Propuesta de metodología para la priorización de distrito de intervención (parte I)

Equipo de Gestión de la información y políticas públicas

El presente documento brinda, a grandes rasgos, la descripción de la metodología utilizada para priorizar un distrito en el que se pueda llevar a cabo una estrategia especializada para la intervención al *microtráfico*.

Este documento explica una técnica para obtejer un puntaje compuesto para una única variable que contiene varias observaciones en el tiempo, con el fin de considerar su información más reciente y al mismo tiempo considerar su variación en el tiempo. Esta técnica puede ser replicada en una gran cantidad de variables numéricas, de manera independiente.

Es necesario que en otro documento se pueda trabajar una metodología para la interacción entre diferentes variables.

# 1. Estructura básica de los datos a utilizar

En términos de ubicación física, se requiere que los datos disponibles estén agregados a nivel distrital. Idealmente, se contará con el código de ubicación geográfica del distrito (ubigeo), tal como es utilizado por el Instituto Nacional de Estadística e Informática (INEI). El ubigeo debe tener su propia columna en el conjunto de datos. En paralelo, se requiere contar con metadatos a nivel distrital, donde se incluyan los nombres de departamento, provincia, distrito y geometrías (para el dibujado de mapas).

A nivel temporal, se requiere conocer el periodo anual al cual hacen referencia los datos. Esta información debe existir como columna en el conjunto de datos, su nombre puede ser periodo\_anual.

Cada medición o variable debe existir en su propia columna en el conjunto de datos. Debe tomarse en cuenta que la medición no debe existir en términos absolutos (por ejemplo: número de denuncias, número de detenidos) sino en términos relativos (tasa de denuncia, porcentaje de población con cierta característica, PBI per cápita, etc). Para ello, resulta útil conocer los detalles técnicos de la medición y qué tanto podrían verse sesgados por el factor poblacional. El conjunto de datos puede tener tantos columnas de medición como sean necesarias.

El siguiente conjunto de datos muestra un ejemplo de la estructura básica. Se puede apreciar que se tiene información para 6 ubigeos en el periodo anual 2018. La variable de medición es el porcentaje de la población distrital en condiciones de pobreza monetaria.

| **ubigeo** | **periodo\_anual** | **pobreza\_monetaria** |
| --- | --- | --- |
| 010506 | 2018 | 40.2 |
| 150407 | 2018 | 10.8 |
| 150811 | 2018 | 18.4 |
| 150125 | 2018 | 22.7 |
| 150702 | 2018 | 9.8 |
| 170103 | 2018 | 4.8 |

Para tener un seguimiento a través del tiempo, se requiere que haya múltiples periodos anuales de medición de las mismas variables. En el siguiente conjunto de datos, se pueden ver 9 observaciones de la variable “Intervenciones de serenazgo por consumo de drogas” (3 ubigeos por 3 periodos anuales).

| **ubigeo** | **periodo\_anual** | **int\_consumo\_drogas** |
| --- | --- | --- |
| 150103 | 2019 | 809 |
| 150103 | 2020 | 1411 |
| 150103 | 2021 | 821 |
| 150105 | 2019 | 62 |
| 150105 | 2020 | 75 |
| 150105 | 2021 | 200 |
| 150120 | 2019 | 459 |
| 150120 | 2020 | 193 |
| 150120 | 2021 | 190 |

Como se mencionó anteriormente, se debe tener cuidado de no utilizar variables con valores absolutos, debido a la alta posibilidad de que en distritos con mayor población haya mayor número de intervenciones. Para evitar este sesgo metodológico se recalcula la variable ponderando el factor población, obteniendo así la “Tasa de intervenciones de serenazgo por consumo de drogas, por 10 mil habitantes”. El siguiente cuadro ilustra la ruta para el cálculo usando la población estimada al 2020.

| **ubigeo** | **periodo\_anual** | **int\_consumo\_drogas** | **pob\_2020** | **tasa\_int\_consumo\_drogas** |
| --- | --- | --- | --- | --- |
| 150103 | 2019 | 809 | 670818 | 12.1 |
| 150103 | 2020 | 1411 | 670818 | 21.0 |
| 150103 | 2021 | 821 | 670818 | 12.2 |
| 150105 | 2019 | 62 | 93111 | 6.7 |
| 150105 | 2020 | 75 | 93111 | 8.1 |
| 150105 | 2021 | 200 | 93111 | 21.5 |
| 150120 | 2019 | 459 | 65139 | 70.5 |
| 150120 | 2020 | 193 | 65139 | 29.6 |
| 150120 | 2021 | 190 | 65139 | 29.2 |

Con ello, puede verse que se reduce la magnitud de diferencia entre en las mediciones al comparar distritos con poblaciones diferentes. El siguiente cuadro muestra cómo luciría la tabla después de ser procesada.

| **ubigeo** | **periodo\_anual** | **tasa\_int\_consumo\_drogas** |
| --- | --- | --- |
| 150103 | 2019 | 12.1 |
| 150103 | 2020 | 21.0 |
| 150103 | 2021 | 12.2 |
| 150105 | 2019 | 6.7 |
| 150105 | 2020 | 8.1 |
| 150105 | 2021 | 21.5 |
| 150120 | 2019 | 70.5 |
| 150120 | 2020 | 29.6 |
| 150120 | 2021 | 29.2 |

Esta tabla servirá como ejemplo para explicar la metodología de cálculo de una medida representativa para cada ubigeo.

# 2. Metodología de agregación de datos

La agregación de datos se realiza para tener condensar múltiples observaciones en un solo valor que sea, en alguna medida específica, representativa de su conjunto (en este caso, para cada ubigeo). Entre las técnicas más conocidas de agregación se encuentran la obtención de recuentos, sumatorias, y medidas de tendencia central (como promedio, mediana, etc).

Debido a que la estructura básica de datos con la que se trabaja para esta investigación contiene mediciones a lo largo del tiempo se ha buscado obtener una medida que pueda representar:

1. El estado más actual de la observación
2. Su tendencia, a través de su cambio en un periodo de tiempo determinado

A continuación se explica cómo obtener ambas agregaciones y sus limitaciones.

## 2.1 Agregaciones básicas

El estado más actual de la observación puede obtenerse fácilmente, extrayendo la medición obtenida en el periodo más reciente. Para obtener la tendencia, se calcula la diferencia entre la observación más antigua y la más reciente en términos relativos. En otras palabras, se obtiene el porcentaje de variación. El siguiente cuadro ilustra este proceso para la tabla de ejemplo.

| **ubigeo** | **mas\_antigua** | **mas\_reciente** | **cambio** | **porc\_variacion** |
| --- | --- | --- | --- | --- |
| 150103 | 12.1 | 12.2 | 0.1 | 0.8 |
| 150105 | 6.7 | 21.5 | 14.8 | 220.9 |
| 150120 | 70.5 | 29.2 | -41.3 | -58.6 |

En este ejemplo en particular, resulta valioso apreciar la diferencia entre el ubigeo “150120” y el ubigeo “150105”. Se puede ver que en las observaciones más antiguas, la medición de “150120” es de 70.5 intervenciones por cada 10 mil habitantes, alrededor de 6 veces la cantidad que en “150129”. Esta diferencia se hace mucho un poco menos evidente en las observaciones más recientes, en la que la tasa de intervenciones en “150120” es alrededor de 2 veces la de “150105”. Si se tomara como referencia la medición más reciente, resultaría más importante centrarse en el ubigeo “150120”.

Por otro lado, se puede ver que no hubo mucho cambio en términos absolutos para “150103”, mientras que en “150120” hubo un decremento de 41.3 unidades. Asimismo, el ubigeo “150105” registro un incremento de 14.8 unidades. Al analizar los cambios en términos relativos (a través del porcentaje de variación), se puede ver que el cambio en “150105” resulta más extremo que en el resto de ubigeos. si se tomara como referencia la variación relativa, habría que centrarse en el ubigeo “150105”.

Con toda esta explicación, se está tratando de argumentar que no es suficiente tomar solo la última medición o solo el porcentaje de variación en el tiempo, ya que ambas agregaciones brindan información importante. Ante ello, es necesario elaborar una agregación compuesta, que se alimente de ambas al mismo tiempo.

## 2.2 Agregación compuesta

La siguiente fórmula ilustra el cálculo necesario para la obtención de un puntaje compuesto. En ella, se asigna el mismo peso (0.5) a cada agregación, de tal modo que Puntaje se alimente de ambas en igual medida. Esta formula contiene un error grave, que se explicará en breve.

La siguiente tabla muestra los resultados obtenidos, ordenados de manera descendente por el puntaje obtenido:

| **ubigeo** | **mas\_reciente** | **porc\_variacion** | **puntaje** |
| --- | --- | --- | --- |
| 150105 | 21.5 | 220.9 | 121.2 |
| 150103 | 12.2 | 0.8 | 6.5 |
| 150120 | 29.2 | -58.6 | -14.7 |

Con estos resultados, el puntaje obtenido indicaría que se debe priorizar intervenir en el ubigeo “150105” y que además existe una gran distancia entre este y el segundo ubigeo con mayor puntaje.

Sin embargo, la fórmula previamente mostrada no toma en cuenta que la observación más reciente de la *tasa de intervenciones por 10 mil habitantes* opera en un nivel de abstracción diferente al *porcentaje de variación* de la tasa. Es decir que aplicar la formula sin pre-procesar las variables equivale a sumar gotas de agua con baldes de agua: se obtiene un número pero este no sirve para extraer conclusiones razonables.

Para el pre-procesamiento de estas agregaciones se ha decidido usar una técnica llamada **normalización**[[1]](#footnote-22). Esta consiste en re-escalar los valores existentes a un rango entre cero (0) y uno (1), donde el valor mínimo observado se convierte en cero (0), el valor máximo observado se convierte en uno (1), y el resto de valores se escalan proporcionalmente. La siguiente tabla muestra la aplicación de este proceso a nuestros datos.

| **ubigeo** | **mas\_reciente** | **porc\_variacion** | **norm\_reciente** | **norm\_variacion** |
| --- | --- | --- | --- | --- |
| 150103 | 12.2 | 0.8 | 0.0 | 0.2 |
| 150105 | 21.5 | 220.9 | 0.5 | 1.0 |
| 150120 | 29.2 | -58.6 | 1.0 | 0.0 |

Puede verse que se obtuvieron las columnas norm\_reciente y norm\_variacion como resultado. Teniendo esto, es posible actualizar la forma de cálculo.

La siguiente tabla muestra el resultado de aplicar la fórmula en el conjunto de datos, ordenados de manera descendente por el puntaje obtenido:

| **ubigeo** | **norm\_reciente** | **norm\_variacion** | **puntaje** |
| --- | --- | --- | --- |
| 150105 | 0.5 | 1.0 | 0.8 |
| 150120 | 1.0 | 0.0 | 0.5 |
| 150103 | 0.0 | 0.2 | 0.1 |

En este caso, se sigue manteniendo el ubigeo “150105” como el del puntaje más alto, pero se reduce la distancia entre este y el segundo lugar. Sumado a ello, el ubigeo “150120” obtiene un mayor puntaje, ya que se toma mejor en cuenta su medición más reciente, en la que tiene el valor más alto.

## 2.3 Posibles mejoras

Es posible que el porcentaje de variación muestre cambios extremos en valores inesperados. Por ejemplo, se puede considerar el siguiente ejemplo:

| **item** | **mas\_antiguo** | **mas\_reciente** |
| --- | --- | --- |
| 1 | 42.2 | 50.6 |
| 2 | 43.8 | 58.6 |
| 3 | 52.8 | 52.3 |
| 4 | 45.4 | 43.7 |
| 5 | 0.5 | 3.0 |

Resulta evidente que el item 5 tiene valores muy bajos en comparación con el resto. Sin embargo, al realizar el cálculo del puntaje con variables normalizadas, se obtiene el siguiente resultado (ordenado en forma descendente según puntaje).

| **item** | **mas\_reciente** | **porc\_variacion** | **norm\_reciente** | **norm\_variacion** | **puntaje** |
| --- | --- | --- | --- | --- | --- |
| 2 | 58.6 | 33.6 | 1.0 | 0.1 | 0.5 |
| 5 | 3.0 | 500.0 | 0.0 | 1.0 | 0.5 |
| 1 | 50.6 | 20.0 | 0.9 | 0.0 | 0.5 |
| 3 | 52.3 | -0.9 | 0.9 | 0.0 | 0.4 |
| 4 | 43.7 | -3.7 | 0.7 | 0.0 | 0.4 |

Se puede ver que el item 5 aparece en segundo lugar, superando a items que tienen valores mucho mayores en su observación más reciente. Esto se debe a que su porcentaje de variación es extremadamente mayor que el del resto. Para contrarestar este efecto, se puede actualizar la fórmula de la siguiente manera:

Donde:

* : Peso asignado al valor más reciente
* : Valor más reciente, normalizado
* : Porcentaje de variación, normalizado

En los cálculos anteriores, tenía un valor de 0.5, lo que le asignaba la misma importancia al valor más reciente y al porcentaje de variación. Si se le asigna un valor de 0.6 a $wt\_\reciente$ se obtienen los siguientes puntajes:

| **item** | **mas\_reciente** | **porc\_variacion** | **norm\_reciente** | **norm\_variacion** | **puntaje** |
| --- | --- | --- | --- | --- | --- |
| 2 | 58.6 | 33.6 | 1.0 | 0.1 | 0.6 |
| 3 | 52.3 | -0.9 | 0.9 | 0.0 | 0.5 |
| 1 | 50.6 | 20.0 | 0.9 | 0.0 | 0.5 |
| 4 | 43.7 | -3.7 | 0.7 | 0.0 | 0.4 |
| 5 | 3.0 | 500.0 | 0.0 | 1.0 | 0.4 |

Como puede verse, de esta manera se reduce el peso que el porcentaje de variación tiene en el puntaje final. Es necesario probar otros límites de la fórmula hasta ir consiguiendo versiones más optimas y más robustas a valores extremos.

1. Para este caso en específico se ha decidio usar la normalización min-max. Existe otra técnica de normalización que escala los valores para obtener un promedio de cero (0) y una desviación estándar de uno (1), que es menos sensible a valores extremos. Se decidió no usar esta otra técnica debido a que los datos ya se encuentran ponderados por población, reduciendo la posible aparición de valores extremos. Además, mantener los valores en una escala del 0 al 1 facilita ponderar el puntaje normalizado por factores que, sumados, equivalgan a 1 también. [↑](#footnote-ref-22)