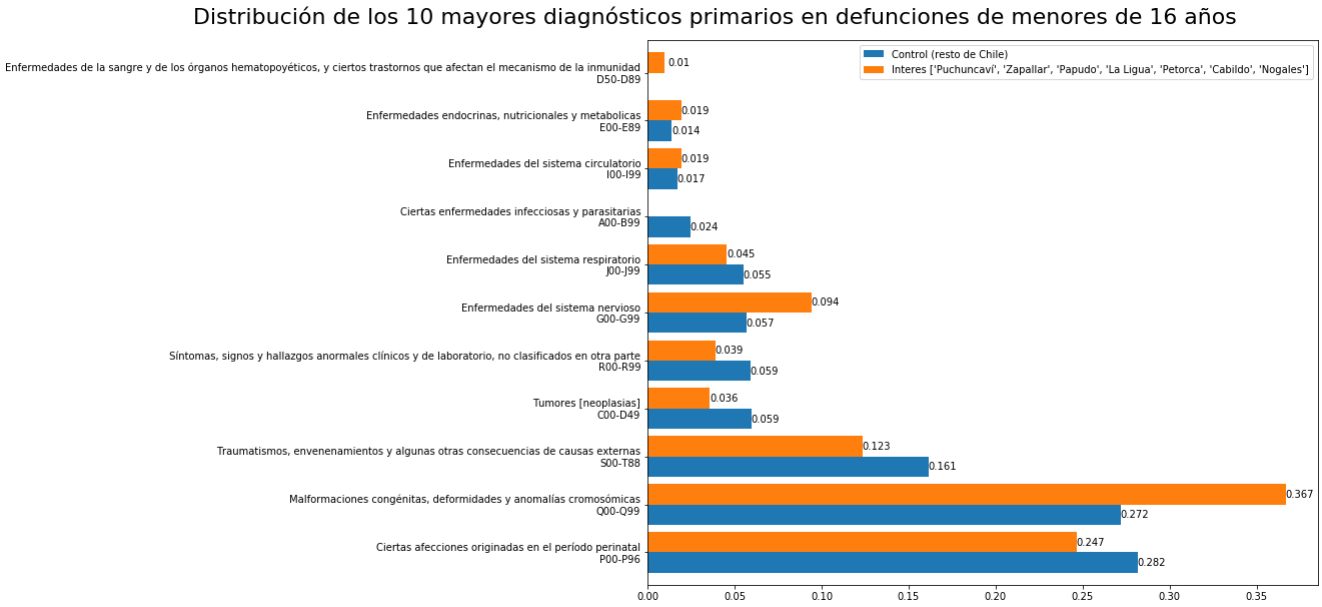

Figura 4: Distribución de los 10 mayores diagnósticos primarios en defunciones de menores hasta 16 años (Puchuncaví-Zapallar-Papudo-La Ligua-Petorca-Cabildo-Nogales)

Grupo 2

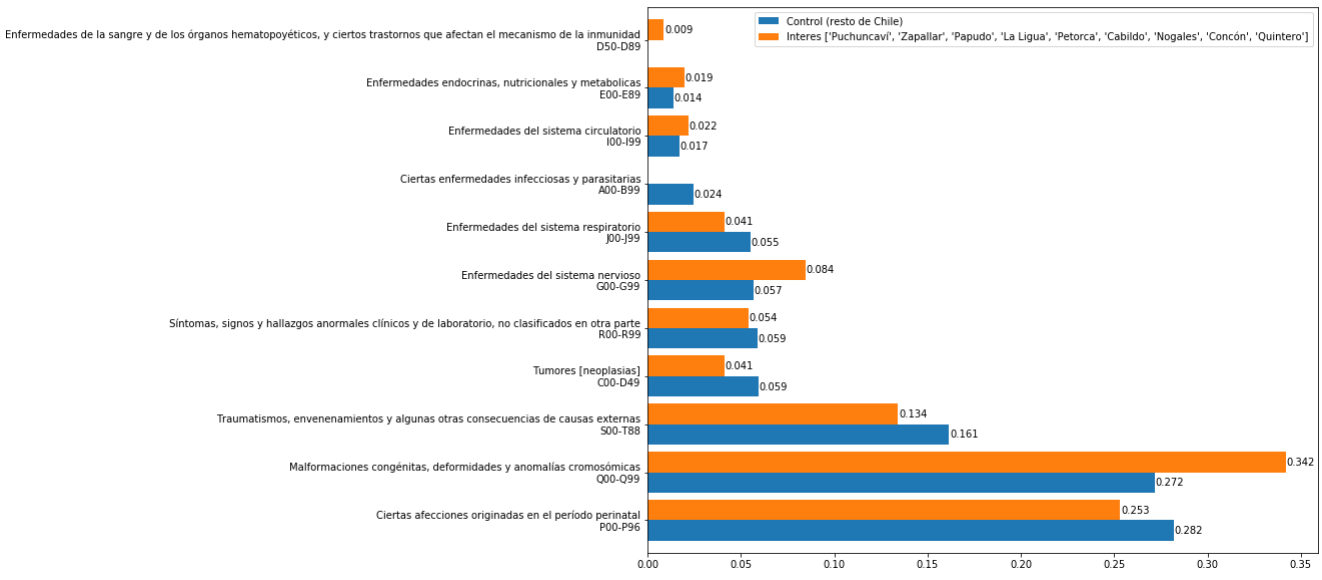

Figura 5: Distribución de los 10 mayores diagnósticos primarios en defunciones de menores hasta 16 años Puchuncaví-Zapallar-Papudo-La Ligua-Petorca-Cabildo-Nogales-Concón-Quintero)

Grupo 3

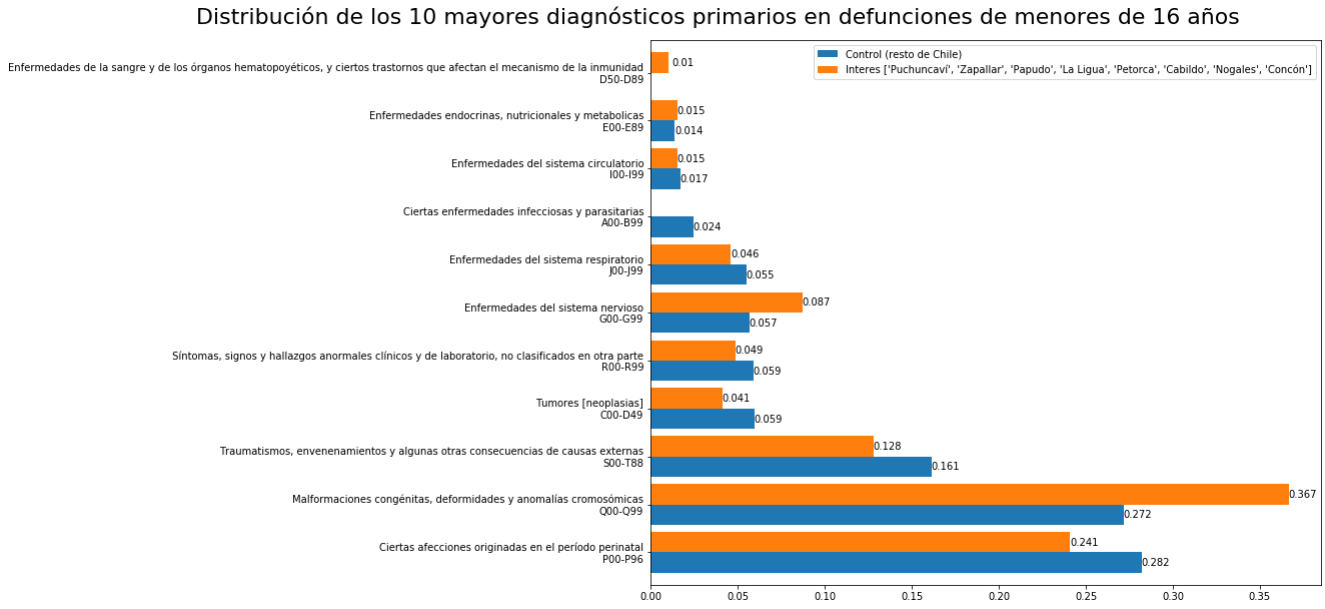

Figura 6: Distribución de los 10 mayores diagnósticos primarios en defunciones de menores hasta 16 años (Puchuncaví-Zapallar-Papudo-La Ligua-Petorca-Cabildo-Nogales-Concón)

Otros valores de interés

En la figura 7 podemos ver la ubicación de las incidencias observadas (línea naranja) con la distribución en el grupo de control del diagnóstico primario *malformaciones congénitas, deformidades y anomalías cromosómicas*.

A continuación cuantificamos la observación anterior con los siguientes números:

- Probabilidad de obtener este resultado, o más, al azar (p-value)
- Incidencia promedio en muestras al azar desde el grupo de control
- Desviación estándar de muestras al azar desde el grupo de control
- Cuantificación en desviaciones estándar respecto del grupo de control, desde los grupos en el promedio de un millón de re-muestreos

```
Grupo 1
['Puchuncaví', 'Zapallar', 'Papudo', 'La Ligua', 'Petorca', 'Cabildo', 'Nogales']
P-value: 0.00010
Promedio de las muestras: 0.27083
Desviación standard de las muestras: 0.02528
Distancia entre el promedio de las muestras y el grupo de interés en desviaciones standard: 3.05

Grupo 2
['Puchuncaví', 'Zapallar', 'Papudo', 'La Ligua', 'Petorca', 'Cabildo', 'Nogales', 'Concón', 'Quintero']
P-value: 0.00039
Promedio de las muestras: 0.27103
Desviación standard de las muestras: 0.02059
Distancia entre el promedio de las muestras y el grupo de interés en desviaciones standard: 3.74

Grupo 3
['Puchuncaví', 'Zapallar', 'Papudo', 'La Ligua', 'Petorca', 'Cabildo', 'Nogales', 'Concón']
P-value: 0.00002
Promedio de las muestras: 0.27085
Desviación standard de las muestras: 0.02239
Distancia entre el promedio de las muestras y el grupo de interés en desviaciones standard: 3.45
```

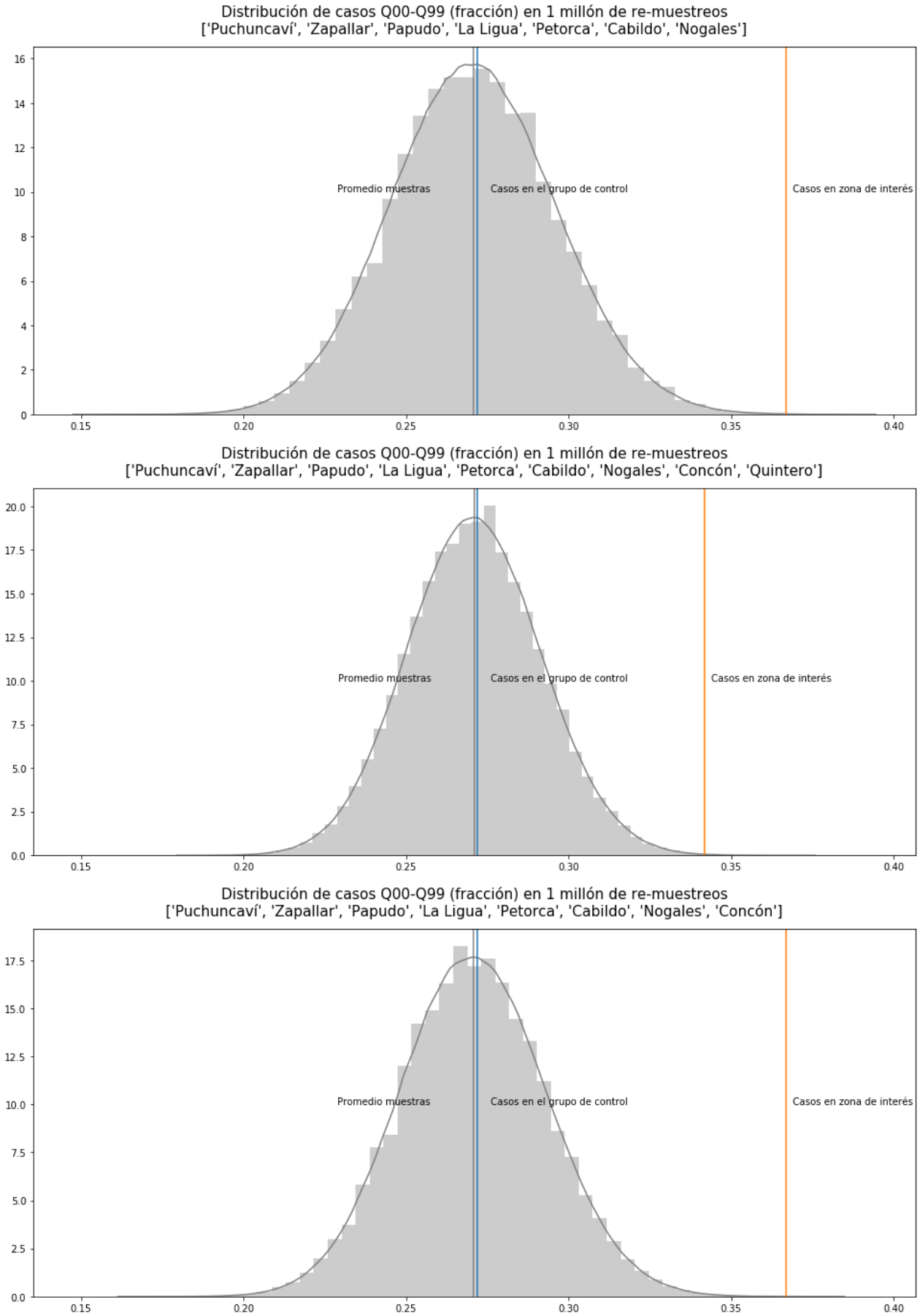

Figura 7: Distribución fracción de casos Q00-Q99 en 1 millón de re-muestreos

Variaciones en el P-value

A razón de haber observado variaciones en el primer dígito no-cero del p-value durante las primeras ejecuciones de 10.000 re-muestreos, se aumento la cantidad de re-muestreos en dos órdenes de magnitud (a un millón). Y para entender como se comporta este p-value respecto a la cantidad de re-muestreos, tomamos sub-muestras del millón de muestras, incrementando su tamaño iterativamente en 500 re-muestreos. Al graficar el p-value en estos distintos tamaños de re-muestreos, se observa que en el n inicial de 10.000 el p-value se lograba estabilizar en su orden de magnitud, pero con un millón se estabilizaba considerablemente más.

Conclusiones

Grupo	Distancia DS	P-Value	Comunas
1	3.05	0.00010	[‘Puchuncaví’, ‘Zapallar’, ‘Papudo’, ‘La Ligua’, ‘Petorca’, ‘Cabildo’, ‘Nogales’]
2	3.74	0.00039	[‘Puchuncaví’, ‘Zapallar’, ‘Papudo’, ‘La Ligua’, ‘Petorca’, ‘Cabildo’, ‘Nogales’, ‘Concón’, ‘Quintero’]
3	3.45	0.00002	[‘Puchuncaví’, ‘Zapallar’, ‘Papudo’, ‘La Ligua’, ‘Petorca’, ‘Cabildo’, ‘Nogales’, ‘Concón’]

Tabla 3: desviaciones estándar por sobre el resto del país de los grupos y su p-value

Tales distancias (3.05, 3.74 y 3.45 desviaciones estándar) entre los valores observados y los promedios del grupo de control (figura 7), así como los p-values observados en el millón de re-muestreos por grupo y su estabilidad observada (figura 8), muestran **una cifra de mortalidad anómala en la zona en estudio**.

Si le restamos la incidencia nacional esperada *(0.272 * 313, 0.272 * 471, 0.272 * 398)* a los grupos de análisis *(0.367 * 313, 0.342 * 471, 0.367 * 398)* podremos estimar que **estamos observando 29.73, 32.97 o 37.8 muertes de menores hasta 16 años en las zonas analizadas, que no observaríamos en el resto de Chile a igual tamaño de muestra**, en el periodo 1998-2016.

Se recomienda enfáticamente seguir observando estos números mientras la fuente de contaminación siga ahí, y durante dos a tres décadas después de que el complejo industrial sea clausurado y la zona, descontaminada.

Se invita a los expertos de las áreas relevantes (salud, bioquímica, ecología, etc.) a investigar la rutas específicas que llevarían al incremento de las defunciones bajo este diagnóstico primario. Se invita, además, a los gobernantes a hacer la prueba de campo, clausurando las fuentes y descontaminando el área, para observar, en algunas décadas, la evolución de la incidencia de este diagnóstico primario en las defunciones de la zona.

Potencial futuro de la metodología

Esta técnica puede ser escalada a nivel nacional para buscar otros fenómenos del mismo tipo, sin especificar una zona en particular, lo que podría revelar problemas de salud pública fuera del “radar” de los investigadores. Para esto se requeriría construir un graph con comunas como nodos, y sus colindacias geográficas como vértices (tal vez con los vectores de comunas de la biblioteca del congreso[1]), e iterar sobre grupos de comunas colindantes con un mínimo de registros totales. De realizarse, se sugiere nombrar *Perico* a tal algoritmo que *treparía por Chile*.

[1] https://www.bcn.cl/siit/mapas_vectoriales/index_html

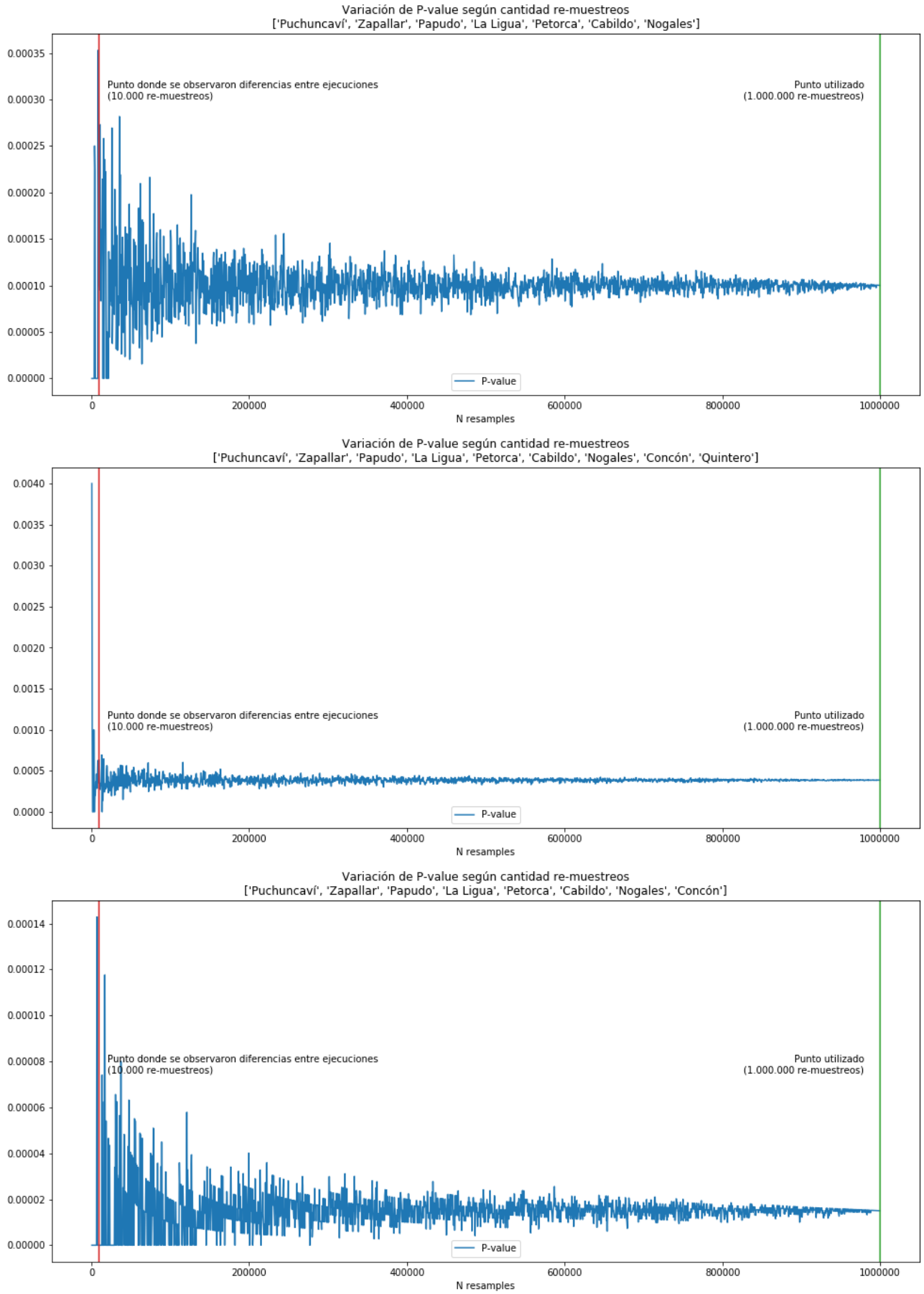

Figura 8: Variaciones en P-values sobre cantidad de simulaciones

Agradecimientos

Al Dr. Yuri Carvajal por la orientación bibliográfica, corrección e implementación de LaTeX. Y a la biblioteca Gabriela Mistral de Ñuñoa por haber sido mi "universidad anfitriona" durante los meses de esta investigación.

Referencias

1. https://github.com/verasativa/defunciones-decoder

2. https://github.com/verasativa/zonacritica/

3. Salud & ambiente: Geografías en sacrificio. Vol. 59, Cuadernos Médico Sociales (Chile). 2019.

4. Philip J Landrigan NJRA Richard Fuller. The lancet commission on pollution and health. *The Lancet*. 2018;391(10119):475.

5. https://github.com/verasativa/CIE-10/

6. Cavieres MF. Estudios sobre la contaminación de puchuncaví en la decada de los 80. Un aporte científico que no fue. *Cuadernos Médico Sociales (Chile)*. 2019;59(1):35.